Evaluating Alternative Criteria for Primary Sampling Units Stratification

Khandaker A. Mansur, and Benjamin M. Reist, U.S. Census Bureau[1]
U.S. Census Bureau, Washington, D.C.20233

Key Words: 2000 Design; Primary Sampling Unit; Stratification; Friedman-Rubin Clustering Algorithm; Criterion Functions

Abstract

Major research projects are being conducted at the Census Bureau to redesign the Demographic Surveys for the 2010s. One project includes research to obtain a method for stratifying the primary sampling units (PSUs). In this paper we revisit the Friedman-Rubin's clustering algorithm that has been used in the last three redesigns for stratification. This clustering algorithm attempts to optimize a criterion function for a fixed number of strata. The most commonly used criteria functions are Minimum variance, Wilks' lambda and Hotelling's trace. These criteria along with the criterion (2000 criterion) used in the 2000 redesign are being studied empirically and their performances and comparisons under these criteria are also being discussed. Results suggested by these criteria show that the minimum variance criterion provides the best stratification for labor force characteristics.

1. Introduction

The Current Population Survey (CPS) is a probability sample survey of the U.S. population conducted monthly by the U.S. Census Bureau for the Bureau of Labor Statistics. Its primary purpose is to provide monthly estimates of labor force characteristics. An important part of the demographic survey redesign is stratification of CPS primary sampling units (PSUs). Stratification clusters PSUs into strata from which a subset of (sample) PSUs is selected. However, strata produced during stratification need to be 'homogeneous', so survey estimates derived from the sample areas will also accurately reflect non-sample areas. The degree of stratum homogeneity and the achieved reduction in survey costs both depend on the capabilities of the PSU stratification.

The goal of the clustering analysis is to find the "best partition" of n objects (PSUs) into g groups (strata). We define the best partition by introducing the numerical valued function known as the criterion function defined for all partitions of the PSUs into g strata, and selecting a partition for which the numerical measure is minimal.

In the last three redesigns, the Friedman-Rubin (FR) clustering algorithm with the criterion function known as the 'trace W', defined later, was used for reducing the first-stage variance component for several variables. We revisit the Friedman-Rubin's clustering algorithm and all criteria that have been used in the last three redesigns for stratification. We examined all criteria including the one (trace W) used in the last redesign to see which criterion performs the best in reducing the between-PSU variances in the CPS. The between-PSU variance in the CPS accounts for the variability due to the selection of one sample PSU per stratum with probability proportionate to size (PPS).

[1] Any views expressed are those of the authors and not necessary those of the U.S. Census Bureau.

2. Criterion Functions

The FR clustering algorithm is used to form clusters of PSUs in such a way that a criterion function is minimal for several variables for a given number of clusters. The most widely known criterion functions (Korthonen 1978) are:

- Minimum variance
- Wilks' lambda
- Hotelling trace

In addition to the above, we examined the criterion function trace W that was used in the last redesigns. In order to examine these results, we investigated them numerically and compared their performances. In our investigation we found that the minimum variance criterion, without a size constraint, produces a good stratification overall.

3. Clustering Criteria Derived from the Scatter Matrix

Assume the populations are multivariate normal with a common covariance matrix and that the population is known. Assume that we have an observation matrix, $[\mathbf{X}]=\lfloor x_{ij} \rfloor$ of order (nxp). The total scatter matrix T is:

$$T= \sum_{k=1}^{g} \sum_{l=1}^{n_k} (X_{lk}-\overline{X})^t (X_{lk}-\overline{X}) \qquad (1)$$

Suppose a given partition of *g* groups, with the number of observations $n_1, n_2, \ldots, n_g$ in each group and $n = \sum_{i=1}^{g} n_i$. Then for the kth group, the row vectors X_{lk} for $l=1,\ldots,n_k$ represent the objects in group G_k with center of gravity vector $\overline{X}$. This total scatter matrix can be partitioned into the within-group scatter matrix

$$W= \sum_{k=1}^{g} W_k = \sum_{k=1}^{g} \sum_{l=1}^{n_k} (X_{lk}-\overline{X}_k)^t (X_{lk}-\overline{X}_k) \qquad (2)$$

where $\overline{X}_k$ is the mean of the n_k observations in G_k group, and the between-group scatter matrix is defined by

$$B= \sum_{k=1}^{g} n_k (\overline{X}_k-\overline{X})^t (\overline{X}_k-\overline{X}) \qquad (3)$$

