Constructing Strata of PSUs for the Residential Energy Consumption Survey

Rachel Harter[1], Patrick Chen[1], Joseph McMichael[1], Edgardo Cureg[2], Samson Adeshiyan[2], Katherine B. Morton[1]
[1]RTI International, 3040 E. Cornwallis Road, Research Triangle Park, NC 27709
[2]Energy Information Administration, 1000 Independence Ave. SW, Washington, DC 20585

Abstract[1]
Textbooks provide guidance on the general principles and desirable properties of defining sampling strata. This paper reviews some basic exploratory methods for determining stratification variables, including principal component analysis, cluster analysis, correlations, regression analysis, and decision trees for reducing the set of potential variables. Although all stratification methods use auxiliary variables available for the entire frame, the decision tree, regression, and correlation approaches also use prior outcome data, which may be available for just a sample of units. The principal component method combined with cluster analysis, on the other hand, focuses on relationships among stratification variables. Using both principal components/cluster analysis and decision trees, we stratify primary sampling units for the Residential Energy Consumption Survey and compare the resulting strata.

Key Words: strata, decision tree, principal components, cluster analysis

1. Introduction

Textbooks of survey sampling define stratification as the process of dividing the target population into separate subpopulations, where each sampling unit is in one and only one stratum. In a stratified design, samples are selected independently in each stratum, and the stratum estimates are combined to form total population estimates. In general, it is desirable for members of a stratum to be as much alike as possible, and for the strata to be as different as possible.

Stratified designs are popular for a number of reasons:

1. Estimates may be desired for subpopulations as well as the total population, or certain subpopulations may be oversampled.
2. Stratification usually results in more precise estimates (lower variance).
3. Sampling less expensive strata at a higher rate can help reduce overall costs (with precision tradeoffs).
4. The strata may be logical subdivisions for managing or equalizing workload.
5. Different sampling or data collection strategies may be appropriate for different subpopulations.
6. Different frames may be available for different subpopulations.
7. Stratified designs help protect against an unrepresentative random sample.

[1] The analysis and conclusions contained in this paper are those of the author(s) and do not represent the official position of the U.S. Energy Information Administration (EIA), the U.S. Department of Energy (DOE), or RTI International.

Stratified designs are very common, and the first three reasons may be the most prevalent justifications. The reasons for stratification are study-specific, however.

This paper will first review some common methods of forming strata, with particular emphasis on methods for identifying auxiliary variables to be used in defining strata. Then we describe the particular survey (Residential Energy Consumption Survey or RECS) for which we defined sampling strata. We review two basic methods that were considered for defining RECS strata for primary sampling units (PSUs), and summarize a simulation that we conducted to investigate the likely outcomes.

2. Defining Strata

Sometimes strata are defined by observed characteristics of the population, such as geographical location (e.g. census region or division, state), administrative grouping (e.g. school district), or traits (e.g. age range, industry). A general principle is that the variable(s) used for stratification must be available for every unit in the frame. (There are ways of dealing with exceptions, which are not covered here.)

When the goal of stratification is to estimate domains, or to oversample some domains, then the domain identifiers are natural choices for defining strata. For example, if estimates of employment are desired by state, then state is a natural stratifier. For another example, if estimates of student reading scores by race are desired, and if student race is available in the frame, then race is a natural choice of stratifier.

If estimates for two sets of crossed domains are desired, such as census division and urban vs. rural, or race and age, both classification variables can be used to define strata. The sample can be allocated to strata using a composite size measure (Folsom, Potter, and Williams 1987) to form an equal probably sample that supports estimation in both sets of domains. Composite measures of size can be defined for multiple crossed characteristic variables.

Beyond the domain variables, stratification often follows a process similar to these steps:

1. Identify goals of the stratification
2. Identify relevant auxiliary variables
3. Clarify assumptions regarding auxiliary variables and variables of interest
4. Look for combinations and transformations of the auxiliary variables to reduce dimensionality
5. Define useful subdivisions of the auxiliary variable values

These steps are illustrated with a simple procedure known as the "cum $\sqrt{f}$" rule (Cochran 1977, pp. 127-131; Dalenius & Hodges 1959). Suppose the goal of stratification is to estimate $\hat{Y}$ or $\hat{\bar{Y}}$ while forming approximately equally-sized strata and sampling approximately equal workloads per stratum. Suppose that the study variable y is highly correlated with a covariate x that is available for all units in the frame. Then x is a likely stratifier for the design. The assumption is that good stratification for x is also good stratification for y. Let $f(x)$ be the density function of the distribution of x. The cum $\sqrt{f}$ rule (and its variations and approximations) have been developed to establish dividing points (stratum boundaries) in the distribution of $f(x)$ to minimize the variance of $\hat{X}$ under Neyman or proportional allocation of sample to strata, based on the idea that $f(x)$ is approximately constant within each stratum.

