A Comparison of Clustering Algorithms used for Multivariate Stratification of Primary Sampling Units[*]

Padraic Murphy[†] and John Chesnut[†]

Abstract
For many demographic surveys it conducts, the Census Bureau uses a two-stage sample design, where the Primary Sampling Units, or PSUs, are counties or groups of counties and the second-stage sampling units are households selected from within the sampled PSUs. To reduce sampling variance in the first stage, we stratify the PSUs. Beginning with the 1980 sample redesign, the Census Bureau has used a method of PSU stratification based on a clustering algorithm described in a 1967 article by Friedman and Rubin. The Friedman-Rubin algorithm has been described as a "greedy hill-climbing heuristic." Alternatively, in the 2010 sample redesign, the Consumer Expenditure Survey used an approach based on k-means clustering and an iterative application of constrained integer programming optimization. We refer to this alternative approach as the "King method" after the primary author of an article in the proceedings of the 2011 Joint Statistical Meetings describing the approach. This paper attempts to compare the two methods by creating stratifications with each method for five different surveys and comparing the results using two alternative evaluation metrics.

Key Words: Clustering, integer programming, greedy algorithm, multi-stage sampling, stratification

1. Introduction

It is well established that stratification of the units in a sampling frame will result in more efficient estimation (e.g., Cochran, 1977), if the strata are defined such that units within a stratum are similar with respect to the quantity of interest. This is true for both stages in a two-stage sample, but we focus here on the first stage - Primary Sampling Unit (PSU) stratification. If one is estimating for a single characteristic, finding an optimal stratification is relatively straightforward. However, most surveys produce estimates for multiple characteristics. An additional complication is that we generally must use proxy variables for the characteristics of interest - for example, if we are estimating current levels of unemployment, a proxy variable could be the most recent historical annual unemployment estimates from the American Community Survey.

Before 1980, those working on sample design for Census Bureau demographic surveys did not expend a great deal of effort on optimizing PSU stratification. For most published estimates, the first-stage contribution to variance was relatively small, so it was satisfactory to use stratifications that were "good enough." However, increased emphasis on estimates for smaller geographies, or for minority subgroups, raised concerns about the first-stage contribution to variance, which could be quite large in those cases (Kostanich, Judkins, Singh, & Schautz, 1981). In the early 1980's, the Census Bureau began working on a systematic approach to PSU

[*] This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[†] U.S. Census Bureau, Washington, DC 20233

stratification, based on what we shall call the Friedman-Rubin clustering algorithm. Development of systems based on Friedman-Rubin continue to the present day, with improved versions being implemented with each sample redesign (1990, 2000, and 2010). The most recent version is a system called the PSU Stratification Program (PSP). We will explain the functioning of the PSP in more detail later in the article.

For the 2010 sample redesign, the Bureau of Labor Statistics (BLS) proposed an alternative approach to PSU stratification, which we shall call the King method, after the primary author of the article in which the method is explained (King, Schilp, & Bergmann, 2011). We will describe the details of the King method later in the article. Based on their research, the BLS chose to use the King method to stratify PSUs for the Consumer Expenditure Surveys (CEX). For their research, they stratified CEX PSUs within each of the four Census regions (Northeast, Midwest, South, and West) using a Friedman-Rubin approach, and compared the results with stratifications produced using the King method. They chose to use a measure of within-stratum homogeneity called trace($\boldsymbol{W}$) - explained later in this article - and found that the King method stratifications had better (lower) trace($\boldsymbol{W}$) values than their Friedman-Rubin stratifications for each region.

While the PSP appears to produce reasonably good PSU stratifications, the 2011 King paper has raised the possibility that it might be possible to find better stratifications, and possibly do so more quickly, at least for areas where the number of PSUs and strata are not too large. The purpose of this research is to compare the PSP with the King method, using the 2010 sample redesign PSU stratification input data for several surveys in addition to CEX, and to evaluate whether the Census Bureau should keep the PSP as its primary PSU stratification tool, replace it with the King method, or perhaps use a combination of the two.

2. Background

2.1 Review of Two Stage Sampling

Most of our demographic surveys usually need state-level estimates in addition to nation-level estimates. In addition, our surveys require in-person data collection. As a result, if we were to select a simple random sample of housing units, the sample would be too dispersed geographically to hire and train enough interviewers to interview all sampled households in a cost efficient manner. Therefore, we delineate large clusters of households that will be the sampling units for the first stage of a two-stage sample, and then in the second stage select the households to be interviewed from within the selected first-stage units. Usually, the first-stage units are single counties (or county equivalents) or groups of geographically contiguous counties. By convention, we call such a first-stage unit a PSU.

To reduce first-stage sampling variance, we stratify the PSUs before selecting the first-stage sample. A special class of PSU's, generally those containing large metropolitan areas, are required to be included in a survey's first-stage sample with certainty. By convention, these are labeled self-representing, abbreviated SR, and each SR PSU is placed in a stratum by itself The remaining PSU's, and the strata that contain them, we label as non-SR, (NSR). Each NSR stratum should contain more PSUs than the number to be selected (otherwise the stratum would be a de facto SR stratum).

In general, we delineate NSR PSUs in such a way that it is practical for one field representative to conduct all interviews for a survey in that PSU. We do this by requiring that a NSR PSU have less than a certain maximum land area, and at least a certain minimum population. By contrast, there are no area or population restrictions for an SR PSU, and the number of field interviewers needed for an SR PSU is dictated by the size of the second-stage sample selected within the SR PSU.

The number of NSR PSUs to select is a decision influenced by a number of factors, including cost, desired precision of estimates, and operational considerations. We will not go into detail about how this is done in practice. Assume that the NSR PSU sample size is given. We could then select a simple random sample of NSR PSUs, and in some cases, that might be a reasonable choice. Usually, though, it makes sense to stratify the NSR PSUs. Furthermore, it turns out that selecting only one or two PSUs per stratum tends to yield the most precise estimates. There is some debate over which is better - one PSU per stratum or two - but we simply note that of the five surveys discussed for this article, four choose to use a one-PSU-per-stratum design, while the remaining survey uses a two-PSU-per-stratum design.

