RTI Press

Occasional Paper

May 2017

Constructing Strata of Primary Sampling Units for the Residential Energy Consumption Survey

Rachel M. Harter, Pinliang (Patrick) Chen,
Joseph P. McMichael, Edgardo S. Cureg,
Samson A. Adeshiyan, and Katherine B. Morton

RTI
INTERNATIONAL

RTI Press publication OP-0041-1705

Suggested Citation

Harter, R. M. Chen, P., McMichael, J. P., Cureg, E. S., Adeshiyan, S. A., and Morton, K. B. (2017). *Constructing Strata of Primary Sampling Units for the Residential Energy Consumption Survey*. RTI Press Publication No. OP-0041-1705. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2017.op.0041.1705

This publication is part of the RTI Press Research Report series. Occasional Papers are scholarly essays on policy, methods, or other topics relevant to RTI areas of research or technical focus.

RTI International
3040 East Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
E-mail: rtipress@rti.org
Website: www.rti.org

The analysis and conclusions contained in this paper are those of the authors and do not represent the official position of the US Energy Information Administration (EIA), the US Department of Energy (DOE), or RTI International.

https://doi.org/10.3768/rtipress.2017.op.0041.1705 www.rti.org/rtipress

Contents

About the Authors

Rachel M. Harter, PhD, is a senior research statistician at RTI International, Division for Statistical & Data Sciences.

Pinliang (Patrick) Chen, MS, is a senior research statistician at RTI International, Division for Statistical & Data Sciences.

Joseph P. McMichael, BS, is a research statistician at RTI International, Division for Statistical & Data Sciences.

Edgardo S. Cureg, PhD, is a lead mathematical statistician at the US Energy Information Administration, Washington, DC.

Samson A. Adeshiyan, PhD, is chief statistician at the National Science Foundation, Arlington, VA. This work was completed when he was a lead mathematical statistician at the US Energy Information Administration, Washington, DC.

Katherine B. Morton, MS, is a senior research statistician at RTI International, Division for Statistical & Data Sciences.

Acknowledgments

The authors gratefully acknowledge the US Energy Information Administration for supporting the research in designing the 2015 Residential Energy Consumption Survey, RTI International for manuscript assistance, and Phillip Kott, John Eltinge, and the RTI Press reviewers for helpful comments.

Abstract

The 2015 Residential Energy Consumption Survey design called for stratification of primary sampling units to improve estimation. Two methods of defining strata from multiple stratification variables were proposed, leading to this investigation. All stratification methods use stratification variables available for the entire frame. We reviewed textbook guidance on the general principles and desirable properties of stratification variables and the assumptions on which the two methods were based. Using principal components combined with cluster analysis on the stratification variables to define strata focuses on relationships among stratification variables. Decision trees, regressions, and correlation approaches focus more on relationships between the stratification variables and prior outcome data, which may be available for just a sample of units. Using both principal components/cluster analysis and decision trees, we stratified primary sampling units for the 2009 Residential Energy Consumption Survey and compared the resulting strata.

Introduction

The work presented here arose in the context of planning the 2015 Residential Energy Consumption Survey (RECS), where primary sampling units (PSUs) were to be stratified prior to selection. Although there are many ways to define strata for sampling, we focus here on two approaches that were proposed for RECS. Both proposals were legitimate, based on different perspectives on the following issues:

- Is it appropriate to use historical data, even if they are not available for all PSUs in the PSU frame?
- How desirable is having stratum definitions that are easily understood?
- Which method works better to improve the precision of estimation for key outcome variable(s)?

This work provides a short introduction to RECS to provide context for the stratification options proposed. We review some basic stratification concepts to build up to the two proposed stratification methods and the issues they raised. We present the results of a comparison test using both stratification methods with data from the 2009 RECS. We conclude by discussing the stratification results that indicated no clear winner between the two proposed stratification methods.

The Residential Energy Consumption Survey

The principal purpose of RECS is to provide Congress and the general public with 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 energy consumed as electricity, natural gas, bulk fuels (fuel oil, propane, and kerosene), and wood. 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.