Most surveys collect data on more than one y variable, although designers often focus on one or two of the most important outcome variables. Cochran (1977) reviewed a number of methods for stratifying for multiple y variables. Here we will focus on a single y. Also, more than one covariate x may be available as stratifiers. There are multiple ways one can select and use covariates to define strata. Ideally the strata defined by the covariates should be internally homogenous with respect to y and as heterogeneous as possible at the stratum level. In general, the stratifiers should be highly correlated with the y variable and capture much of the y's variability.

Subject matter experts are a great source of information about potential stratifiers. They know what variables are available, what variables are correlated, and what has been used in the past. The expertise may be readily available in the literature. Sometimes expert advice is sufficient for identifying stratification variables. If the y variables are available from another source or a prior survey, it is useful to compute simple correlations with the auxiliary variables to identify potential stratifiers.

The main assumption of using auxiliary variables to define strata is that the variation in the x variables used to define strata captures a large portion of the variations in the y variables. When the variation within a stratum is small, the variance of the total estimate, which is summed from the independently sampled strata, will also be smaller than if the sample were drawn randomly from the entire frame without stratification.

Whether or not to use historical or alternative y variables is a philosophical issue closely tied to the assumptions one is willing to make. Historical y values generally are not available for all units in the frame, or are available only in aggregate. With historical y values for some units, the potential stratifiers could be used to predict $\hat{y}$ for all units in the frame, and the cum $\sqrt{f}$ rule could be applied to the distribution of $\hat{y}$ (Eltinge 2015). When historical y values are used in some way, the assumption is that the correlations between the auxiliary variables and the historical y variable(s) are strong predictors of the correlations between the auxiliary variables and the y variables to be collected in the survey.

If the potential stratifiers are continuous variables, or if many cross classifications of the variables are possible, then the next step is to decide how many strata to use and how to reduce the possible subdivisions of the frame based on the auxiliary variables. First we focus on variable reduction and/or level reduction to capture most of the variability in the y variable(s) with fewer stratifying variables or consolidated ranges of their values.

A well-known variable reduction technique is principal component analysis. Principal component analysis can be used to find the best combinations of variables to use as stratifiers. The first principal component can be the one x variable used with the cum $\sqrt{f}$ rule, for example. Or the first two or three principal components can be used to capture multiple dimensions of the variability. The principal components are actually combinations and transformations of the original x variables. The focus of principal components is the relationships among the auxiliary variables, assuming that dividing the frame into homogeneous groups based on the principal components will also create relatively homogeneous groups in terms of the y variable(s).

Other methods can be used to reduce the number of variables. Expert judgment is commonly used. When historical y values are used, correlations, regression analysis, and decision trees can help identify the most relevant stratifiers. Variables with little or no correlation with the *y* variable are unlikely to be useful stratifiers. Correlations alone, however, will not necessarily identify interactions among the *x* variables. Parameter estimates from regression models can be used to test the significance of the *x* variables, but multicollinearity is also a consideration. Both decision trees and stepwise regressions reduce multicollinearity by identifying explanatory variables sequentially. Decision trees have the added advantage of determining likely subdivisions of the auxiliary variables to best explain the variation in the *y* variable(s). The number of nodes can be pre-specified, if desired, and the nodes correspond to the strata. Alternatively, the decision trees can be grown to their full extent, and then trimmed back to the desired number of strata. Because nodes can have as few as one frame member, growing the full tree and then trimming may be preferred over pre-specifying the number of nodes.

Cluster analysis can be used to categorize the frame based on the values of the stratifiers. Although stratifiers can be collapsed and combined in an ad hoc manner, cluster analysis provides scientific justification for the strata. The U.S. Census Bureau uses the Friedman-Rubin clustering algorithm (Friedman & Rubin 1967) for stratification for its demographic surveys, most notably the Current Population Survey (Judkins & Singh 1981; Mansur & Reist 2010). Cluster analysis can be performed on the original *x* variables or the principal components. Cluster analysis will place every frame unit into one and only one stratum, but the stratum boundaries are not necessarily well-defined; if there are any later additions to the frame, they might not be assigned to one stratum uniquely. For most applications this is not an issue.