2.2 The Stratification of NSR PSUs - A Constrained Clustering Problem

One may view stratification of NSR PSUs in the first stage of two-stage sampling as a clustering problem. We are trying to form clusters of NSR PSUs that are similar with respect to the characteristics of interest for the survey. However, there are two constraints that we impose that make PSU stratification different from the classic clustering problem. The first is that since we are selecting a sample of one or two NSR PSUs to represent other non-selected NSR PSUs, we must have at least one more PSU in each stratum than the number we are selecting. (So if we are selecting one PSU per stratum, there must be at least two PSUs in each NSR stratum; there must be at least three if we are selecting two per stratum). The other is not immediately obvious: we require that the NSR PSU strata be approximately the same size, in terms of some appropriate measure. Often, the PSU measure of size is population; it could also be number of housing units or addresses, or population age 16+, etc. The reason for this second requirement has to do with the variance of sample estimates. If the sizes of PSU strata are very different, and we are selecting PSUs with probability proportional to size, the stratum weights will be very different. This will cause our estimates to be less reliable than if the strata are similar in size. Therefore, when forming NSR PSU strata, we require that (for example) the size of each stratum deviate no more than ten percent from the average stratum size.

We point out this difference between natural clustering problems and PSU stratification (a constrained clustering problem) because it means that the algorithms and computer software developed to address the natural clustering problems cannot be used to solve the constrained clustering problem, at least not directly. As we shall see, however, an algorithm for natural clustering may prove to be a good starting point for constrained clustering.

2.3 How Many Possible PSU Stratifications Exist?

To answer this question, we borrow from the field of combinatorics the concept of an "r-associated Stirling Number of the second type" (Broder, 1984). By definition, the r-associated Stirling Number $S_r(N, L)$ is the number of ways of partitioning N objects into L groups such that every group contains at least r objects. For our purposes, we are concerned with the cases where $r = 2$ or 3. It should be clear that the following are true for both of these values of r and any whole number values of N and L, assuming $N \geq L \geq 1$:

$$S_r(N, L) = \begin{cases} 0 & \text{if } N < rL \\ 1 & \text{otherwise} \end{cases}$$

It can be shown (Broder, 1984) that the following recursion formula is true for $L > 1$:

$$S_r(N+1, L) = L \times S_r(N, L) + \binom{N}{r-1} \times S_r(N-r+1, L-1)$$

For instance:

$$S_2(5, 2) = 2 \times S_2(4, 2) + \binom{4}{1} \times S_2(3, 1) = 2 \times 3 + 4 \times 1 = 10$$

So there are 10 ways of partitioning 5 objects into 2 groups of at least 2. Programming this recursion formula is straightforward, and one may calculate the number of possible stratifications given any combination of N, L, and r. To illustrate, Table 1 shows the number of possible stratifications for selected surveys and stratification areas given their 2010 counts of NSR PSUs and strata. While some of the search space is taken out of consideration by the stratum size constraints, the size of the remaining space is still of a similar order of magnitude. One can see that enumerating and evaluating all possible stratifications is impractical.

Table 1 Number of Possible Stratifications for Selected Surveys by Stratification Area

			2010	2010	Number of
		Min	NSR	NSR	Possible
	Stratification	PSU	PSU	Strata	Stratifica-
Survey	Area	Count	Count	Count	tions
Current Pop. Survey	Lousiana	2	29	8	1.329×10^{21}
Current Pop. Survey	Penn.	2	30	9	2.244×10^{22}
National Crime Vict. Survey	Florida	2	36	9	2.680×10^{28}
Survey of Income and Program Participation	Oregon	3	24	3	4.638×10^{10}

2.4 Development of the PSU Stratification Program (PSP)

In 1967, Friedman and Rubin published an article about cluster analysis proposing a systematic heuristic procedure for searching for the best partition of n objects into g groups. The article also discusses a number of "best" mathematical criteria and relates these to statistical theory (Friedman & Rubin, 1967). In 1981, a group at the Census Bureau proposed using a modified version of the procedure described by Friedman and Rubin to stratify PSUs for the Current Population Survey (CPS) (Kostanich et al., 1981). Building on the proposal by Kostanich et al., as well as other sources, the Census Bureau developed a computer program to search for size-constrained PSU stratifications; and this program was used for PSU stratification in the sample redesign based on the 1980 Census, for several surveys (Jewett & Judkins, 1988). Essentially the same program was used again for the 1990 sample redesign (Ludington, 1992).

During the 1990 sample redesign, a separate program was developed to address the issue of widely varying interviewer travel costs in Alaska that could not be taken into account using the PSU stratification program mentioned above (Ludington, 1992). Based on this Alaska PSU stratification program, the Census Bureau developed the Stratification Search Program (SSP) and used it to stratify PSUs in all states in the 2000 sample redesign (SSP_2K). A modified version of the SSP was used for the 2010 sample redesign, called the PSU Stratification Program (PSP_2010 or PSP).

2.5 The Pseudo-Assignment Algorithm for PSU Stratification - "the King Method"

In the field of operations research, there is a type of linear programming problem known as the Transportation Problem (TP). A special sub-type of the TP is the Assignment Problem, where one may think of supply nodes as workers or machines, and destination nodes as tasks or jobs. In a balanced assignment problem, all of the supply and demand quantities are equal to 1, so one is trying to find the best way of assigning (for example) tasks to workers in a one-to-one manner that minimizes cost (or maximizes productivity). If there are more tasks than workers, or vice versa, the problem is unbalanced, and the problem must be balanced by adding "dummy" nodes in order to use the efficient methods developed to solve Assignment Problems, such as the Hungarian Method (Winston, 1994).

If PSUs are viewed as "jobs" and PSU strata are viewed as "workers," one may attempt to set up the PSU stratification problem as an unbalanced assignment problem. Unfortunately, the size constraints imposed on strata do not fit the definition of a true assignment problem, and so the special algorithms for such problems are not applicable. However, it is still possible to set up the PSU stratification problem as a general Linear Programming problem (LP) - more specifically, a binary integer linear program (BIP) - if one can find a way to express the "cost" of assigning a PSU to a stratum. In working on PSU stratification for the 2010 sample redesign of the Consumer Price Index and Consumer Expenditure surveys, staff at the Bureau of Labor Statistics (BLS) devised a way of doing this as one step in an iterative process (King et al., 2011). They refer to the LP they set up as a "pseudo assignment problem."