In addition, 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.

Key Points

- Two methods of stratification were proposed for the primary sampling units of the Residential Energy Consumption Survey.
- Some basic goals and principles of stratification are reviewed.
- When the goal is to improve precision in estimation, stratification variables that capture much of the variability in the survey variables of interest are good choices for defining strata.
- With multiple stratification variables available, principal components combined with cluster analysis can reduce dimensionality and define strata.
- Alternatively, if historical survey data are also available, decision tree analysis can be used to define strata.
- Both methods created reasonable test strata for the Residential Energy Consumption Survey data, and neither method consistently outperformed the other in reducing the variance of estimated population totals. Ultimately, choosing a method is a matter of prioritizing the goals and assumptions.

Together with the Commercial Building Energy Consumption Survey and the Manufacturing Energy Consumption Survey, RECS completes the trio of surveys conducted by the US Energy Information Administration (EIA) to describe the consumption of energy within the overall US economy. More information on RECS is available at the EIA Web site (http://www.eia.gov/consumption/residential).

The target population for the 2015 RECS is all housing units that are primary housing units, defined as being occupied by a household at least half of the year, in the 50 states and the District of Columbia. Vacation homes, seasonal housing units, and group quarters, such as dormitories, nursing homes, prisons, and military barracks are excluded from the study; however, housing units 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 housing units 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 (number of occupied housing units in the 2013 1-year American Community Survey estimates) and with minimal replacement (Chromy, 1979). At the second stage, four secondary sampling units (SSUs) defined by census block groups were selected within each sampled PSU with probability proportional to size (occupied housing units) and with minimal replacement. For the third stage of selection, housing unit sampling frames were constructed, and a systematic sample of housing units was selected from the housing unit sampling frame in each segment.

The 2015 RECS is required to produce estimates for the nation, 19 geographic domains (subdivisions of Census Divisions), and 4 large states, with a specified level of precision. The goal of the stratification on RECS is to help reduce the variability in the estimates for these geographical areas while controlling the costs.

Given the estimation goals, the first-level stratification of PSUs partitions the United States into the 19 geographic domains, as was done for the 2009 RECS. This paper focuses on the stratification of PSUs (PUMAs) within geographic domains for RECS.

Fundamentals of Stratification

Textbooks of survey sampling define *stratification* as the process of partitioning the target population into separate subpopulations for sampling, called *strata*, 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.

Goals of Stratification

Stratified designs are popular for several reasons:

- Stratification usually results in more precise estimates (lower variance) for key variables of interest in the target population.
- Estimates may be desired for domains (subpopulations of interest) as well as the total population, and strata that correspond to domains, at least approximately, may be oversampled.
- Sampling less expensive strata at a higher rate can help reduce overall costs, with reductions in precision.
- The strata may be logical subdivisions for managing or equalizing data collection workload.
- Different sampling or data collection strategies may be appropriate for different subpopulations, leading to subpopulation strata.
- Different frames may be available for different subpopulations, for subpopulation strata.
- Stratified designs help protect against the possibility of an unrepresentative random sample.

Stratified designs are very common, and the reasons for stratification are study specific. For the RECS, stratification beyond geographic domains reduces variance in estimates so that precision requirements can be met for national and geographical domains with smaller sample sizes (lower costs) than for less complex designs.

Process of Stratification

Stratification often follows a general process that includes these steps:

1. Identify goals of the stratification.
2. Identify relevant stratification variables.
3. Clarify assumptions regarding stratification variables and variables of interest.
4. Look for combinations and transformations of the stratification variables to reduce dimensionality.
5. Define useful subdivisions of the stratification variable values.
6. Determine the desired number of completed cases from each stratum (allocation) and select enough sample from each stratum to achieve the desired allocation.

Having defined the goals for RECS stratification above (step 1), this paper focuses on steps 2–5. This paper does not cover Step 6 for sample allocation and selection after the strata are defined. We focus primarily on a single *y* for RECS, total energy consumption. Additional *y* variables were used only for evaluation purposes.