Using the terminal nodes of decision trees as strata is an alternative approach, combining the variable reduction and the grouping of variable values into a single step. Decision trees set very clear boundaries for stratum definitions. The boundaries are not necessarily intuitive, however.

The issue of optimal stratification has been considered by several researchers. Dalenius (1950, 1957) and others developed methods of approximating optimal stratum boundaries under Neyman or proportional allocation. Lavallée and Hidiroglou (1988) developed an algorithm for stratifying skewed distributions based on stratum boundary work by Sethi (1963). Stratification does not need to be optimal to be useful, however. The standard errors of estimates from a stratified sample will rarely exceed those from a simple random sample of the same size, thus even imperfect stratification does not damage the survey estimates (Lavrakas, 2008). In other words, any reasonable stratification is likely to yield improvements.

Sometimes the stratum boundary methods themselves help inform the number of strata that should be used. In general, subdomains for reporting make useful strata. Otherwise, with geographic subdomains, the number of strata depends largely on the usefulness of the *x* variables in capturing the variability in the *y* variables. Cochran indicated that, unless the correlations between *x* and *y* variables were extremely high (>95%), there is unlikely to be useful gain from more than six strata. Or, as Lohr (1999, p. 110) indicated, "The less information, the fewer strata you should use." For RECS, the number of strata within a domain was approximately 5.

Frequently the stratifying variables are used in their current form, without transformation, even if the values of the variables are combined into ranges. For example, age may be consolidated into ranges, but otherwise maintained as a variable defined as a simple count of the years a person has lived. One advantage of maintaining the original variable forms is that the strata are easier to describe. Readily understandable strata may be important for studies that are in the public domain or that have to be explained to funders. Most methods retain the form of the variables, even if the values divide the strata. Principal components, however, transform the variables to be orthogonal—no multicollinearity; however the interpretation of the principal components is not necessarily clear.

3. Application to the Residential Energy Consumption Survey

The principal purpose of the Residential Energy Consumption Survey (RECS) is to provide Congress and the general public periodic estimates of the total number of occupied, primary residential units in the 50 States and the District of Columbia, as well as the total amount and cost of electricity, natural gas, bulk fuels (fuel oil, propane, and kerosene), and wood they consume. RECS also provides estimates of how much of the residential sector's total energy consumption is for space heating, air conditioning, water heating, refrigeration, and other end uses.

Additionally, RECS reports information on energy-related characteristics of the residential sector. These characteristics include building structure and square footage, household demographics, appliance inventories and usage patterns, and measures of energy efficiency.

Together with the Commercial Building Energy Consumption Survey (CBECS) and the Manufacturing Energy Consumption Survey (MECS), RECS completes the trio of surveys that the U.S. Energy Information Administration (EIA) conducts to describe the consumption of energy within the overall U.S. economy. More information on RECS is available at the EIA website (http://www.eia.gov/consumption/residential/).

The target population for the 2015 RECS is all housing units (HUs) occupied as primary HUs in the 50 states and the District of Columbia. Vacation homes, seasonal HUs, and group quarters, such as dormitories, nursing homes, prisons, and military barracks, are excluded from the study; however, HUs on military installations are included. The 2015 RECS includes two components: a Household Survey and a Rental Agent Survey. For the latter, the sample rental agents are identified by HUs responding to the Household Survey that are occupied by renters or where owners indicate that some or all utility or bulk fuel energy bills are not paid directly by the household. Therefore, sample design focuses on the selection of housing units.

A stratified three-stage sample design was used for the 2015 RECS Household Survey. At the first stage, a stratified sample of 200 PSUs from the Census Bureau's Public Use Microdata Areas (PUMAs) was selected with probability proportional to size and with minimal replacement (Chromy, 1979). At the second stage, secondary sampling units (SSUs) defined by census block groups (CBGs) were selected within each sampled PSU with probability proportional to size and with minimal replacement. Prior to selecting the sample, SSUs were sorted by urban/rural, socioeconomic status (SES) indicator, and proportion of newly constructed HUs within the SSU. From this well-ordered frame, four SSUs were selected within each PSU, for a total of 800 SSUs. For the third stage of selection, HU sampling frames were constructed using one of three methods: (1) address based sampling (ABS) mailing list only; (2) ABS supplemented by RTI's frame-linking procedure, Check for Housing Units Missed (CHUM) (McMichael et al., 2008, 2013); and (3) traditional field enumeration. This hybrid method of constructing HU frames was used to obtain high coverage while controlling costs. A systematic sample of HUs was selected from the HU sampling frame in each segment.

