MODIFICATION OF FRIEDMAN-RUBIN'S CLUSTERING ALGORITHM FOR USE IN STRATIFIED PPS SAMPLING

Donna Kostanich, David Judkins, Rajendra Singh, and Mindi Schautz
U.S. Bureau of the Census

I. INTRODUCTION

Major research activities are currently underway at the Census Bureau to improve the quality of the Current Population Survey (CPS). Obtaining a method which reduces the first-stage component of variance is a high priority project that has been and is still being investigated. This paper deals with some of the research conducted thus far on the use of a particular clustering algorithm to achieve this variance reduction. The following section gives a brief discussion of the survey's objectives and the reason for the concern in reducing this component of variance. Section III of this paper discusses the general form of this variance component as it relates to both the design of the CPS and a general clustering situation. Also presented in this section are the modifications made to Friedman and Rubin's clustering algorithm (see {2}) for the above stated purpose. A description of the computer program that has been created to satisfy our modifications along with other related options is given in Section IV. Empirical results on the effect of these modifications under various realistic sampling situations are shown in Section V.

II. BACKGROUND

One of the major purposes of the CPS is to produce labor force statistics at both the national and state levels with the primary objective of obtaining reliable estimates on the number of unemployed. The emphasis of the CPS had recently, during the latter part of 1978, shifted towards producing more reliable unemployment estimates at a substate level. For a more in-depth discussion on the design and objectives of the CPS during the last decade see {4}. More recently (i.e. early 1981), the aim of the CPS has shifted back to obtaining reliable state estimates. This shift was caused by budgetary problems. Additionally, there has been considerable interest in minority estimates. The above mentioned changes in objectives, especially the anticipated shift towards substate estimates, had created concern over the first-stage contribution of variance. For national estimates, this component of variance due to the selection of primary sampling units (PSUs) is relatively small compared to the total variance. This proportion increases for minority estimates at the national level. For some states, however, and even more so for some substate areas, this component of variance can be quite large. Therefore, major research is underway in order to reduce the between-PSU variance for several variables at the state level.

III. GENERAL

The between-PSU variance in the CPS results from the selection of one sample PSU per stratum with probability proportionate to size (PPS). In the current CPS design, PSUs are generally defined to be counties or a contiguous group of counties with the exception that minor civil divisions form PSUs in some northeastern states. The results presented in this paper always consider single counties as PSUs. This component of variance for estimated levels with a given characteristic is given by the following formula:

$$(1)\ Var_B = \sum_{h=1}^{g} \sum_{i=1}^{n_h} \frac{P_{hi}}{P_h} \left(\frac{P_h}{P_{hi}} U_{hi} - U_h \right)^2$$

where g = the number of strata

n_h = the number of PSUs in the h^{th} stratum

P_{hi} = the population of the i^{th} PSU in the h^{th} stratum

P_h = the population of the h^{th} stratum $\left(\text{i.e. } P_h = \sum_{i=1}^{n_h} P_{hi} \right)$

U_{hi} = the actual number of persons with a certain characteristics in the i^{th} PSU of the h^{th} stratum

and U_h = the number of persons with a certain characteristic in the h^{th} stratum $\left(\text{i.e. } U_h = \sum_{i=1}^{n_h} U_{hi} \right)$

The above formula (1) can be rewritten in terms of proportions by letting

$X_{hi} = U_{hi}/P_{hi}$ and $X_h = \frac{U_h}{P_h}$. This gives:

$$(2)\ Var_B = \sum_{h=1}^{g} \sum_{i=1}^{n_h} P_{hi} P_h \left(X_{hi} - X_h \right)^2$$

Consequently, the purpose of this study is to produce a stratification procedure that simultaneously reduces (2) for several key labor force statistics for a given number of strata. Although g is not actually known, it can be assumed to be fixed since its value depends heavily on the desired stratum workload which is fixed and dependent on the size of the stratum. Other factors including reliability requirements and the value of Var_B achieved for unemployed also affect the value of g.