Step 2

For precision goals in estimation, members of a stratum should be as much alike as possible, and the strata should be as different as possible. The stratification variables *x* used to define the strata should be highly correlated with the outcome variable *y*. When the goal is to reduce the variance of estimates of the outcome variable, the main assumption of using stratification variables is that the variation in the *x* variable(s) captures a large portion of the variation in the *y* variable. 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.[1] That is, the variance of the estimated total is the sum of the within-stratum variances.

Stratification variables *x* may be 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). Potential stratification variables for RECS are shown in the application section below.

A general principle is that the stratification variable(s) must be available for every unit in the frame; although methods exist for dealing with exceptions, we do not cover them here.

When the goal of stratification is to estimate domains or to oversample some domains, then domain identifiers are natural choices for stratification variables. For RECS, because estimates of energy consumption are desired for geographic domains, the RECS domain is a natural stratification variable. If estimates for two sets of crossed domains are desired, such as census division and urban versus rural, both classification variables can be used to define strata.

Subject matter experts are a great source of information about potential stratification variables. They know what variables are available, what variables are correlated, and what variables have been used in the past. This 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 previous survey, as was the case for RECS, it is useful to compute simple correlations with potential stratification variables to better understand the variables' relevance.

Step 3

Whether to use historical *y* variables (from a prior survey) or alternative *y* variables (from another source) as stratification variables is a philosophical issue closely tied to the assumptions one is willing to make. We considered this issue at the heart of our investigation for the RECS study presented in this paper.

Generally, historical *y* values are not available for all units in the frame or are available only in aggregate, as is the case for alternative *y* variables from other sources. This is essentially a problem of missing values in the stratification variables. A common approach is to impute the missing values in some way. In the case of missing values in historical *y* variables, some other potential stratification variables can be used to predict the historical $\hat{y}$ for all units in the frame, and $\hat{y}$ can then be used as a stratification variable (J. Eltinge, personal communication, July 15, 2015).

Even if historical *y* values are not used to define strata directly, they can nevertheless be used to estimate correlations or model relationships with other potential stratification variables. When historical *y* values are used in some way, the assumption is that the correlations between the stratification variables and the historical *y* variable(s) are strong predictors

[1] Godfrey et al. (1984) argue that using ratio or calibration estimation also provides the benefits of using *x* in estimating $\hat{y}$. Consequently, the stratification should focus on the distribution of the residuals when modeling y in terms of *x*, not on the distribution of *x* itself. More often, however, stratification is done assuming a Horvitz-Thompson estimator will be employed. We make that assumption here.

of the correlations between the stratification variables and the current *y* variables to be collected in the survey.

Steps 4 and 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 can be 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 need to be explained to funders.

If the potential stratification variables 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 partitions of the frame based on the stratification variables. The goal of variable reduction is to capture most of the variability in the *y* variable(s) with fewer stratifying variables (step 4). The goal of level reduction is to consolidate ranges of the stratification variable values to reduce the possible number of strata while retaining the most meaningful partitions (step 5). These steps may be combined.

Principal Components and Cluster Analysis

Principal component analysis is a well-known variable reduction technique suitable for finding the best combinations of variables to use as stratification variables. The first principal component can be the sole *x* variable, or the first two or three principal components can be used to capture multiple dimensions of the variability. The principal components are linear combinations and transformations of the potential *x* stratification variables. Principal components transform the variables to be orthogonal; there is no multicollinearity, but the interpretation of the principal components is not necessarily clear. The focus of principal components is on the relationships among the stratification variables themselves, 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), even though the *y* variable is not involved in the stratification in any way.

Cluster analysis can be used to categorize the frame based on the values of the stratification variables, providing scientific justification for the strata. The US Census Bureau uses the Friedman-Rubin clustering algorithm (Friedman & Rubin, 1967) for stratification in 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 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, however, this is not an issue.

One of the methods proposed for RECS was to use principal component analysis followed by cluster analysis on the principal components.