For each partition of *n* objects into *g* groups, the total scatter matrix T can be written as the sum of within and between-groups scatter matrices (Friedman and Rubin 1967):

$$T=B+W \qquad (4)$$

The expressions of the clustering criteria based on the scatter matrices are:

$$\text{Minimum variance} = \operatorname{tr}(WT^{-1})$$

$$\text{Wilks' lambda} = \frac{|W|}{|T|} = \left|WT^{-1}\right|$$

$$\text{Hotelling trace} = \operatorname{tr}(BW^{-1})$$

$$\text{2000 criterion} = \operatorname{tr}(W)$$

The tr(W) can be written as:

$$\mathrm{tr}(W_k) = \sum_{l,m=1}^{n_k} \frac{(X_{lk} - \overline{X}_{mk})(X_{lk} - \overline{X}_{mk})}{n_k}, l < m$$

$$\mathrm{tr}(W) = \sum_{k=1}^{g} \sum_{l,m=1}^{n_k} \frac{(X_{lk} - \overline{X}_{mk})(X_{lk} - \overline{X}_{mk})}{n_k}, l < m \qquad (5)$$

From eq. (1), we write

$$\mathrm{tr}(T) = \mathrm{tr}(B) + \mathrm{tr}(W) \qquad (6)$$

Since T is constant over all the partitions, minimizing tr(W) is equivalent to maximizing tr(B).

$$\left|W^{-1}T\right| = \left|I + W^{-1}B\right| \qquad (7)$$

$$\mathrm{tr}\,(BW^{-1}) = \mathrm{tr}(TW^{-1}) - p \qquad (8)$$

where p is the rank of W. Equation (5) shows that minimizing the minimum variance, $\mathrm{tr}(WT^{-1})$ is equivalent to maximizing $\mathrm{tr}(BT^{-1})$ that is expressed as:

$$\mathrm{tr}(BT^{-1}) = \sum_{k=1}^{g} n_k (\overline{X}_k - \overline{X})^t \, T^{-1} (\overline{X}_k - \overline{X}) \qquad (9)$$

Since $|T|$ is independent of grouping in equation (1), the criterion $|W|$ is equivalent to the Wilks' lambda criterion. So, minimizing $\left|WT^{-1}\right|$ is equivalent to minimizing $|W|$. Minimizing $|W|$ is another widely used criterion suggested by Friedman and Rubin (1967). From equation (5), we see that minimizing $\mathrm{tr}(TW^{-1})$ is equivalent to maximizing $\mathrm{tr}\,(BW^{-1})$, the Hotelling trace.

The special case of two groups has been considered by Scott and Symons (1971). He shows that minimizing $|W|$ is equivalent to maximizing $\mathrm{tr}(BT^{-1})$. Using this makes computation easier but unfortunately, the result does not extend in a simple way to more than two groups.

Indeed, these criterion functions can be presented in terms of eigenvalues $\lambda_1 \ldots\ldots \lambda_q$ of WT^{-1}:

$$\text{Minimum variance} = \mathrm{tr}(WT^{-1}) = \sum_{i=1}^{q} \lambda_i + (p - q)$$

$$\text{Wilks' lambda} = \frac{|W|}{|T|} = \left|WT^{-1}\right| = \prod_{i=1}^{q} \lambda_i$$

$$\text{Hotellings trace} = \mathrm{tr}(BW^{-1}) = \sum_{i=1}^{q} \lambda_i^{-1}$$

where q is the rank of B and W is a positive definite with a rank p.

These eigenvalues are solutions of the determinant equation $\left|W-\lambda T\right|=0.$ All eigenvalues of this equation are known to be invariant under nonsingular linear transformations of the original data matrix. In fact they are the only invariants of W and B under such transformations (Friedman and Rubin1967).