In their 2011 article, King et al. describe how they used the King method to stratify NSR PSUs for the 2010 sample redesign of the Consumer Expenditure surveys in each of the four Census regions (Northeast, Midwest, South, and West) and compared the results with stratifications obtained using a Friedman-Rubin-type hill-climbing algorithm similar to that used by the Census Bureau's PSP. They found that the stratifications produced by the King method have lower trace($\boldsymbol{W}$) values than the stratifications from their Friedman-Rubin procedure for all four regions. There is no direct comparison of computation times mentioned in the paper, but they point out that the King method "uses commercially available software, is explainable, and fast." In any event, the BLS chose to use the King approach to stratify PSUs for the Consumer Price Index and the Consumer Expenditure surveys in the sample redesign based on the results of the 2010 Census. Note that the BLS eventually decided to stratify CEX NSR PSUs at the Census division level, rather than at the Census region level; they were able to successfully apply their PSU stratification method with this new set of parameters.

3. Methodology and Results

3.1 Description of the PSP and Modifications

The PSP, created for the 2010 sample redesign, is a suite of Linux scripts and SAS® programs designed for Census Bureau demographic surveys to stratify PSUs in the redesign of their samples following the 2010 Census. It was based on the PSU stratification system used for the 2000 Redesign called the Stratification Search Program (SSP). We will not attempt to describe the SSP and then explain the changes made to get to the PSP. Rather, we will simply describe what the PSP does, as built for 2010. We will, however, discuss some modifications we have made in the course of our research - the PSP results presented in this article come from our modified version of the 2010 PSP, not the one actually used in the 2010 sample redesign.

The PSP has two major components. The first component stratifies PSUs within specified sub-national areas. For the Current Population Survey (CPS), the National Crime Victimization Survey (NCVS), and the Survey of Income and Program Participation Survey (SIPP), these sub-national areas were states. For the American Housing Survey (AHS) and the Consumer Expenditure Survey (CEX), these areas were Census divisions. The second PSP component takes the results of the area PSU stratifications for all sub-national areas and combines them to produce national estimates of variance for the stratification variables and any evaluation variables designated by the PSP user.

Our research focuses on the first component of the PSP, stratification of PSUs within a sub-national area. Each such area is handled independently by the PSP area component. The PSP area component performs the following steps. The algorithm first creates a specified number of initial stratifications satisfying the stratum size constraint. Then it performs a series of hill climbing passes where all single PSU moves to a different stratum are considered that satisfy the size constraint, and the new stratification with the lowest criterion value is selected. The hill climbing passes are repeated until no improvement is observed for the minimizing criterion. Next a series of exchange passes are performed considering all exchanges of a PSU with another in a different stratum. This is also repeated until no improvement is observed. From all of the final stratifications created from the specified number of random starts, the one with the smallest criterion value is selected. The PSP is flexible in that the user can specify their preferred minimizing criterion.

During our research into the functioning of the PSP, we found some issues that if addressed could improve the PSU stratification results. In particular, the method of finding initial stratifications to start the PSP algorithm works, but often fails to find a stratification that meets size constraints, and frequently produces duplicate stratifications (in the sense that the groupings of PSUs are the same, even though the labeling of the strata may mask this). Further, our research indicates that the number of initial stratifications typically used (and the maximum allowed by the 2010 PSP) may be too small. We found that with a larger number of initial stratifications (50 to 100), the PSP was able to identify better stratifications that may have been missed with fewer initial starts.

To address the issue of initial stratifications, we created a modified version of the PSP that attempts to find 50 distinct stratifications that meet the size constraints to begin with (rather than using the "smoothing" algorithm described in 2010 PSP documentation). Note our emphasis on "distinct" is because two stratifications that have the same actual groupings but label the strata differently are not allowed in our framework (but were in the 2010 PSP). We ensure distinct stratifications by linking the labeling/numbering of the strata to the sorted order of the PSU identification codes. That is, each stratum may be identified with the PSU in the stratum that has the earliest PSU identification code. Note this means that the first stratum is always the one that has the first PSU.

3.2 Description of King Method, and modifications

To describe the King method, assume the following information is given:
N = the total number of PSUs to be stratified
L = the number of strata to be formed
R = the number of stratification variables
M_i = the measure of size for PSU i, for $i = 1$ to N
Y_{ij} = the value of stratification variable j for PSU i, for $i = 1$ to N, $j = 1$ to R
p = the allowed proportion each stratum measure of size may differ from the average

The first step in the King method is to standardize each stratification variable by

- Calculating the mean and standard deviation of the variable across all PSUs in the given area
- For each PSU, subtract the area mean from the PSU value, and divide this difference by the area standard deviation
- Assign the resulting values to an array (matrix) X_{ij}, so we have

$$X_{ij} = (Y_{ij} - \mu_{Y_j})/\sigma_{Y_j}$$

The next step in the King method is to use a k-means clustering algorithm (Clarke, Fokoué, & Zhang, 2009, pp. 409-411) to form an initial stratification. In SAS®, we do this using the FASTCLUS procedure. The strata (clusters) formed by k-means are not necessarily optimal. Furthermore, the algorithm does not provide a way to control the number of PSUs in each stratum or the sum of PSU sizes in a stratum. Therefore, the stratification produced by k-means is generally not a "good" PSU stratification. However, it does make a good starting point. Using the initial stratification formed with k-means, we next calculate the centroid for each stratum. Each centroid may be represented by a vector $\boldsymbol{K_h}$ in $\mathbb{R}^R$-space:

$\boldsymbol{K_h} = [k_{h1}, k_{h2}, ., k_{hR}]^T$ where $k_{hj} = (\sum_{i \in S_h} X_{ij})/N_h$

S_h= the set of PSUs in stratum h

N_h= the number of PSUs in stratum h

Let $c_{ih} = \sqrt{\sum_{j=1}^{R}(X_{ij} - k_{h_j})^2}$

That is, c_{ih} is the Euclidean distance between the (standardized) point in $\mathbb{R}^R$-space representing PSU i and the stratum centroid K_h. At this point, we define a linear integer programming problem, which King refers to as a pseudo-assignment problem, where the assignment problem is a sub-type of the transportation problem, following conventions in the field of operations research. We will not explain the origins of this terminology here; but the interested reader can find a good exposition of the transportation and assignment problems in any basic Operations Research textbook, such as (Winston, 1994, pp. 338-379). For each PSU, i=1 to N, and each stratum h,h=1 to L, define a binary variable u_{ih} such that

$$u_{ih} = \begin{cases} 1 & \text{PSU } i \text{ is in stratum } h \\ 0 & \text{otherwise} \end{cases}$$

Following conventions, the integer programming model may then be stated as:

Minimize $\sum_{i=1}^{N} \sum_{h=1}^{L} c_{ih} u_{ih}$

Subject to:

$\sum_{h=1}^{L} u_{ih} = 1$ for every i (a PSU is in exactly 1 stratum)
$\sum_{i=1}^{N}(M_i u_{ih}) \geq (1-p)((\sum_{i=1}^{N} M_i)/L)$ for every h (lower bound on stratum size)
$\sum_{i=1}^{N}(M_i u_{ih}) \leq (1+p)((\sum_{i=1}^{N} M_i)/L)$ for every h (upper bound on stratum size)

If we want to ensure that all NSR PSUs remain truly non-self-representing, and if there are any unusually large PSUs, we might also add the following set of constraints:

$\sum_{i=1}^{N} u_{ih} \geq B$ for every h where

$$B = \begin{cases} 2 & \text{if selecting 1 PSU per stratum} \\ 3 & \text{if selecting 2 PSUs per stratum} \end{cases}$$

In (King et al., 2011), they do not include this final set of constraints, but the set of lower bound size constraints likely make it redundant given the actual sizes of the PSUs they stratified in their example.

One then uses standard methods to solve this binary integer programming problem (e.g., the branch-and-bound method) as implemented by one's favorite software package. In SAS®, we use PROC OPTMODEL with the MILP (Mixed Integer Linear Program) solver. The solution consists of an optimal assignment of the binary decision variables u_{ih}. Here, u_{ih}=1 means that PSU i is assigned to stratum h.

Since the PSUs may now be assigned to different strata than in the original grouping produced by k-means, we re-calculate the stratum centroids. But now, the distances from PSUs to centroids are different, so the assignments may no longer minimize the objective function. Therefore, we solve the binary integer programming problem again, replacing the old values of c_{ih} with the new values. Keeping track of the objective function value, repeat this process until the absolute change in the objective function value from one iteration to the next is below some predetermined threshold. In rare instances, the stratifications in subsequent iterations may alternate between two groupings such that the absolute difference in objective function values remains at a constant greater than the threshold. To be safe, one should put in a check to stop iterating if the same absolute difference repeats more than two or three times. This was not mentioned in (King, Schilp, and Bergmann, 2011), probably because they did not encounter this odd situation.

3.3 Modifications to the King Algorithm

We modified the King method only with respect to the first step, generating an initial stratification, but we came up with two alternative ways of doing that. First, we generated multiple initial stratifications by doing 50 different random sorts of the PSUs before applying the k-means clustering since SAS documentation states that FASTCLUS can be sensitive to the initial ordering of records. It turns out that while we did get multiple distinct initial stratifications in some survey areas, the integer programming steps in the King algorithm modified all of the initial stratifications to the same final stratification. We speculate this is because randomly sorting PSU records resulted in only minor differences in FASTCLUS results (if there was any variation at all); and the constraints imposed in the integer programming step quickly removed those minor differences.

For our second modification, as an alternative to k-means initialization, it seemed natural to use the 50 size-constrained initial stratifications we created for our modified PSP as starting points for the King method integer programming step. Unlike the k-means initial stratifications, these did NOT all lead to the same final stratification. Furthermore, in most survey areas, the best of the final stratifications obtained this way were better than the single k-means result. We speculate that the size (and minimum PSU count) constraints result in more local optimum points for the objective function that are missed by starting with the unconstrained k-means procedure.

3.4 Metrics

The goal of PSU stratification is to find the most homogeneous groupings of PSUs possible given the constraints. Evaluating the overall level of within-stratum homogeneity across all of the strata requires that one define a metric to measure homogeneity. Many such metrics have been suggested in the literature, three of which are used in the 2010 PSP as criterion functions in the Friedman-Rubin approach. Note that the Mean Coefficient of Variation MCV criterion is based on the $BetVar$ function described in the 1981 article by Kostanich, et al. The MCV criterion option in the PSP allows the user to specify the relative influence of each stratification variable. This is done by specifying a vector of weights $\boldsymbol{f}$ that sum to one. Thus, the mean can be expressed as the dot product of the vector of weights and vector of CV values. In their 1967 article, Friedman and Rubin discuss trace($\boldsymbol{W}$), Hotelling's

Trace, and the ratio of the determinants of $\boldsymbol{T}$ and $\boldsymbol{W}$, where $\boldsymbol{T}$ is the total scatter matrix, and $\boldsymbol{W}$ is the within-clusters component of scatter. Friedman and Rubin express a slight preference for the ratio of determinants, for reasons one can read about in their paper, if interested. All surveys that used the PSP in the 2010 sample redesign chose to use the MCV criterion function.

For the purposes of this research, we chose to look at four different metrics, the three used as criterion functions in the PSP, and the ratio of determinants mentioned above.

- trace($\boldsymbol{W}$)
- $\mid \boldsymbol{W} \mid / \mid \boldsymbol{T} \mid$ (Ratio of Determinants)
- trace($\boldsymbol{WT^{-1}}$) (Korhonen, 1978)
- $MCV = \boldsymbol{CV} \cdot \boldsymbol{f}$ (Mean Coefficient of Variation)

We note here that only the MCV metric implicitly takes into account the different measures of size of the PSUs and the strata. The remaining three are all based on the scatter matrix decomposition (cf. Friedman & Rubin, 1967) and their formulas do not explicitly include the PSU measures of size. (Stratification variables could be calculated using PSU measure of size in some way, which would mean that the elements of the scatter matrix include PSU measure of size implicitly. However, in practice most stratification variables used by the surveys we discuss are PSU-level totals of some kind, without reference to PSU measure of size).

Further, note that while we used all four of these metrics in the course of our research, we end up focusing on only two of them, MCV and trace($\boldsymbol{W}$), when discussing results.

3.5 Scaling

When we began studying the PSP and the King method, one of the first differences we noticed is that the PSP does not have an option for scaling the stratification variables, while the King method automatically scales variables using the "standard" method (subtracting the mean and dividing by the standard deviation). We suspect that the lack of a scaling option in the PSP was an oversight; but since scaling is "built in" to the MCV function, and all the surveys using the PSP chose the MCV option, there was no impact on the 2010 PSU stratification results for those surveys. Still, if one chooses to use a different criterion function, it is important to consider that NOT scaling can result in differential influence among the stratification variables. For example, suppose that the unit for stratification variable A is "dollars" while the unit for B is "thousands of dollars." For a metric like trace($\boldsymbol{W}$), A will have much more influence than B, perhaps unintentionally. To correct for this, one may choose a scaling method that equalizes the influence across all stratification variables. In addition to the "standard" method used in the King approach, there are a variety of possible scaling methods one might choose. In their 1981 article, Kostanich et al. specifically mention "size proportional scaling" and "equal size scaling" which calculate scaling factors based on the relative variances of the variables under different sets of assumptions about the PSU measures of size. For this research, we also looked at "unit scaling" which transforms all variables to have values in the (0,1) interval.