In attempts to accomplish the above, a clustering algorithm suggested by Friedman and Rubin in {2} was used as a basis for reducing this first stage variance component for several variables. In general, their clustering algorithm attempts to optimize a criterion function for a fixed number of clusters. One of the criterion functions suggested in {2} to be minimized is the within cluster sums of squared errors for several variables, which is referred to as the trace W criterion. Note that W is defined to be the within cluster dispersion matrix. Using notation defined above, an element of W, $w_{j,j'}$, is defined as:

$$w_{j,j'} = \sum_{h=1}^{g} \sum_{i=1}^{n_h} \left(X_{hij} - \bar{X}_{hj} \right) \left(X_{hij'} - \bar{X}_{hj'} \right),$$

for j, j' = 1, 2, ... p
where the additional subscript j is for designation of the P variables and

$$\bar{X}_{hj} = \left(\sum_{i=1}^{n_h} X_{hj} \right) / n_h$$

Thus,

$$(3) \text{ Trace } W = \sum_{j=1}^{P} \sum_{h=1}^{g} \sum_{i=1}^{n_h} \left(X_{ihj} - \bar{X}_{hj} \right)^2$$

Note that if the between PSU variance given in (2) was summed over the key labor force statistics, this function would be very similar to that given in (3) above. The only differences are the P_{hi} and P_h terms and manner in which the cluster means are computed. Thus, it seems appropriate to modify (3) to obtain a sum of variances in (2) to use as the criterion function to minimize in order to reduce the between-PSU variance. This new criterion function will be defined as:

$$(4) \text{ Betvar} = \sum_{j=1}^{P} \text{Var}_{B,j}$$

where $\text{Var}_{B,j}$ is the quantity stated in formula (1) above for the j^{th} variable and p is the number of variables to be used in the clustering algorithm. This appears to be a worthwhile addition because of the effect of differing PSU sizes (P_{hi}'s) on both the squared deviations and the clusters means. This modification of trace W was then used as the criterion function along with parts of Friedman and Rubin's clustering algorithm to form a stratification algorithm. It was later discovered that an identical criterion function had been used with a very similar clustering algorithm for sampling purposes (see {1}).

The algorithm specified in {2} actually consists of three separate passes; a hill climbing pass, a forcing pass and a reassignment pass. The hill climbing pass examines each object (PSU) one at a time and moves this object to the group that produces the most improvement in the criterion function. If no improvement occurs the object remains in its current cluster. A one move local minimum occurs when an entire pass of the objects produces no moves. The other two passes are heuristic procedures which attempt to obtain a near global minimum. The forcing pass attempts to optimize the criterion by placing groups of objects from one cluster into the other clusters. For the reassignment pass the algorithm considers moving all objects into different clusters at one time.

Due to the limited availability of programming resources, the effectiveness of these two passes was evaluated using the within cluster sums of squared errors given in formula (3) above. These results, presented in {3}, show that a number of initial starts will produce results similar to those obtained using the forcing and reassignment passes. Therefore, it was decided that programming resources could be more efficiently utilized by incorporating random initial starts and other modifications into the program rather than using these two passes.

The next section of this paper describes more thoroughly the program created for the stratification algorithm including other additions that have been made to assist in this research as it applies to sampling.

IV. STRATIFICATION PROGRAM

This stratification program, developed at the Census Bureau, basically consists of the hill-climbing pass in {2} and a criterion function (4) for between-PSU variances. Other options have also been incorporated into this program for the primary purpose of forming strata for PPS sampling. However, this program can be used for the more conventional clustering problem due to a number of options available to the user.

The program is run on a Univac 1100/44 machine. The program size is 39-43k, with 'k' being defined as 1024 36 bit words. The size of the program depends on the options incorporated in a particular run. The program language is ALGOL.

The clustering program can be altered through the selection of available options, which involve using alternate pieces of code in the program.

The following are the options that alter the clustering program itself:

1) OPTION FOR MAXIMUM/MINIMUM CLUSTER SIZE CONSTRAINTS. Any reasonable maximum and minimum cluster sizes can be assigned, which are referred to as size constraints. Size constraints have been incorporated into the algorithm in order to balance interviewer workloads.

Only one minimum and two maximum size constraints may be specified. One maximum size constraint can be used to exclude unusually large elements or PSUs, which will then be sampled as self-representing PSUs, and the other can be used as the maximum size constraint in the clustering algorithm.