In the CPS, one sample PSU is chosen from each stratum with PPS. For PPS sampling (Kostanich 1981), the within-stratum sum of squares scatter matrix (eq.2) can be written as:

$$w_{i,j} = \sum_{h=1}^{g} \sum_{k=1}^{n_h} \left(\frac{P_h}{P_{hk}} U_{hki} - U_{hi}\right)\left(\frac{P_h}{P_{hk}} U_{hkj} - U_{hj}\right) \qquad (10)$$

where g= the number of strata

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

P_{hk} =the measure of size of the k^{th} PSU in the h^{th} stratum

P_h = the measure of the h^{th} stratum

U_{hki} = the value of the i^{th} stratification variable in the k^{th} PSU in the h^{th} stratum

U_{hi} = the value of the i^{th} stratification variable in the h^{th} stratum

And the between-strata variance scatter matrix (eq.3) is:

$$b_{i,j} = \sum_{h=1}^{g} n_k \overline{U}_{hi} \overline{U}_{hj} \qquad (11)$$

where $\overline{U}_{hi}$ = the mean of the i^{th} stratification variable in the h^{th} stratum

$\overline{U}_{hj}$ = the mean of the j^{th} stratification variable in the h^{th} stratum

4. Friedman-Rubin Algorithm

There are two types of clustering algorithms: hierarchical and non-hierarchical. The hierarchical algorithms seek a family of stratifications such that there is a stratification for every possible number of strata and such that every stratum in each stratification is contained in exactly one stratum of every higher stratification. The non-hierarchical algorithms seek exactly one stratification with a prescribed number of strata. We study here only one non-hierarchical algorithm called the FR algorithm.

The FR algorithm is characterized by the iterative reallocation of PSUs to strata in such a way as to optimize the criterion function. The FR algorithm can optimize any one of the above three different criterion functions. There are three different procedures for determining which reallocations to try. These procedures are referred to as the hill climbing procedure, the exchange procedure, and the size adjustment procedure[2].

[2] This procedure was used only by the Census Bureau for its last redesigns and not used in this paper. It was stated here for the future research.

In the hill climbing procedure, individual PSUs are moved one at a time from one stratum to another in an attempt to reduce the criterion functions. A one move local minimum occurs when an entire pass of the objects produces no moves.

The exchange procedure also attempts to minimize the criteria by selecting pairs of PSUs from different strata and interchanging them.

The size adjustment procedure performs all possible moves and exchanges of PSUs from one stratum into another in order to reduce the disparity in stratum sizes. The one resulting in the smallest variance increase per person is chosen. After the size adjustment procedure, all strata should have populations within the size constraint. One goal of using the size adjustment procedure is to keep the strata roughly equal in size. This is important because it allows the design to be self-weighting while keeping the amount of work (interviewers) constant across PSUs. Due to our time constraints, the size adjustment procedure was not performed in our study.

5. Clustering Program

The clustering program developed for this paper was written in SAS/IML®. It used five initial randomly assigned stratifications, one hill climbing pass, one exchange pass and was also written so that the user could specify which of the four criterion functions should be used in the optimization. It is important to note that the procedures for forming initial strata; number of strata and cluster sizes given in sections 5.1, 5.2 and 5.3 respectively were not used in this paper due to our time constraints, but it would be useful for the future research on stratification.

5.1 Forming Initial Strata

Initial strata can be formed by allocating PSUs to strata using a random number generator.

Assignment of one PSU to each stratum:

(1) Randomly generate a stratum number, k, between 1 and s using the seed generated from the computer.

(2) Randomly select a PSU to assign to the randomly selected stratum. Randomly select another PSU from the remaining PSUs and assign it to the following stratum, k+1. Continue assigning randomly selected PSUs to consecutive strata until each stratum contains one PSU.

Assignment of remaining elements:

(1) Randomly select a stratum and assign the next remaining PSU to that stratum if the addition does not cause the total size of the stratum to exceed the upper size constraint.

(2) If the assignment cannot be made without exceeding the size constraint, generate another stratum number and repeat step (1).

5.2 Number of Strata

The number of strata is somewhat arbitrary and can be chosen to be between 4 and 15, depending on the size of the state. However, the number of strata to be formed in each state is an approximation of the number that would be needed to satisfy the anticipated requirements on the state estimate of the number of unemployed for the CPS.