The 2015 RECS is required to produce estimates for the nation, for 19 geographic domains (subdivisions of Census Divisions), and four large states with a specified level of precision. The goal of the stratification is to help reduce the variability in the estimates while controlling the costs.

Given the estimation goals, the first-level stratification of PSUs divides the United States into the 19 geographic domains, as for the 2009 RECS. The remainder of this RECS example is concerned with the stratification of PSUs (PUMAs) within geographic domains.

Within each geographical domain, there are variations in energy consumption among PSUs. For example, a PSU with a high proportion of detached single-family HUs has higher energy consumption than a PSU with a low detached single HUs proportion. Associating PSU characteristics to energy consumption helps to divide PSUs into groups with similar energy consumption. Based on the 2009 RECS total household energy consumption estimates, we identified 10 PSU (PUMA-level) characteristics from the National Oceanic and Atmospheric Administration and the American Community Survey that are correlated with energy consumption (and often with each other). In some cases we restructured the variables to be categorical.

1. Average Heating Degree Days
2. Average Cooling Degree Days
3. Urban/Rural
4. Housing Unit Type
5. Own vs. Rent
6. Year Built
7. Housing Unit Size
8. Housing Unit Income
9. Latitude
10. Major Heating Fuel Type

We used the last cycle of RECS, for which 2009 was the reference year, to develop strata of PSUs for the 2015 RECS. In brief, we merged the most current available values of these PUMA-level variables to the 2009 RECS microdata and assigned the 2009 sample HUs to strata defined in two ways. One method used the historical total household energy consumption as the dependent y variable for the housing unit and the PUMA-level x variables in a decision tree analysis. The other method determined the principal components of the PUMA-level x variables, and applied cluster analysis at the PUMA level to the first two or three principal components.

Both methods assume that stratification that captures much of the variability in the x variables will also capture much of the variability in the current y variables. Using the historical y values in the decision tree method also assumes that strong relationships between the current x variables and historical y variable indicate that relationships with the current y variable will be strong, even in PUMAs for which we have no 2009 RECS observations. Both methods assume that the 2009 RECS sample design is ignorable for developing and testing the strata for the 2015 RECS.

The decision tree method used the 2009 RECS household total energy consumption as the dependent variable and 18 PUMA-level variables (functions of the variables listed above, see Table 1) as independent variables. The Chi-square Automatic Interaction Detection (CHAID) algorithm (Kass 1980) was used to grow the decision tree. CHAID uses chi-squared statistics to identify optimal splits and allows multiple node splitting. It sequentially chooses the independent (predictor) variables that have the strongest interaction with the dependent variable. Comparing to regression models, CHAID is not susceptible to the collinearity among independent variables because one variable is selected to split the tree at a time (Sambandam). CHAID classified HUs into groups (terminal nodes) with similar PUMA characteristics associated with energy consumption. Figure 1 shows the decision tree stratification of PUMAs in Domain 1, New England, which is composed of 109 PUMAs. The strata are defined by the proportion of housing units with three or more bedrooms, median household income, the proportion of housing units built in 1970 or later, and the proportion of single-family detached homes. The PUMAs in their assigned strata are shown on the map in Figure 2.