2) OPTION FOR INITIAL ALLOCATION OF PSUs TO CLUSTERS. The following are the three available procedures for assigning the initial allocation of the elements or PSUs to the clusters:

a) This allocation is defined by an external procedure (such as intuition, principal components, graphing, etc.). The number of elements to be in each cluster, n_h, is also defined by the external procedure. A cluster code is placed on each record and the file is then sorted using this field. The first n_1 elements are put into the first cluster and the next n_2 elements are placed into the second cluster, etc.

b) Another method is to first sort the file by the predominant variable(s), such as unemployment. The initial allocation is determined by forming approximately equal sized clusters from this sorted file.

c) A random clustering of the elements may also be achieved by using a procedure that incorporates a random number generator. The elements are randomly placed into clusters so that the clusters meet any reasonable minimum or maximum size constraints. First, PSUs or elements are assigned to meet the minimum size constraint. If, after the minimum size constraint is met, some elements are not yet assigned to a cluster, then the procedure attempts to randomly assign all remaining elements to a cluster. If at any time a clustering scheme cannot be devised such that all elements are included in one of the groups and all clusters satisfy the size constraints, the process is stopped and started again by randomly assigning all elements. The number of random starts to use is left up to the user.

3) OPTION FOR CRITERION FUNCTIONS. The available functions to be minimized are:

a) Trace W; suggested by Friedman and Rubin and also given in (3)

b) Betvar; function of variances given in (4)

c) Determinant W'; W' is specified in Section V.

The Determinant W' option is similar to the Determinant W criteria given in {2} but incorporates the PSU sizes into the within cluster dispersion matrix W. Some results showing the applicability of this function as a criterion to minimize are included in the next section of this paper.

4) OPTION TO STANDARDIZE VARIABLES. The variables being used to determine the clusters may be standardized by one of the following methods (no standardization is also an option):

a) probability proportionate to size scaling. For this option the scaling factor, sf_j, applied to all observations on the j^{th} variable, is determined by:

$$sf_j = \frac{\max_k \sqrt{\sum_{i=1}^{N} P_1 P_{1i} (X_{1i,k} - X_{1,k})^2}}{\sqrt{\sum_{i=1}^{N} P_1 P_{1i} (X_{1i,j} - X_{1,j})^2}}$$

for j and k = 1, 2, ... p

where p is the number of variables, N is the total number of PSUs or elements and all PSUs are assumed to be in one cluster (i.e. h = 1 implies that $\Sigma n_h = N$ and $\Sigma P_h = P_1$). Note that P_1 is the population of the entire NSR area.

b) equal size scaling. For this option all observations or elements are considered equally important in determining the scaling factors sf_j, applied to all observations on the j^{th} variable. These factors are obtained by:

$$sf_j = \frac{\max_k \sqrt{\sum_{i=1}^{N} (\bar{X}_{1i,k} - \bar{X}_{1,k})^2}}{\sqrt{\sum_{i=1}^{N} (X_{1i,j} - \bar{X}_{1,j})^2}}$$

where j, k, p and N are defined above in a). Scaling methods a.) and b.) above, respectively, equalize the between-PSU variance or the sums of squares for all variables given that all elements are in one cluster.

5) OPTION TO CHOOSE NUMBER OF CLUSTERS. The number of clusters to be formed is determined by the user; however, the actual number of clusters formed may be internally reduced by:

a) the removal of any individual element whose size (P_{hi}) is greater than the maximum allowable cluster size;

b) the creation of an initial empty cluster through the random starts procedure. This was allowed in order to avoid complexities, since an empty cluster would not meet the minimum size constraint. The possibility of this occurring is very slight if one chooses a reasonable number of clusters and reasonable size constraints.

6) OPTION TO CHOOSE PREFERENCE FACTORS. Any preference factors or variable weights may be assigned to each variable. If factors are indicated here, they are incorporated into the Betvar criterion function in order to weight the contribution from some variables more than others. Consequently, this forces a greater reduction in between-PSU variance for the more heavily weighted variables as determined by the user. If the variables were not standardized initially, the contribution of each variable to the Betvar criterion would differ. These contributions would be dependent on the general magnitude of the variances rather than the importance of the actual variable. If this option is not specified, all preference factors are assumed equal to one.