Once variables are scaled to have equal influence, the user may at that point choose to give some stratification variables more weight. In the PSP, in fact, the user is required to provide preference factors for the stratification variables; but the PSP only uses the preference factors with the MCV criterion function option. Again, this may simply be an oversight, and an artifact of the decision by all surveys involved to use the MCV criterion function for their final stratifications. (It may also be the case that some surveys attempted to use the other criterion function options, but found the results difficult to interpret and therefore chose to go with the option that gave more sensible results. We speculate that including scaling options and applying preference factors with the non-MCV criterion functions in the PSP might have resulted in some surveys choosing one of the other criterion functions).

3.6 Comparison of Methods

Considering all of the possible combinations of variable scaling and criterion functions we looked at, there are a total of 20 variants of our modified PSP we might attempt. However, since the MCV is dimensionless, scaling does not change MCV results as long as the set of values in question all remain non-negative. In fact, if changing the scale of variables results in changes in MCV, this indicates the MCV is not meaningful. Therefore, we only apply MCV with unscaled stratification variables. (If the original set of values for any stratification variable includes both positive and negative values, it is not appropriate to use the MCV at all - but most surveys used variables with exclusively nonnegative values). We believe the same logic applies to the Korhonen criterion function trace($\boldsymbol{WT^{-1}}$). The trace($\boldsymbol{W}$) criterion function, on the other hand, is not dimensionless, and scaling does have an impact. The scaling method that seems to make the most sense in this case is standard scaling. Preliminary analysis using the 20 variants mentioned above empirically supports these conclusions; and it also showed that the two other criterion functions we looked at - the ratio of determinants of the T and W matrices, and Hotelling's trace - are redundant. For the King method, there is only one scaling option (standard) and one criterion (trace($\boldsymbol{W}$)). However, we do have the option of initializing k-means clustering without size constraints or using the random set of size-constrained initial stratifications created for our modified PSP. Note that the PSP always uses random initial stratifications.

To summarize, there are five methods we compare in this section: three using PSP variants, and two using the King method with different initialization options:

1. PSP, no scaling, MCV criterion function

2. PSP, no scaling, Korhonen criterion function

3. PSP, standard scaling, trace($\boldsymbol{W}$) criterion function

4. King with random initial stratifications, standard scaling

5. King with k-means initial stratifications, standard scaling

Also, we have two types of comparison. The first uses the MCV criterion function as an evaluation metric. The second uses the trace($\boldsymbol{W}$) criterion function, with standard scaling. Note that either type of metric can be applied to any stratification, regardless of what method was used to stratify.

For a given survey and area (state or Census division), we first apply each of the five PSU stratification methods in the list above. To compare the methods, we pick a measure of within-PSU-stratum homogeneity to be our evaluation metric. We evaluate each of the five final stratifications using this evaluation metric, and rank the methods accordingly. We then repeat the comparison/ranking with the other evaluation metric.

3.7 Data

For the 2010 sample redesign, each survey selected a set of stratification variables and corresponding datasources. Our study uses the same stratification variables and data. Table 2 lists the datasources selected by each survey.

Table 2 Data Sources for the Stratification Variables for each Survey

Data	Surveys
2000 Census Short/Long Form Summary Files	SIPP
2010 Census Summary File	AHS, CPS, NCVS, SIPP
2010 ACS 5-Year Housing Unit/Person Files	AHS, CPS, NCVS, SIPP
2007 ACS 3-Year Housing Unit/Person Files	CEX
2010 Urban-Rural Block Level File	AHS, CPS, NCVS, SIPP
2010 BLS Census of Employment and Wages - County by Industry Division Code Level File	CPS
2009 Uniform Crime Reporting County Level File	NCVS
U.S. Census Bureau's TIGER/Line® Shapefiles	CEX

3.8 Using MCV as an Evaluation Metric

To begin our discussion of the MCV-based comparison of methods, we look at results for CPS, Alabama. We applied each of our five methods to the set of CPS stratification variables used to stratify the NSR in Alabama in the sample redesign based on the 2010 Census. In the actual redesign, planners tried a range of stratum counts; we just use the number of strata they settled on. For CPS Alabama, 44 NSR PSUs were partitioned into 10 strata.

Figure 1 shows the distribution of MCV metric values for each of the five methods, using box-plots. Note that all methods started (for this particular survey stratification area - CPS, Alabama) with 50 distinct initial stratifications. All three PSP methods ended with 50 distinct final stratifications, as well. For the King Random method, nine of the final stratifications duplicated other final stratifications, so only 41 distinct stratifications remain. For the King k-means method, as happened with all survey stratification areas, all of the initial stratifications collapsed to one final stratification.

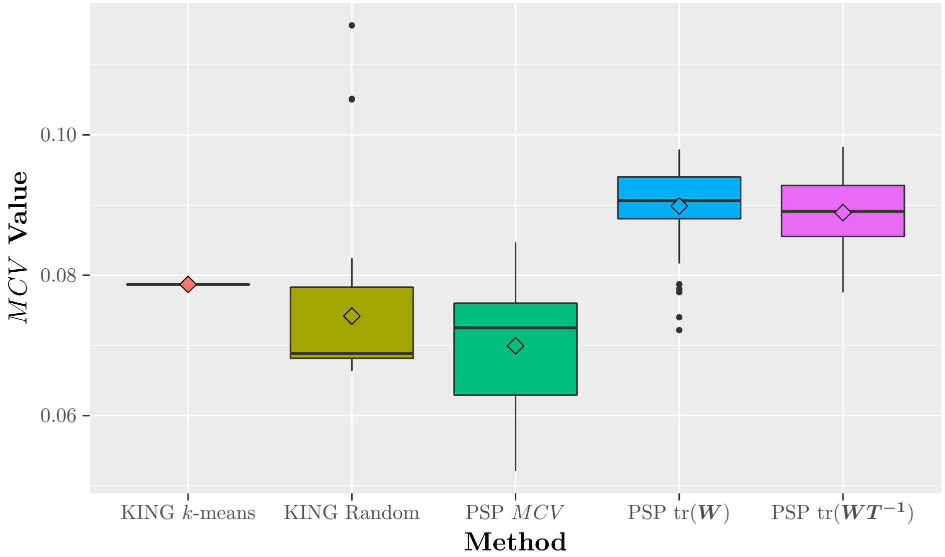

Figure 1: MCV Distributions by Stratification Method, Current Population Survey - Alabama