Decision Tree Method

Alternatively, when historical *y* values are available, correlations, regression analysis, and decision trees can help identify the most relevant stratification variables. Variables with little or no correlation with the *y* variable are unlikely to be useful stratification variables. 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. That is, when multiple variables can be used to explain the variability in the outcome variable(s), the interactions and multicollinearity make it difficult to determine the best choices for partitioning the frame. Complete cross-classifications of all such stratification variables can lead to too many strata to be practical—small frame sizes and sample allocations that cannot be met. Decision trees reduce multicollinearity by sequentially identifying *x* variables with the strongest interaction with *y* to split the tree. Decision trees have the added advantage of determining likely subdivisions of the stratification variable values to best explain the variation in the *y* variable(s). That is, the terminal nodes of the tree form the strata. The number of nodes can be prespecified, if desired, or 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 prespecifying the number of nodes.

Using the terminal nodes of decision trees as strata combines 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 second proposal for RECS was to use decision trees with historical *y* data modeled by the stratification variables, and use the nodes as strata.

Optimization

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.

Number of Strata

Sometimes, the stratum boundary methods themselves help inform the number of strata that should be used. Otherwise, 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%), useful gain from more than six strata (1977) is unlikely—or, as Lohr (1999, p. 110) noted, "The less information, the fewer strata you should use." For RECS, the number of strata within a domain was approximately five.

Application and Test of Stratification Steps for RECS

Step 1

For this step, the purpose of stratification beyond geographic domains was to improve precision while containing costs. Below, we summarize the remaining steps for stratification of PSUs, following the two proposed methods of selecting the stratification variables and determining the stratum boundaries. We tested both stratification methods using 2009 RECS outcome variables.

Step 2

Within each geographical domain, energy consumption varies among PSUs. For example, a PSU with a high proportion of detached single-family housing units has higher energy consumption than a PSU with a low proportion of detached single-family housing units. Associating PSU characteristics to energy consumption helps to divide PSUs into groups based on 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.

The 10 characteristics are listed below:

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 previous 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 housing units to strata defined in the two proposed ways. One method used the historical (2009) 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 was used to determine 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.

Step 3

Both methods assumed that stratification capturing much of the variability in the *x* variables also captures much of the variability in the current *y* variables. Using the historical *y* value for total energy consumption in the decision tree method also assumed that strong relationships between the current *x* variables and the historical *y* variable indicated that relationships with the current *y* variable would be strong—even in PUMAs for which we had no 2009 RECS observations. Both methods assumed that the 2009 RECS sample design was ignorable for developing and testing the strata for the 2015 RECS.

Steps 4 and 5

Decision Tree Method

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. Compared with regression models, it is not susceptible to collinearity among independent variables because one variable is selected to split the tree at a time (Sambandam, 2003). CHAID classified housing units into groups (terminal nodes) based on 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 were 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.

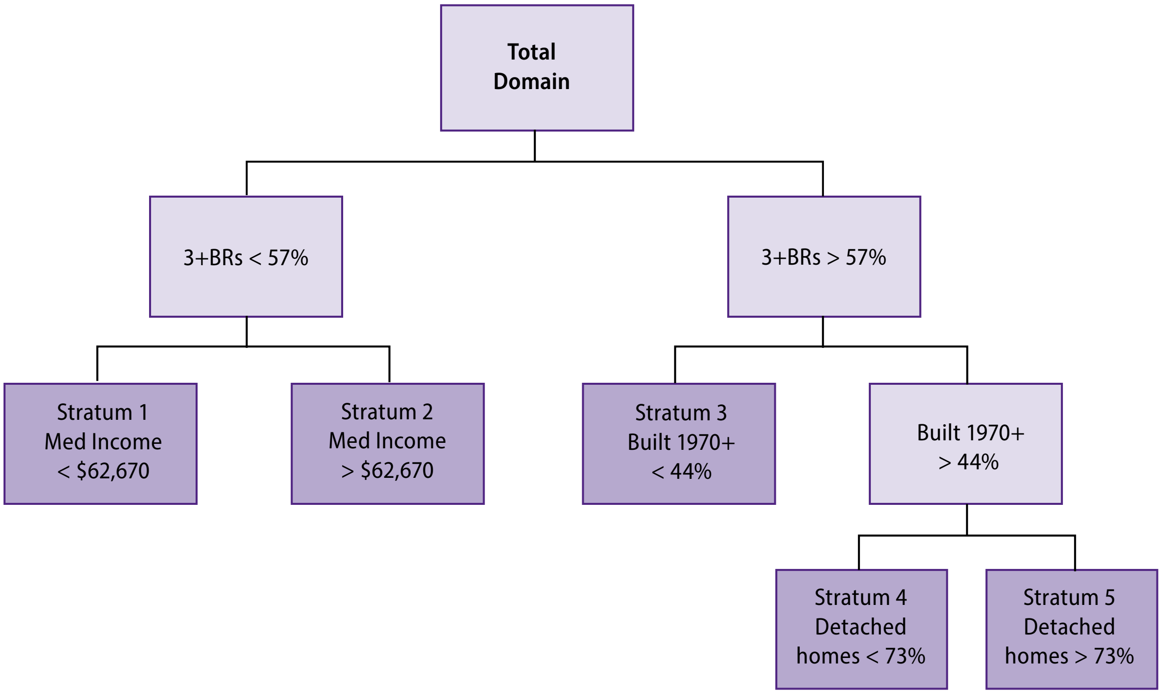
Table 1. Loading factors for principal components in Domain 1