V. EMPIRICAL RESULTS

Some preliminary results showing the effectiveness that the above mentioned modifications have on reducing the between-PSU variance are presented in the attached tables. Strata were formed within three test states with PSUs being defined as counties. Not all counties in a state were included in this testing; only those considered to be nonself-representing (NSR) were used to form strata. Counties excluded were assumed to be self-representing and were designated as such since their population would yield a sample large enough to support an interviewer. For example, counties included in the Pittsburgh SMSA were not stratified and would be in sample with certainty. These three states were arbitrarily selected for testing as they seemed to be somewhat representative of the country. For each state the algorithm was tested using four key labor force variables, which are important in the CPS. Each of these states includes the number of unemployed and the number in the civilian labor force (CLF) as variables 1 and 2, respectively, and the resulting between-PSU variances are designated by $Var_{B,1}$ and $Var_{B,2}$. Variables 3 and 4 represent similar minority characteristics of unemployed and CLF respectively, where Spanish was used in Colorado and blacks in Mississippi and Pennsylvania. The number of strata formed in each state is an approximation of the number that would be needed to satisfy the anticipated reliability requirements on the state estimate of the number of unemployed for the redesigned CPS.

The NSR PSUs in each of these states were clustered into strata using the following criteria:

1.) $\text{Betvar} = \sum_{j=1}^{4} Var_{B,j}$

2.) $\text{Trace } W = \sum_{j=1}^{4} \sum_{h=1}^{g} \sum_{i=1}^{n_h} (X_{ihj} - \bar{X}_{hj})^2$

3.) Determinant W'; where W' can be written as:

$$\begin{bmatrix} Var_{B,1} & Cov_{B,1,2} & Cov_{B,1,3} & Cov_{B,1,4} \\ Cov_{B,1,2} & Var_{B,2} & Cov_{B,2,3} & Cov_{B,2,4} \\ Cov_{B,1,3} & Cov_{B,2,3} & Var_{B,3} & Cov_{B,3,4} \\ Cov_{B,1,4} & Cov_{B,2,4} & Cov_{B,3,4} & Var_{B,4} \end{bmatrix}$$

by defining the covariance terms as:

$$Cov_{B,j,j'} = \sum_{h=1}^{g} \sum_{i=1}^{n_h} P_{hi} P_h (X_{hij} - X_{hj})(X_{hij'} - X_{hj'})$$

4.) Betvar = $\sum_{j=1}^{4} pf_j \, Var_{B,j}$

where the pf_j's represent preference factors discussed in section IV of this paper, for which the following values were assigned:

$pf_1 = 2$ and $pf_2 = pf_3 = pf_4 = 1$

This was tested in order to examine the effect of weighting the most important labor force variable, unemployed, more heavily.

5.) Betvar = $Var_{B,1}$

6.) Betvar = $Var_{B,2}$

7.) Betvar = $Var_{B,3}$

8.) Betvar = $Var_{B,4}$

Note that these last four criteria produce a stratification based on only one of the four key labor force variables and the results give some indication of the amount of reduction in between-PSU variance that can be achieved for each variable.

Each of the above eight criteria were used to cluster NSR PSUs into strata in each of the three test states. Also for each criterion and state a stratification was produced by imposing three types of size constraints on the strata, none, loose and tight. Three random starts[1] were implemented for each type of size constraint and the resulting average between-PSU variances, $\overline{Var}_{B,j}$, were obtained and are presented in the attached tables. Note also that for a given size constraint, identical initial starts were used in clustering the PSUs with each of the eight criteria. For all stratifications the actual variables were first standardized with scaling factors determined by the probability proportionate to size method discussed in section IV of this paper.

Tables 1, 2, 3 and 4 present the average between-PSU variances that were obtained from the first four criteria given above, respectively. The average between-PSU variances resulting from stratifying on each variable separately (i.e. criteria 5.)-8.) above) are shown in Table 5. For purposes of comparison, the percent reductions in between-PSU variance for each variable are also given in these tables. These percent reductions are based on what the between-PSU variance would be if the NSR PSUs were not stratified and g PSUs were selected with probability proportionate to size with replacement. These nonstratified between-PSU variances, $Var^*_{B,j}$, were calculated for each test state by:

$$Var^*_{B,j} = \frac{\sum_{i=1}^{N} P_{1i} \, P_1 \, (X_{1ij} - X_{1j})^2}{g} \quad \text{for } j = 1, 2, 3 \text{ and } 4$$

where N represents the number of NSR PSUs in the state.