The maximum number of strata (G_{mx}) required while meeting specified reliability requirements can be computed from the following formula:

$$(G_{mx}) = \frac{(N_h/N_{16+})Y_\bullet + (T_{wl}\,.\,b\,.\,\sigma^2_{Bmx})}{(T_{wl}\,.\,SI_{mx})}$$

where N_h = number of housing units (HUs)

N_{16+} = number of civilian noninstitutional persons 16+

T_{wl} = target workload for an NSR PSU[3]

b = adjustment factor for between-PSU variance

σ^2_{Bmx} = maximum between PSU variance on unemployed

SI_{mx} = maximum sampling interval required

$Y_\bullet$ = total number of persons 16+ over all NSR PSUs in the stratum

The maximum stratum size (MX_S) is:

$$MX_S = 1.2 \max \sum_{j=1}^{k} Y_j / G_{mx} \qquad \text{and}$$

The minimum stratum size (MN_S) is:

$$MN_S = \ 0.7\ [(\sum_{j=1}^{k} Y_j / G_{mx}), \max(\text{NSR PSU})]$$

where $\sum_{j=1}^{k_g} Y_j$ =total NSR pop, k= Number of NSR PSU, max(NSR PSU) = the largest Y_j in any NSR PSU

5.3 Maximum and Minimum Cluster Size

Any reasonable maximum and minimum cluster size can be assigned, which are referred to as size constraints. The following formulas should be used to compute size constraints for the strata.

(1) Upper Size Constraint

The maximum size for any stratum, k is calculated as:

MAXSIZE=MAXIN/SIZERATIO
where MAXIN is the maximum stratum size

SIZERATIO is a predetermined value with values generally between 0 and 1. We use either 3/4 or 5/6, the value of the SIZERATIO. SIZERATIO allows the size constraints to vary; the smaller the value of SIZERATIO, the more loose the size constraints will be.

[3] A self-representing (SR) PSU is treated as a separate stratum. They are usually the most populous PSUs in each state and are selected for sample with certainty. The remaining strata are formed by combining PSUs that are similar in characteristics such as unemployment, proportion of HUs with three or more persons, etc. The single PSU randomly chosen from each of these strata is NSR (non-self representing) because it represents not only itself but the entire stratum.

(2) Lower Size Constraint

The lower size constraint for any stratum is given by:
MINSIZE= MININ* SIZERATIO
where, MININ is the minimum stratum size.

6. Numerical Results

Results based on the above mentioned criterion functions were compared in this paper. We simulated and tested the 2000 FR algorithm for stratification on all 42 NSR PSU states using four stratification variables that are important to the labor force: number of female head households; 3+ persons households; number of unemployed females; and number of unemployed males. It is worth mentioning that female head households and 3+ persons households were seen to be highly correlated with the labor force variables. The stratification variables were taken from the 2000 decennial census short and long form data. We studied exactly the same number of strata for each state that were used in the 2000 design. Using common random starts in each state, each of the above four criteria was used to cluster NSR PSUs into strata for each state. We used the FR algorithm, five random numbers and one iteration of both the hill climbing and exchange procedures to find a locally optimal stratification.

Tables 1, 2, and 3 present the between-PSU variances that were obtained from the four criteria mentioned above. For purposes of comparison, the percent variance reductions nationally in the between-PSU variance for each variable under each criterion are shown in Table 1. These percent variance reductions are based on the between-PSU variance that are obtained from the above mentioned criteria and the unstratified between-PSU variance where the number of g PSUs were selected with PPS with replacement.

In Table 1, the largest reductions (44%) in between-PSU variances for the civilian labor force occurred with the 2000 criterion. Largest reductions (75% and 95%) for the female head households and 3+ persons households respectively occurred also with the 2000 criterion. On the other hand, the largest reductions (67%, and 29%) in between-PSU variances for the variables unemployed, and unemployment rate respectively occurred with the minimum variance criterion. We noticed that the 2000 criterion worked slightly better for 70% of the states compared to 67% of the states under the minimum variance criterion for the unemployed black . Overall, the 2000 criterion produced the best stratification for the larger characteristics that are highly correlated with the labor force characteristics and the minimum variance criterion generally produced the best stratification for most labor force characteristics.