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	Environment	Building size/type	Building age

Principal Components and Cluster Analysis

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

Figure 1. Decision tree and node strata for Domain 1 (New England)


Total Domain
3+BRs < 57%
3+BRs > 57%
Stratum 1 Med Income < $62,670
Stratum 2 Med Income > $62,670
Stratum 3 Built 1970+ < 44%
Built 1970+ > 44%
Stratum 4 Detached homes < 73%
Stratum 5 Detached homes > 73%

Cluster analysis using Ward's minimum variance method (Ward, 1963) was applied to the principal components that had eigenvalues 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 for an apples-to-apples comparison. 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

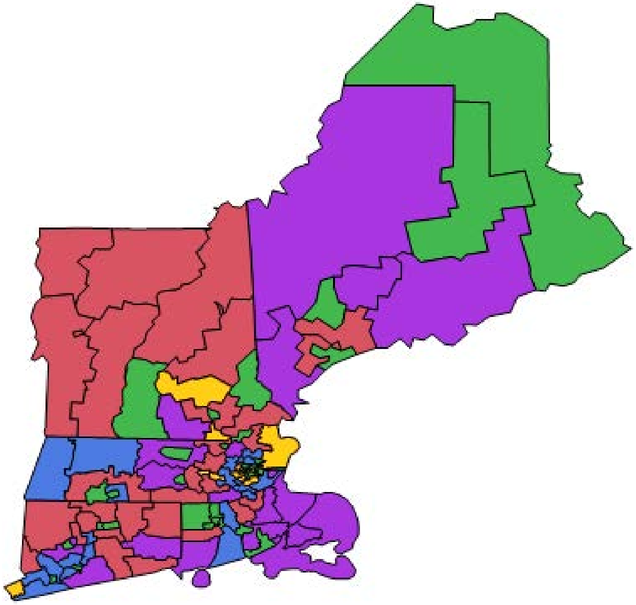
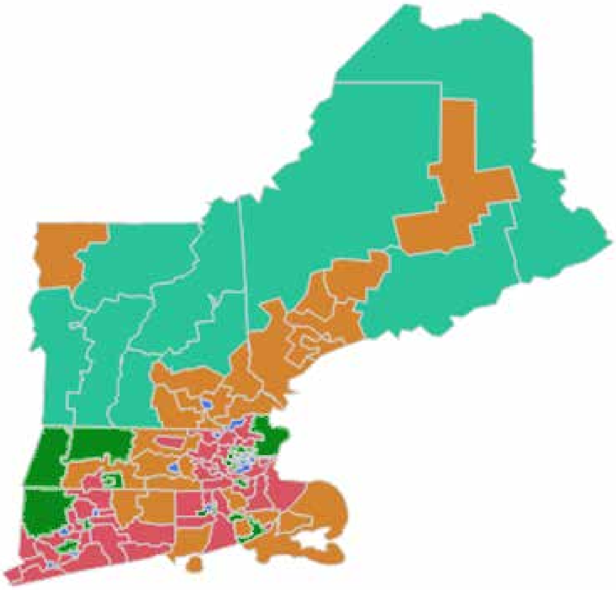