Chart A summarizes the nonstratified between-PSU variances for each state and variable and also gives other stratification parameters used in implementing the algorithm.

The effect of incorporating probabilities of selection in the clustering criterion can be seen by comparing the results in Table 1 with those in Table 2. Greater reductions in between PSU variances occurred with the Betvar criterion for all states and all variables when no size constraints were imposed on the strata. In fact, in one instance the Trace W criterion produced an increase in the between-PSU variance for a particular variable. This happened in Colorado to variable 2, CLF, when no size constraints were imposed; as can be seen from the negative percent reduction in Table 2. When size constraints are imposed on the strata, generally the Betvar criterion produces gains over the Trace W criterion although these gains are smaller. In Colorado and Mississippi, with tight size constraints, one of the variables, variable 3 - Spanish unemployed in Colorado and variable 1 - unemployed in Mississippi, achieved greater variance reduction with the Trace W criterion; however, for all variables concerned the Betvar criterion produced a better stratification.

For Pennsylvania, when loose size constraints were imposed the Trace W criterion produced an overall better stratification (i.e. three variables achieved a greater variance reduction). However with tight size constraints half the variances were smaller using the Trace W criterion the others were reduced more with the Betvar criterion. This preliminary comparison shows quite strongly that the modifications made to the Trace W criterion will produce significant reductions in the first stage variance component when no size constraints are imposed. It also appears that in general and under certain situations smaller gains in variance reduction can be achieved even with size constraints on strata.

The variances for stratifications produced by minimizing the Determinant W' are shown in Table 3. The applicability of this criterion to stratification is of interest since it is a function of both between-PSU variances and within cluster covariances. By comparing these results with those obtained by minimizing only the between-PSU variances, (see Table 1), it appears that the within cluster covariances do not in general improve the stratification. Most cases where greater variance reductions are achieved with the Determinant, the Betvar criterion does not produce substantially larger variances. However, an unusual situation did occur for unemployed in Pennsylvania, smaller variances were consistently obtained from the Determinant for all three types of size constraints.

Also included in this study is an example of the effect of preference factors, pf_j's, on the stratifications. As previously discussed the pf_j's are incorporated in the Betvar criterion in order to weight particular variables more heavily than others. This example simply weights the $Var_{B,1}$ (variance for unemployed) twice that of the others. As anticipated, greater reductions in the between-PSU variance for unemployed (i.e. see Tables 1 and 4) did occur, but by differing amounts. The decreases were generally substantial, particularly when no size constraints were imposed. It appears that the between-PSU variances for the other three variables can also increase considerably.

The results presented in Table 5 somewhat indicate the amount of variance reduction that is possible for a variable under various conditions, since these stratifications were based on only one variable. It should be noted that this procedure will not produce the minimal variance because results are somewhat dependent on the initial clustering due to the fact that the algorithm only produces a local minimum. However, these variances can serve as a basis for determining the effectiveness of a multivariate stratification. Also worthy of mention is that greater reductions in between-PSU variances can be achieved if strata sizes are allowed to vary. This can be seen in all the stratifications produced, but is especially emphasized when only one stratification variable is used (i.e. Table 5).

Although, the results presented in this paper are not conclusive, research is currently focused on ways to achieve greater variance reduction when size constraints are imposed on strata. This is being investigated under the assumption that the Betvar criterion along with preference factors will be the most practical way to form strata for the CPS.

FOOTNOTE

[1]Some of the average between-PSU variances for Mississippi were obtained from only two random starts.

BIBLIOGRAPHY

{1} Dahmström, P. and Hagnell, M. "Multivariate Stratification of Primary Sampling Units in a General Purpose Survey." Scand. J. Statist. 1978: 5, 53-56.

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

{3} Judkins, D. R. and Singh, R. P. "Using Clustering Algorithms to Stratify Primary Sampling Units." Presented at the 141th Annual Meeting of the American Statistical Association, August 1981.

{4} Moore, T.; Bettin, P.; Kostanich, D.; and Shapiro, G. "Overview of Current Population Survey Sample Design." Presented at 139th Annual Meeting of the American Statistical Association, August 1979.