Table 1: Percent Reduction in National Between-PSU Variance

Variable	*Wilks lambda*	*Hotelling trace*	*Minimum variance*	*2000 Criterion*
Civilian Labor Force	34%	21%	34%	**44%**
Female Head Households	56%	13%	68%	**75%**
3+ Person Households	75%	55%	39%	**95%**
Unemployed	52%	10%	**67%**	60%
Unemployed Black	48%	10%	67%	**70%**
Unemployment Rate	23%	5%	**29%**	**29%**

Table 2 shows that when compared to other criteria, the 2000 criterion produced significant variance reductions for most states (36%) for the female head households and 93% of the states for 3+ persons households. The minimum variance criterion produced largest variance reductions in 34% of the states for the civilian labor force, 69% of the states for unemployed, 41% of the states for black unemployed, and 74% of the states for unemployment rate. Table 2 also shows that the 2000 criterion worked best for the characteristics that are highly correlated with labor force characteristics and the minimum variance criterion produced the best stratification for all labor force characteristics under consideration.

Table 2: Percent of States with Best Between-PSU Variance

Variable	*Wilks Lambda*	*Hotelling trace*	*Minimum variance*	*2000 Criterion*
Civilian Labor Force	26%	10%	**33%**	31%
Female Head Households	31%	5%	29%	**36%**
3+ Person Households	5%	2%	0%	**93%**
Unemployed	7%	0%	**69%**	24%
Unemployed Black	19%	5%	**40%**	36%
Unemployment Rate	14%	7%	**74%**	33%

Table 3 shows that under the 2000 criterion, the variance reductions of 38% the states are equal to or more than the average state variance reduction; and the minimum variance criterion produces the same amount of variance reductions for 34% of the states for the civilian labor force. Since 38% and 34% are close to each other, we can reasonably say that the minimum variance criterion may work as good as the 2000 criterion for the civilian labor force. For the Unemployed characteristic, under the minimum variance criterion, 63% of the states have variance reductions and under the 2000 criterion 51% of the states have variance reductions that are equal to or more than the average state variance reduction. We also notice that the minimum variance criterion works better for all labor force characteristics under consideration. On the other hand, the 2000 criterion works better for the characteristics that are highly correlated with the labor force characteristics.

Table 3: Average State Between-PSU Variance Reduction

Variable	*Wilks lambda*	*Hotelling trace*	*Minimum variance*	*2000 Criterion*
Civilian Labor Force	29%	19%	34%	**38%**
Female Head Households	49%	23%	55%	**62%**
3+ Person Households	69%	60%	32%	**92%**
Unemployed	42%	11%	**62%**	51%
Unemployed Black	31%	13%	**44%**	42%
Unemployment Rate	28%	14%	**39%**	33%

7. Conclusion

Based on the comparison study, we see that the minimum variance and the 2000 criterion outperform the Wilks' lambda and Hotelling trace criteria in both state and national measures; the 2000 criterion generally outperforms the minimum variance criterion for the variables that are highly correlated with labor force characteristics; and the minimum variance criterion outperforms the 2000 criterion for labor force characteristics. We conclude that the minimum variance criterion produces the best stratification for CPS if no size constraint is imposed in strata and the number of strata is fixed.

8. Future Research

It is worth mentioning that based on previous work done by many authors, we suspect that the Wilks lambda criterion may work better than any other criterion if size constraints are imposed and strata sizes are allowed to vary. We are planning to incorporate size constraints into the algorithm to verify this assumption.

The variables being used to determine the clusters should be standardized otherwise, the contribution to each variable to the criterion may differ.

References

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

Korhonen P.K., (1978), "Experiments with Cluster Analysis Criteria based on the within Groups Scatter Matrix," Comstate 1978, Proc. In Computational Statistics, 3rd Symposium held in Leiden 1978, Physica-Verlag, Wien, 266-272.

Kostanich D.L, Judkins D.R, Singh R.P, and Mindi S., (1981),"Modification of Friedman-Rubin's Clustering Algorithm for use in stratified PPS Sampling." Paper presented at the 1981 American Statistical Association Meetings.

Scott A.J. and Symons M.J, (1971),"Clustering Methods Based on Likelihood Ratio Criteria," Biometrics 27, No.2, 387-398.