Figure 3. Cluster analysis strata



The Test

For the RECS test, both methods were used to stratify the frame of PSUs. Using sample observations (housing units) from the 2009 RECS as if they had been selected with these two stratifications, the variances of estimates were tested on three other *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 variance of the stratified estimated total is the sum of the stratum variances, so only the within-stratum component contributes to the variance of the estimated total. Therefore, the stratification that leads to a smaller within-stratum variance and a larger between-strata variance across variables and domains does a better job of capturing the variability in the *y* variables and leads to better precision overall.

Table 2 shows the ratio of the decision tree within-stratum variance to the corresponding cluster analysis variance. That is, we calculated the within-stratum variance of the *y* variables using the 2009 sample data and both methods of stratification. Then we computed the ratio of the within-stratum variance for the decision tree strata divided by the within-stratum variance for the cluster analysis strata. Because the outcome variables had vastly different variances, the ratios effectively scaled the variances, making comparisons easier. If the within-stratum ratio was less than 1, then the decision tree method was better at creating homogeneous strata for that outcome variable; if the ratio was greater than 1, then the cluster analysis method was better. The table shows that neither method was uniformly better across outcome variables. The decision tree method tended to define more homogeneous strata for electricity consumption and energy dollars, but the cluster analysis method was better for natural gas.

Because the ratios were all near 1, both the decision tree approach and cluster analysis approach seemed to be reasonable options for the 2015 RECS. If ratios had differed more widely from 1, we might have favored one method over the other for supporting the precision requirements with a smaller sample.

Table 2. Ratios of within-stratum variance components: decision tree strata to cluster strata

Variables	Within-stratum variance ratio*
Domain 1 (New England)	
Total electricity usage (thousand BTU)	0.96
Total natural gas usage (hundred cubic feet)	1.10
Total energy cost (dollars)	**0.93**
Domain 2 (New York)	
Total electricity usage (thousand BTU)	0.98
Total natural gas usage (hundred cubic feet)	1.01
Total energy cost (dollars)	**0.99**
Domain 16 (California)	
Total electricity usage (thousand BTU)	0.98
Total natural gas usage (hundred cubic feet)	1.01
Total energy cost (dollars)	**0.98**

*Ratios < 1 favor the decision tree strata; ratios > 1 favor the cluster strata.

Discussion

Most surveys collect data on more than one outcome variable *y*, although designers often focus on one or two of the most important outcome variables either for simplicity or because the precision requirements are expressed in terms of one or two key outcome estimates. Cochran (1977) reviewed a few methods for stratifying for multiple *y* variables. For example, if a small number of *y* variables is considered key, the stratification can be determined for each variable independently, and often a compromise stratification can be obtained that meets most of the stratification goals. For the RECS test, the decision tree method was optimized for the 2009 total energy consumption *y* variable because total energy consumption was deemed the most important outcome variable for the 2015 RECS. Note that the principal components/cluster analysis approach does not directly depend on any *y*. Testing the strata on the 2009 total energy consumption—because we did not yet have 2015 total energy consumption—naturally would have favored the decision tree strata. For this reason, the strata were tested on three other *y* variables.

The RECS test compared two methods of stratification that were based on different assumptions about the value of using historical *y* variables. The *y* variable from a previous time period was missing for some PSUs in the frame, which ordinarily would exclude the historical *y* as a stratification variable in the usual sense. There was no guarantee that the usefulness of the historical *y* in PSUs sampled for the 2009 RECS carried over to unsampled PSUs, and no way to test the relationships in unsampled PSUs. On the other hand, historical data allowed for examination of relationships between stratification variables and the historical outcome variable in the PSUs with sample. Furthermore, the use of the historical outcome variable supported use of model-based methods, such as decision trees, to define the strata.