Each box plot shows for each method the minimum, maximum, mean, first quartile (Q1), median, and third quartile (Q3) of the MCV criterion function values corresponding to the final stratifications produced by the method. On the box plots, the lower endpoint of the lower 'whisker' (or the smallest outlier value below the lower whisker) corresponds to the minimum. Similarly, the endpoint of the line at the top of the upper whisker (or the largest outlier value above the upper whisker) corresponds to the maximum. The top and bottom of the box correspond to Q1 and Q3, respectively. The line in the middle of the box corresponds to the median, and the diamond corresponds to the mean. We make the following observations:

- The PSP MCV method is clearly the best one here, as measured by the MCV evaluation metric. Not only did it produce the final stratification with the lowest MCV metric value (0.0521) but most of its final stratifications have lower MCV metric values than most of the other four methods' final stratifications. The only other method that is even close is the King Random method; notice it actually has a slightly lower median (0.0681) than the PSP MCV median (0.0725). Still, we would use the minimum-valued stratification produced by any method; and based on these runs (with 50 initial stratifications for each method) the odds clearly favor the PSP MCV method.

- The King Random method's best result is better than the single King k-means result; and in fact, most of the Random method's results are better. We shall see in the summary of results across all survey stratification areas (below) that this is true in general, suggesting that using our size-constrained random initializations might be an improvement over using k-means to initialize.

- The PSP trace($\boldsymbol{WT^{-1}}$) and PSP trace($\boldsymbol{W}$) variants appear to produce similar results.

We ran the five PSU stratification methods for a large number of survey stratification areas using the surveys' 2010 sample redesign PSU stratification variable choices. We compared results across the five methods based on this MCV evaluation metric.

To more easily summarize the comparisons, we defined a variable we will call the "minimum-metric relative ratio" or $MMRR$. For each metric and survey stratification area, we see which method has the smallest of the five minimum metric values. Then for method i, we define the minimum-metric relative ratio for that method in that survey stratification area as the ratio of the method i minimum to the smallest minimum.

$$MMRR_i = \frac{u_i}{\min\limits_{j=1to5} u_j}$$

where u_i = minimum evaluation metric value among all final stratifications for method i.

Thus, the method with the best result always has an $MMRR$ value of 1.00, while the other four have values greater than or equal to 1.00.

In our CPS Alabama example, the $MMRR$ values for the five methods (in the order presented in Figure 1) are 1.51, 1.27, 1.00, 1.38, and 1.49. To show the results of these comparisons across all survey stratification areas, we summarize the data using boxplots, shown in Figure 2. (Note: The number of observations - 84 - is the number of survey stratification areas for which we were able to complete 50 initial-stratification runs for all methods. For certain survey stratification areas (e.g., CPS Texas) the number of PSUs and strata were so large that a single run took multiple days.

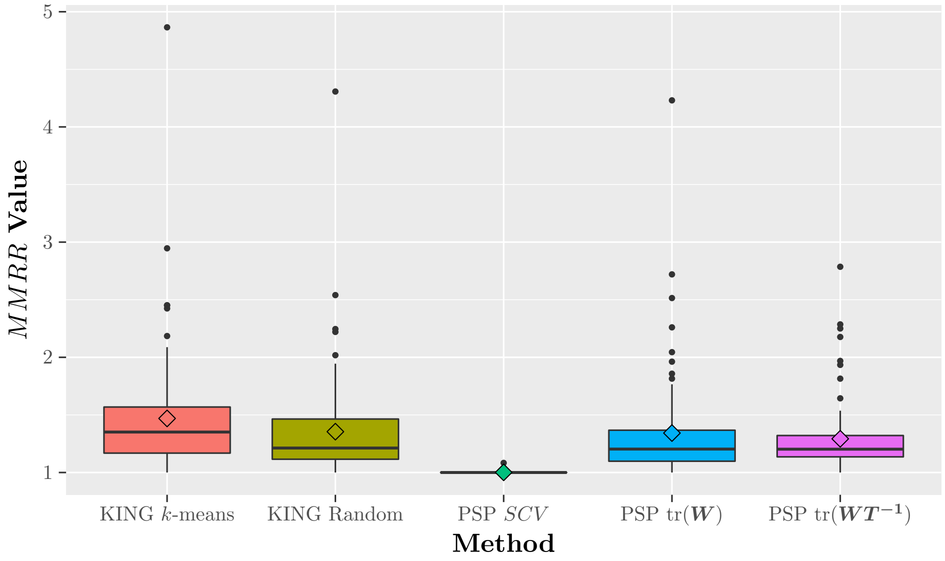

Figure 2: Distribution of Minimum-Metric Relative Ratio of MCV Values by Stratification Method - All Surveys and Geographic Areas

The CPS results for Alabama are typical of results for all surveys and stratification areas we looked at. That is, the PSP using the MCV criterion function almost always finds lower-MCV stratifications than any of the other four methods. There are two cases - NCVS in FIPS state 19 (Iowa) and CPS in FIPS state 35 (New Mexico) - where one of the other methods found slightly better stratifications. Reasons for these anomalies are not immediately apparent; but even in these cases the stratifications found by the PSP using the MCV criterion function are only slightly less good than the best stratification.

3.9 Using trace($\boldsymbol{W}$) as an Evaluation Metric

If we use the trace($\boldsymbol{W}$) criterion function as the evaluation metric instead of the MCV criterion function, the comparison between the five methods is quite different. Going back to the example used in the previous section, we look at the distributions of trace($\boldsymbol{W}$) values for the final stratifications produced by each method for CPS, for the Alabama stratification area. These distributions are shown graphically in Figure 3.