Chart A

Summary of Stratification Parameters and Nonstratified Variances

Parameter \ State	Colorado	Mississippi	Pennsylvania
Number of PSUs(N)	55	79	58
Number of Strata(g)	5	12	8
$Var^*_{B,1}$ (x 10^5)	30.00	105.82	870.63
$Var^*_{B,2}$ (x 10^5)	1,164.60	1,677.33	42,007.50
$Var^*_{B,3}$ (x 10^5)	15.00	126.13	9.06
$Var^*_{B,4}$ (x 10^5)	1,522.78	4,815.42	2,676.00
Scaling factors: sf_1	7.124	6.746	6.946
sf_2	1.143	1.694	1.000
sf_3	10.080	6.179	68.090
sf_4	1.000	1.000	3.962

Table 1. Reductions in Variance Due to Stratification (Based on Betvar[1/] Criterion)

State and Size Constraints	$\overline{Var}_{B,1}$ [2/] (% Reduction)	$\overline{Var}_{B,2}$ [2/] (% Reduction)	$\overline{Var}_{B,3}$ [2/] (% Reduction)	$\overline{Var}_{B,4}$ [2/] (% Reduction)
Colorado				
None	11.31 (62%)	401.98 (65%)	3.62 (76%)	332.59 (78%)
Loose	17.58 (41%)	786.27 (32%)	6.69 (55%)	553.62 (64%)
Tight	19.14 (36%)	763.44 (34%)	6.35 (58%)	737.14 (52%)
Mississippi				
None	17.69 (83%)	316.56 (81%)	9.32 (93%)	588.10 (88%)
Loose	16.49 (84%)	394.77 (76%)	9.55 (92%)	621.08 (87%)
Tight	46.31 (56%)	656.94 (61%)	31.72 (75%)	1,347.82 (72%)
Pennsylvania				
None	131.67 (85%)	8,165.93 (81%)	.71 (92%)	94.87 (96%)
Loose	338.28 (61%)	13,836.20 (67%)	4.35 (52%)	1,293.74 (52%)
Tight	411.30 (53%)	16,346.03 (61%)	4.35 (52%)	1,297.09 (52%)

1/ This criterion is defined as: $\text{Betvar} = \sum_{j=1}^{4} Var_{B,j}$

2/ The average between-PSU variances are x 10^5

Table 2. Reductions in Variance Due to Stratification

(Based on Trace W Criterion)

State and Size Constraints	$\overline{Var}_{B,1}$ 1/ (% Reduction)	$\overline{Var}_{B,2}$ 1/ (% Reduction)	$\overline{Var}_{B,3}$ 1/ (% Reduction)	$\overline{Var}_{B,4}$ 1/ (% Reduction)
Colorado				
None	16.60 (45%)	1,281.79 (-10%)	4.98 (67%)	434.11 (71%)
Loose	19.43 (35%)	838.89 (28%)	8.58 (43%)	880.57 (42%)
Tight	20.80 (31%)	784.42 (33%)	6.02 (60%)	786.55 (48%)
Mississippi				
None	18.91 (82%)	424.23 (75%)	12.02 (90%)	767.65 (84%)
Loose	23.30 (78%)	541.75 (68%)	12.94 (90%)	698.51 (85%)
Tight	42.88 (59%)	844.31 (50%)	31.28 (75%)	1,538.99 (68%)
Pennsylvania				
None	172.63 (80%)	8,381.78 (80%)	1.75 (81%)	282.03 (89%)
Loose	216.40 (75%)	9,449.39 (78%)	2.25 (75%)	960.91 (64%)
Tight	455.61 (48%)	12,355.56 (71%)	3.94 (57%)	1,346.34 (50%)

1/ The average between-PSU variances are x 10^5.