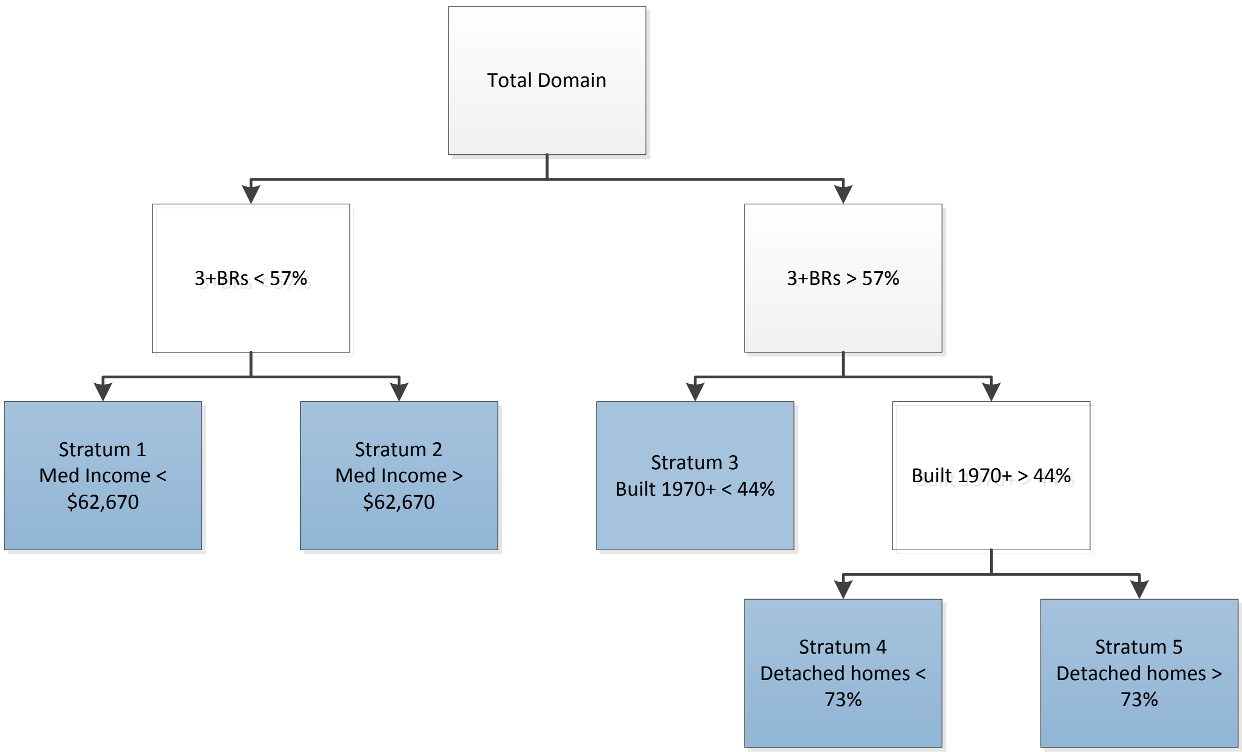

Figure 1. Decision Tree and Node Strata for Domain 1 (New England)

The same 18 PUMA-level variables were reduced by computing principal component analysis. Table 1 summarizes the loading factors for the major principal components in Domain 1.

Table 1. Loading Factors for Principal Components in Domain 1

X variable	Factor 1	Factor 2	Factor 3
Cooling degree days	0.81	-0.02	-0.39
Heating degree days	-0.89	-0.03	0.28
% in urban areas	0.89	-0.01	-0.13
% in urban areas/clusters	0.88	-0.02	-0.02
% single family detached	-0.23	0.83	0.42
% single family	-0.17	0.85	0.43
% owner-occupied	-0.20	0.87	0.37
% built 1970+	-0.28	0.38	0.84
% built 1980+	-0.28	0.35	0.87
% built 1990+	-0.27	0.32	0.87
% built 2000+	-0.21	0.20	0.89
% 3+ bedrooms	-0.03	0.92	0.25
% 4+ bedrooms	0.17	0.88	0.11
Median household income	0.37	0.76	0.02
PUMA latitude	-0.82	-0.15	0.23
% natural gas heating fuel	0.62	-0.20	-0.53
% electricity heating fuel	0.51	-0.51	-0.20
% other heating fuel	-0.68	0.31	0.52
Interpretation of Principal Components from Largest (Shaded) Loading Factors	Climate	Building Size/Type	Building Age

Cluster analysis using Ward's minimum variance method (Ward 1963) was applied to the principal components whose eigenvalues were greater than 1. For this investigation, we forced the number of clusters to match the number of terminal nodes from the decision tree in the same domain. Figure 3 shows the strata defined through the clustering for Domain 1. The strata defined by the two methods are clearly different.

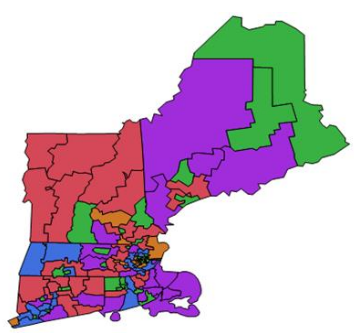

Figure 2: Decision Tree Strata

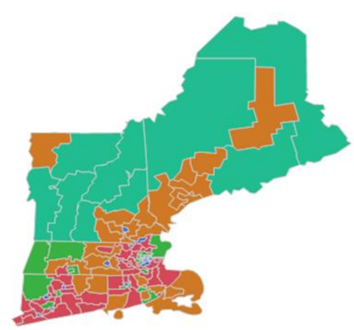

Figure 3: Cluster Analysis Strata

To test the two stratification methods, we used additional *y* variables from the 2009 RECS: household electrical usage, household natural gas usage, and total household dollars spent on energy. For three geographic domains, we decomposed the variability of these other *y* variables into within-stratum and between-strata components. The stratification that leads to a smaller within-stratum variance and a larger between-stratum variance across variables and domains does a better job of capturing the variability in the *y* variables.