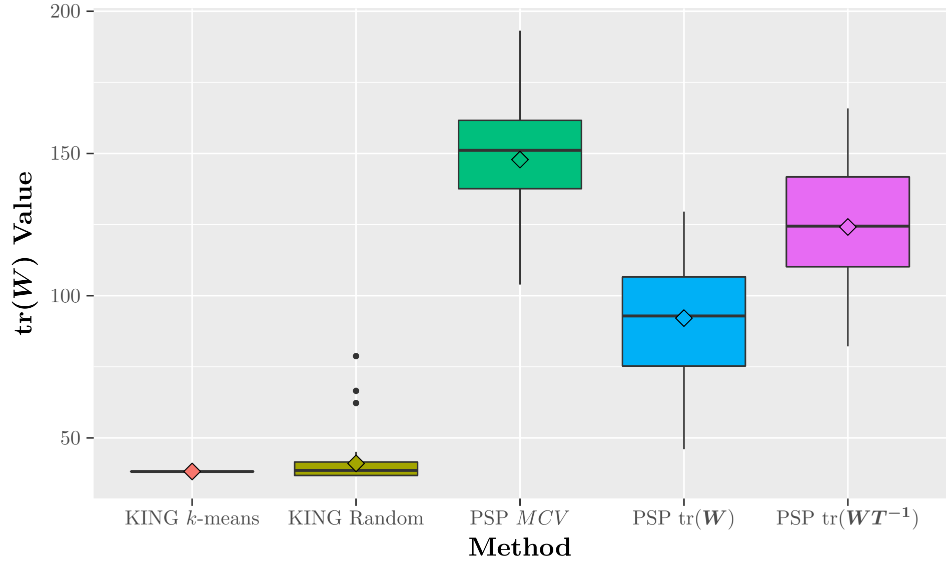

Figure 3: trace($\boldsymbol{W}$) Distributions by Stratification Method Current Population Survey - Alabama

We make the following observations:

- The King Random method is best here, as measured by trace($\boldsymbol{W}$).
- The range of trace($\boldsymbol{W}$) values for the King Random method stratifications is relatively much narrower than the range for any PSP method.
- The King k-means method trace($\boldsymbol{W}$) value is only slightly higher than the best one found by King Random.

The trace($\boldsymbol{W}$) evaluation results for CPS Alabama are broadly typical of what we see for other survey stratification areas; but there are more extreme exceptions to the general trend than we saw with the MCV evaluation metric.

Of the 84 survey stratification areas we looked at, the King Random method does not have the lowest trace($\boldsymbol{W}$) minimum for nine of them. For all nine, the method with the lowest trace($\boldsymbol{W}$) minimum is the PSP using trace($\boldsymbol{W}$). For seven of these, the King Random method minimum is less than two percent larger than the lowest minimum. For CPS Florida (the example shown here), its minimum exceeds the lowest minimum by 103 percent. For CPS Louisiana, and NCVS South Carolina, the King Random method minimum exceeds the lowest minimum by 40 and 171 percent, respectively. Again, these are anomalies, and somewhat puzzling.

As we did for the MCV evaluation metric, we compared results based on the trace($\boldsymbol{W}$) evaluation metric across the five methods for 84 survey stratification areas, and calculated relative minimum ratios. The distribution of the relative minimum ratios for each method are displayed using boxplots in Figure 4.

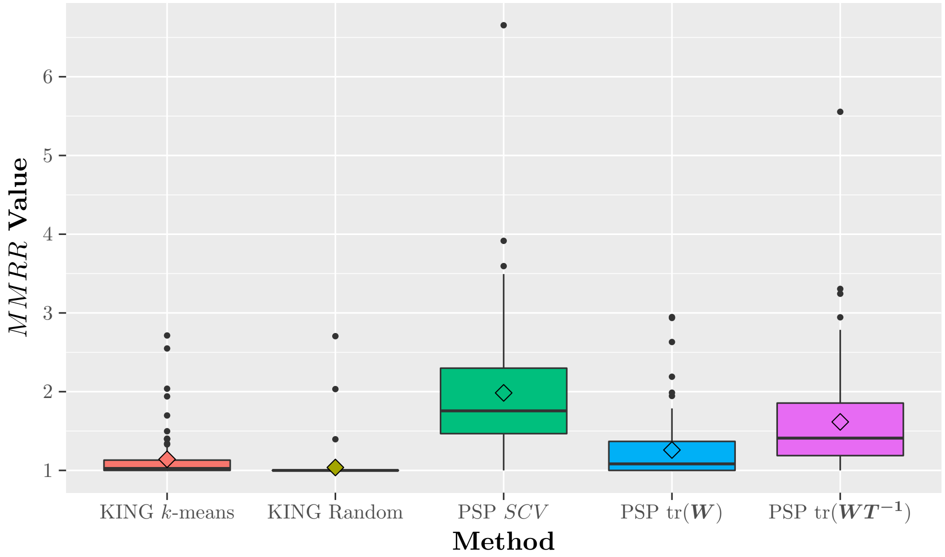

Figure 4: Distribution of Minimum-Metric Relative Ratio of trace($\boldsymbol{W}$) values by Stratification Method - All Surveys and Geographic Areas

It seems clear from this data that the King Random method will find stratifications with the lowest trace($\boldsymbol{W}$), as compared with the other four methods, most of the time. However, the anomalies noted above suggest more investigation is needed to better understand what conditions might cause the King Random method to not perform as well.

4. Conclusions

When using MCV as the evaluation metric, the PSP using the MCV as the minimizing criterion consistently finds better stratifications than all other methods examined, including both King variants. In the particular example we looked at (CPS Alabama,) the best (minimum) value for the PSP using MCV was 0.05, as compared with minimums ranging between 0.07 and 0.08 for the other four values. Looking at distributions of the $MMRR$ values we used to summarize comparisons between methods across the 84 instances of survey stratification areas to which we applied the five methods, the PSP using MCV had an $MMRR$ value of 1.00 for 82 survey-areas. In the two areas where it was greater than 1.00, the values were 1.0012 and 1.0832.

When using trace($\boldsymbol{W}$) as the evaluation metric, the King variant using random size-constrained initial stratifications (King Random) usually finds the best stratification among the methods examined; but we did see three survey-areas where the best stratification found by PSP using trace($\boldsymbol{W}$) had drastically lower trace($\boldsymbol{W}$)

values than the King Random method. One example was CPS Florida, where the minimum metric value for PSP using trace($\boldsymbol{W}$) was 17.8, compared to 36.1 for King Random. In general, however, the King Random method outperformed the other four methods, as we saw in the $MMRR$ distributions when evaluating with trace($\boldsymbol{W}$). For 75 of the 84 survey-areas we looked at, King Random had an $MMRR$ value of 1.00. Of the remaining nine survey-areas, the $MMRR$ value for six of them was less than 1.02; the other three had values of 2.7 (NCVS South Carolina), 1.4 (CPS Louisiana) and 2.0 (CPS Florida).