Table 3. Reductions in Variance Due to Stratification

(Based on Determinant W´ Criterion)

State and Size Constraints	$\overline{Var}_{B,1}$ 1/ (% Reduction)	$\overline{Var}_{B,2}$ 1/ (% Reduction)	$\overline{Var}_{B,3}$ 1/ (% Reduction)	$\overline{Var}_{B,4}$ 1/ (% Reduction)
Colorado				
None	13.32 (56%)	593,62 (49%)	7.30 (51%)	603.69 (60%)
Loose	21.42 (30%)	728.16 (37%)	6.29 (58%)	538.48 (65%)
Tight	17.71 (41%)	833.58 (28%)	7.68 (49%)	676.42 (55%)
Mississippi				
None	21.44 (80%)	562.96 (66%)	17.50 (86%)	765.08 (84%)
Loose	25.80 (76%)	710.18 (58%)	25.48 (80%)	1,077.20 (78%)
Tight	44.29 (58%)	880.79 (47%)	39.99 (68)	2,153.36 (55%)
Pennsylvania				
None	111.57 (87%)	10,149.13 (76%)	1.01 (89%)	128.21 (95%)
Loose	290.29 (67%)	16,659.13 (60%)	4.39 (52%)	1,330.92 (50%)
Tight	312.78 (64%)	15,577.57 (63%)	4.41 (51%)	1,331.46 (50%)

1/ The average between-PSU variances are x 10^5.

Table 4. Reductions in Variance Due to Stratification

(Based on Betvar 1/ Criterion with Preference Factors)

State and Size Constraints	$\overline{Var}_{B,1}$ 2/ (% Reduction)	$\overline{Var}_{B,2}$ 2/ (% Reduction)	$\overline{Var}_{B,3}$ 2/ (% Reduction)	$\overline{Var}_{B,4}$ 2/ (% Reduction)
Colorado				
None	6.64 (78%)	707.98 (39%)	3.74 (75%)	394.14 (74%)
Loose	15.07 (50%)	893.96 (23%)	7.19 (52%)	632.71 (58%)
Tight	16.81 (44%)	876.26 (25%)	7.37 (51%)	796.96 (48%)
Mississippi				
None	8.74 (92%)	502.01 (70%)	11.03 (91%)	733.92 (85%)
Loose	16.14 (85%)	675.78 (60%)	22.14 (82%)	1,492.69 (69%)
Tight	26.89 (75%)	1,127.18 (33%)	37.75 (70%)	2,665.30 (45%)
Pennsylvania				
None	78.02 (91%)	11,389.25 (73%)	1.08 (88%)	141.16 (95%)
Loose	245.31 (72%)	17,270.57 (59%)	5.35 (41%)	1,404.12 (48%)
Tight	343.12 (61%)	24,191.70 (42%)	4.52 (50%)	1,330.17 (50%)

1/ This criterion is defined as: Betvar = $\sum_{j=1}^{4} pf_j \, Var_{B,j}$ with $pf_1 = 2$ and $pf_2 = pf_3 = pf_4 = 1$.

2/ The average between-PSU variances are x 10^5.

Table 5. Reductions in Variance Due to a One-Variable Stratification (Based on a different Betvar Criterion for Each Variable)

Criterion	$Var_{B,1}$	$Var_{B,2}$	$Var_{B,3}$	$Var_{B,4}$
State and Size Constraints	$\overline{Var}_{B,1}$ 1/ (% Reduction)	$\overline{Var}_{B,2}$ 1/ (% Reduction	$\overline{Var}_{B,3}$ 1/ (% Reduction)	$\overline{Var}_{B,4}$ 1/ (% Reduction)
Colorado				
None	1.83 (94%)	54.04 (95%)	.57 (96%)	74.68 (95%)
Loose	13.24 (56%)	570.95 (51%)	6.08 (59%)	528.04 (65%)
Tight	14.64 (51%)	708.69 (39%)	6.01 (60%)	639.23 (58%)
Mississippi				
None	1.20 (99%)	18.27 (99%)	.96 (99%)	40.63 (99%)
Loose	21.99 (79%)	299.24 (82%)	20.76 (84%)	309.81 (94%)
Tight	18.68 (82%)	318.99 (81%)	23.95 (81%)	795.68 (83%)
Pennsylvania				
None	15.58 (98%)	965.73 (98%)	.07 (99%)	9.03 (99.7%)
Loose	214.83 (75%)	10,509.41 (75%)	4.55 (50%)	1,224.35 (54%)
Tight	427.08 (51%)	14,385.56 (66%)	4.05 (55%)	1,226.74 (54%)

1/ The average between-PSU variances are x 10^5.

Note that the average between-PSU variances for the four variables represent different stratifications.