Table 2 shows the ratio of the decision tree variance components to the corresponding cluster analysis components for easier comparison. The tables demonstrate that neither method is uniformly better. The decision tree method tends to capture more variability for electricity consumption and energy dollars, but the cluster analysis method is better for natural gas.

Table 2. Variance Decomposition Ratios

Domain 1 (New England)	**Within Strata**	**Between Strata**
Total Electricity Usage (thousand BTU)	0.96	2.27
Total Natural Gas Usage (hundred cubic feet)	1.10	0.25
Total Energy Cost (dollars)	0.93	3.62
Domain 2 (New York)		
Total Electricity Usage (thousand BTU)	0.98	1.13
Total Natural Gas Usage (hundred cubic feet)	1.01	0.79
Total Energy Cost (dollars)	0.99	1.04
Domain 16 (California)		
Total Electricity Usage (thousand BTU)	0.98	1.41
Total Natural Gas Usage (hundred cubic feet)	1.01	0.62
Total Energy Cost (dollars)	0.98	1.36

Both the decision tree approach and cluster analysis approach are reasonable. Without time and budget constraints, perhaps stratification could have been improved. For the 2015 RECS within-domain stratification of PUMAs at the PSU selection stage, the decision tree approach was used.

4. Concluding Remarks

Some aspects of stratification were glossed over in the general discussion. And the simulation conducted for RECS was useful but not conclusive. Many studies have neither the time nor the resources for extensive work on stratification. Stratifiers are often error-prone, and misclassification of frame units into strata is not uncommon. In general, the goal is not to find the “best” stratification, if indeed it could be known. Rather, the goal is to find reasonably good stratification that will support the analytical needs while reducing the variance of the estimates compared with simpler designs. Stratification need not be “best” or “perfect” to perform well.

Even so, this paper is intended to provide some structure to the stratification process, or at least some considerations for the sample designer. The stratification of PSUs for RECS illustrates the basic steps and compares results for two reasonable stratification methods.

References

Cochran, W.G. (1977). *Sampling Techniques: third edition*. New York: John Wiley & Sons.

Chromy, J. R. (1979). Sequential sample selection methods. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 401-406.

Dalenius, T. (1950). The problem of optimum stratification. *Scandinavian Actuarial Journal*, 1950 (3-4), 203-213.

Dalenius, T. (1957). *Sampling in Sweden*. Almqvist and Wicksell, Stockholm.

Dalenius, T. and Hodges, J.L. Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Eltinge, J. (2015). E-mail memorandum to R. Harter, July 15, 2015.

Folsom, R.E., Potter, F.J., and Williams, S.R. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. In *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 792-796.

Friedman, H. P. and Rubin, J. (1967). On Some Invariant Criteria for Grouping Data. *Journal of the American Statistical Association*, 62, 1159-1178.

Judkins, D. R., & Singh, R. P. (1981). Using clustering algorithms to stratify primary sampling units. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 279-284.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29 (2), 119-127.

Lavrakas, P.J., ed. (2008). *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: Sage Publications, Inc.

Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Mansur, K.A. and Reist, B.M. (2010). Evaluating Alternative Criteria for Primary Sampling Units Stratification. In *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 4664-4672.

McMichael, J., Ridenhour, J., & Shook-Sa, B. (2008). A robust procedure to supplement the coverage of address-based sampling frames for household surveys. In *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 4329-35.

McMichael, J. P., Shook-Sa, B. E., Ridenhour, J. L., & Harter, R. (2014, April). The CHUM: A frame supplementation procedure for address-based sampling. In *2013*

Research Conference Papers (http://fcsm.gov/events/papers2013.html). Washington, DC.

Sambandam, R. (date unknown). Cluster Analysis Gets Complicated. White paper, TRC Market Research. (http://www.trchome.com/component/content/article?id=146:cluster-analysis).

Sethi, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58, 236–244.