Looking at other possible evaluation metrics, we find that MCV tends to be negatively correlated with Korhonen's criterion and the ratio of determinants, as well as trace($\boldsymbol{W}$); while Korhonen's criterion, ratio of determinants, and trace($\boldsymbol{W}$) tend to all be positively correlated with each other. Given this observation, we think that trace($\boldsymbol{W}$) (coupled with standard scaling) is a good proxy for either Korhonen's criterion or the ratio of determinants (with no scaling). Therefore, from the set of metrics we have looked at, considering only MCV (with no scaling) and trace($\boldsymbol{W}$) (with standard scaling) is sufficient in evaluating competing stratification methods. Between MCV and trace($\boldsymbol{W}$), we prefer MCV because it explicitly takes into account the PSU and stratum measures of size. The fact that scaling is not necessary to equalize the influence of stratification variables is another plus for the MCV metric. A possible drawback for the MCV metric is that one must be careful about the scale of stratification variables. The coefficient of variation (the basis of the MCV metric) is meaningless for a variable whose values come from a relative scale rather than an absolute scale. The classic example to illustrate this is to consider temperature measured in Farenheit and Celsius. The set of Celsius readings 0,10,20,30,40 corresponds to the set 32,50,68,86,104 in Farenheit. These are physically exactly the same temperatures, but the coefficients of variation for these two sets are 0.79 and 0.42, respectively. However, converting both sets to their absolute scales - Kelvin and Rankine, respectively - one will get the same c.v. value for both sets, 0.0539.

If an argument can be made that trace($\boldsymbol{W}$) is the better metric, then the King Random method usually finds better stratifications (as measured by trace($\boldsymbol{W}$)) than the other methods considered. However, the PSP using trace($\boldsymbol{W}$) with standard scaling often does almost as well; and in a few cases finds a stratification with a much lower trace($\boldsymbol{W}$) than the King Random minimum, as we saw above.

Between the King Random and King k-means variants, we find the Random variant tends to find slightly better stratifications. On the other hand, since it appears that varying the initial sort order of PSUs has no impact on the final result of the k-means variant, one need only create one initial stratification for the King k-means method, so this variant will produce results almost as good in less time than the Random variant. However, for most of the survey stratification areas we looked at, computation time was a matter of minutes for any of the five methods, even with 50 distinct initial stratifications. For small to moderately sized PSU stratification problems, then, the time saved by starting with k-means would not be worth the higher trace($\boldsymbol{W}$) value one is likely to get, as compared with the King Random method. For very large problems (such as AHS in Census Regions 3-8, or any of CPS/NCVS/SIPP in Texas) one run of the King method from a single initial stratification may take much longer. In this case, it might be preferable to sacrifice the possibility of a lower trace($\boldsymbol{W}$) in order to only do one run. (Keep in mind that everything said in this paragraph assumes that trace($\boldsymbol{W}$) is the preferred evaluation metric. If not, using the PSP with the MCV criterion function will almost certainly give better results than either King variant).

Creating a set of unique initial stratifications that satisfy size constraints (which differs from the method used by the 2010 PSP) improves PSP results, since no initial starts are discarded. The same set of initial stratifications appears to help the King method (as compared with initializing using k-means) as observed above. Comparing the PSP with the King method in terms of computation time more generally, we conclude that there is not much difference for small or medium size sets of PSUs. However, for larger sets of PSUs, the computation time for the King method appears to be longer than for the PSP by orders of magnitude. As noted above, we found that the sets of PSUs for AHS in Census Divisions 3-8, and the sets of PSUs in Texas for CPS, NCVS, and SIPP required much more time for the King method to stratify than for the PSP.

Some stakeholders have expressed concern that the staff resources required to do the work of researching, writing specifications, programming, maintenance, and other tasks associated with the in-house development of the PSP was excessive in the 2010 sample redesign. Furthermore, they are concerned that if substantial changes are needed, updating the PSP for the 2020 sample redesign could also require more resources than would be justified by the added value of the result. The 2011 article by King et al. suggested that their approach relies only on software that is available off the shelf, and is therefore less resource intensive than a PSP-like approach. However, it is our feeling that the mathematical knowledge required to properly use that off the shelf software is complex enough that the resource requirements for using the King method would be roughly equivalent to those for the PSP. Moreover, we feel that the PSP will not require substantial changes. Given our preference for the MCV metric, and the demonstrated advantage of the PSP over the King method when evaluating stratifications with the MCV, we believe those concerns are unwarranted.

References

Broder, A. Z. (1984). The r-Stirling Numbers. *Discrete Mathematics*, *49*(2), 241-259.

Clarke, B., Fokoué, E., & Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning.* New York: Springer.

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

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

Jewett, R. S., & Judkins, D. R. (1988, 11). Multivariate Stratification With Size Constraints. *Journal of Scientific Statistical Computing*, *9*(6), 1091-1097.

King, S., Schilp, J., & Bergmann, E. (2011). Assigning PSUs to a Stratification PSU. In *Proceedings of the Joint Statistical Meetings*, Survey Research Methods Section (p. 2235-2246). Alexandria, VA: American Statistical Association.

Korhonen, P. K. (1978). Experiments with Cluster Analysis Criteria Based on the Within Groups Scatter Matrix. In *Proceedings of the Computational Statistics*, 3rd Symposium held in Leiden (p. 266-272). Physica-Verlag, Wien: Computational Statistics.

Kostanich, D., Judkins, D., Singh, R., & Schautz, M. (1981). Modification of Friedman-Rubin's Clustering Algorithm for Use in Stratified PPS Sampling. In *Proceedings of the Joint Statistical Meetings*, Survey Research Methods Section (p. 285-290). Alexandria, VA: American Statistical Association.

Ludington, P. (1992). Stratification of Primary Sampling Units for the Current Population Survey Using Computer Intensive Methods. In *Proceedings of the Joint Statistical Meetings*, Survey Research Methods Section (p. 752-757). Alexandria, VA: American Statistical Association.

Winston, W. L. (1994). *Operations Research: Applications and Algorithms, third edition (C. Hinrichs, Ed.)*. Belmont, California: Wadsworth Publishing Company.