The decision tree method that used historical consumption data performed better for total electricity consumption and total energy costs, which are highly correlated with total consumption. Natural gas has a much smaller correlation with total consumption, which may explain why the decision tree based on total consumption did not work as well for this variable.

It would be instructive to evaluate the two stratification methods on several more outcome variables, and to calculate the correlations among the variables to see if the correlations explain which variables are better with and without historical total energy consumption in the stratification process. Conversely, the decision tree process could be repeated using other historical outcomes such as total electricity or total natural gas in the decision tree to see the impact on estimation of total energy consumption.

Changes in the energy industry might alter the assumed relationships over time. For example, with the drop in natural gas prices in recent years, the relationships between various fuels and total consumption may have shifted, changing the preferred stratum definitions. With two cycles of data available for testing, the first cycle could be used to determine strata on the second cycle, and the estimates could be calculated on the second cycle to see how well the relationships hold up over time. That is, the within-stratum variance ratios for the two methods could be tested on 2015 outcome variables.

For ease of explaining the stratum definitions, the decision tree approach can be clearly described or diagramed, as in Figure 1. The clusters constructed from principal components are more challenging to explain because of the variable transformations in the principal components and the fuzzy cluster boundaries.

Some aspects of stratification were glossed over in this presentation, and the test conducted for RECS was inconclusive for reducing variance; even so, the test was useful because it demonstrated that neither stratification was consistently better than the other for RECS, and we need not spend more time on the decision. For the 2015 RECS within-domain stratification of PUMAs at the PSU selection stage, the decision tree approach was used. The correlations among the variables and other assumed relationships influenced the decision about whether to use historical outcome data in defining strata. For RECS—with total energy consumption as the primary outcome variable, with expected high correlation between current total energy consumption and both prior consumption and the stratification variables, and with the greater ease of articulating the decision tree stratum definitions—the decision tree approach was selected over the cluster analysis method, even though the historical outcome variable was not available for every PSU in the frame.

Many studies have neither the time nor the resources for extensive work on stratification. Stratification variables 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, but rather to find reasonably good stratification that will support the analytical and operational goals better than simpler designs. Stratification need not be "best" or "perfect" to perform well (Lavrakas, 2008).

Although the advantages of one stratification method over another may be minor, the mixed results of the RECS test illustrate the need to think about the goals of stratification, the key outcome variables, and their relationships with the stratification variables. That is, steps 1–3 in the stratification process deserve due consideration. Determining useful potential stratification variables is more important than the specific way in which the stratification variables are used.

References

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Chromy, J. R. (1979). Sequential sample selection methods. In *JSM Proceedings, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 401–406.

Dalenius, T. (1950). The problem of optimum stratification. *Scandinavian Actuarial Journal, 1950*(3–4), 203–213. https://doi.org/10.1080/03461238.1950.10432042

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

Friedman, H. P., & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association, 62*(320), 1159–1178. https://doi.org/10.1080/01621459.1967.10500923

Godfrey, J., Roshwalb, A., & Wright, R. L. (1984). Model-based stratification in inventory cost estimation. *Journal of Business & Economic Statistics, 2*(1), 1–19. https://doi.org/10.1080/07350015.1984.10509365

Judkins, D. R., & Singh, R. P. (1981). Using clustering algorithms to stratify primary sampling units. In *JSM Proceedings, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 279–284.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 29*(2), 119–127. https://doi.org/10.2307/2986296

Lavrakas, P. J. (Ed.). (2008). *Encyclopedia of survey research methods*. Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781412963947

Lavallée, P., & 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., & Reist, B. M. (2010). Evaluating alternative criteria for primary sampling units stratification. In *JSM Proceedings, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 4664–4672.

Sambandam, R. (2003). *Cluster analysis gets complicated* [White paper]. Retrieved August 22, 2016, from 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. https://doi.org/10.1080/01621459.1963.10500845

www.rti.org/rtipress **RTI Press publication OP-0041-1705**