Maximizing Retention of Primary Sampling Units (PSUs) in a Two-PSU Per Stratum Design

Jay J. Kim and Danielle R. Corteville, Bureau of the Census
Patrick E. Flanagan, Bureau of Transportation Statistics
Jay Kim, SRD, Bureau of the Census, Washington D.C. 20233
jay.jong.ik.kim@census.com

Key Words : Primary Sampling Units, Redesign, PSU overlap

I. Introduction[1]

The Bureau of the Census redesigns its surveys every ten years after its decennial population census. This is to capture in the new design the changes that occurred during the past decade in the demographic, geographic and economic status of the population. The Bureau has been constructing the infrastructure for designing the 2000 surveys since mid-90. Its demographic surveys use multi-stage designs. In the first stage, the primary sampling units (PSUs) are stratified and one or two PSUs are selected from each stratum. In the second stage, ultimate sampling units are chosen from each selected sample PSU.

In selecting PSUs, there is an advantage of retaining as many 1990 sample PSUs as possible in the 2000 sample. The reasons are as follows. First, by using in the 2000 field operations the field representatives (field reps) experienced in the 1990 surveys rather than newly hired ones, we can control nonsampling errors better. Second, we can save costs associated with survey operations by not spending up to $5,000.00 for training a new hire. In order to maximally retain the 1990 field reps, we have to pick the same sample PSUs in the 1990 and 2000 redesigns to the maximum level. This means the 2000 PSU selection will be conditional on whether the PSU being considered was the 1990 sample PSU or not. Note that the (unconditional) selection probability for each PSU in a stratum is based on the size of the PSU. In general PSUs are selected based on the size or estimated size, i.e., sampling based on the probability proportional to size (PPS) or sampling with the probability proportional to estimated size (PPES) is used for selecting PSUs. Thus even if we try to maximize the retention level of the 1990 PSUs, we have to maintain the same 2000 unconditional selection probability for each PSU.

Census Bureau maximizes the PSU overlap between the 1990 and 2000 redesigns for the Current Population Survey (CPS), National Crime Victimization Survey (NCVS) and Survey of Income and Program Participation (SIPP). CPS and NCVS are one PSU per stratum (PSU/stratum) designs. SIPP uses a two PSUs/stratum design. This paper concerns itself with two PSUs/stratum design for SIPP.

Keyfitz (1951) considered this problem for one PSU/stratumdesign and obtained a limited solution. The situation he considered was that the composition of the strata in terms of PSUs remains the same over two designs (they will be called "initial and current sample," respectively) and only the sizes of PSUs change. Raj (1968) showed that Keyfitz's problem can be reformulated to a linear programming problem. Causey, Cox and Ernst (1985) extended it to a very general situation and formulated it as a transportation problem, a special case of linear programming. Causey, Cox and Ernst approach can be used for more than one PSU/stratum design. This approach assumes that the initial sample of PSUs is selected independently from stratum to stratum. Also when both the initial and current samples pick two PSUs per stratum and the number of PSUs in a stratum in the current design is large, large computer space is needed to run the linear programming software. Because of the second constraint, Ernst and Ikeda (1995) developed a reduced size transportation algorithm for SIPP which picked two PSUs per stratum. In the case of two PSUs per stratum, the linear programming problem can become as large as $2^n \times \binom{n}{2}$, where n is the number of overlapping PSUs. In their algorithm, they reduced the problem to $[\binom{n}{2} \% n \% 1] \times \binom{n}{2}$. This algorithm can not be used if there is no independence in selecting PSUs from stratum to stratum. Ernst (1986) developed an algorithm which does not require independence. The approach in this paper calculates the joint probability of selection of PSUs in the initial sample as if there is complete independence. Ernst's algorithm (1986) is used

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

for this project.

II Review of Ernst's Procedure (1986)

We pick PSUs stratum by stratum. Let S denote the stratum at hand in the current design, from which we select two sample PSUs. We assume there are n ($n \geq 2$) PSUs in S, which is denoted by $s_1, s_2, ..., s_n$. Thus we have $S = \{s_1, s_2, ..., s_n\}$. Once the PSUs in S are identified, we trace them back to the initial sample. Suppose they are from r strata in the 1990 design. We denote the strata by $T_1, T_2, .., T_r$. We denote by y_i the probability that T_i is associated with S. We will have at least one overlapping (common) PSU[2] between S and T_i. Let I_{ij}[3], $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, u_i$ be a PSU or pair of PSUs in $S \cap T_i$. Note that the PSUs in $S \cap T_i$ do not need to have been selected in the initial sample. We simply consider all overlapping PSUs between S and T_i. We denote by p_{ij} the probability that $I_i = I_{ij}$ where I_i is the possible outcome[4] in the initial design. For the current design, we denote $N_1, N_2, .., N_K$ for all possible pairs of PSUs[5] which can be formed from $S = \{s_1, s_2, ..., s_n\}$. Note $N_k = \{N_{k1}, N_{k2}\}$, as we deal with two PSUs per stratum. We denote by p_k the probability that $N = N_k$, where N is the actual outcome in the current design and Durbin (1967)-Brewer (1963) formula is used to calculate p_k. We also define

$$x_{ijk} = P(T = T_i, I_i = I_{ij}, N = N_k) \qquad (1)$$

where x_{ijk} is the joint probability that the initial stratum selected is T_i, the overlapping PSU(s) between S and T_i is I_{ij} and the pair of PSUs selected from S is N_k.

We define c_{ijk} as the conditional expected number of PSUs in N_k that were in the initial sample given that $T = T_i$ and $I_i = I_{ij}$. In the context of linear programming we will call it the cost. Then the objective function of linear programming is

Maximize

$$\sum_{i=1}^{r} \sum_{j=1}^{u_i} \sum_{k=1}^{K} c_{ijk} x_{ijk} \qquad (2)$$

which is the unconditional expected number of overlapping PSUs in the initial and current designs. As this is a linear programming problem, we need constraints which are

$$\sum_{i=1}^{r} \sum_{j=1}^{u_i} x_{ijk} = p_k, \qquad k = 1, 2, 3, \ldots, K$$

$$i = 1, 2, 3, \ldots, r; \; j = 1, 2, 3, \ldots, u_i$$

$$\sum_{k=1}^{K} x_{ijk} = y_i p_{ij}, \quad i = 1, 2, 3, \ldots, r; \; j = 1, 2, 3, \ldots, u_i$$

$$\sum_{i=1}^{r} y_i = 1$$

Once x_{ijk} is determined which maximizes the objective function above, we will select N_k depending upon the sampling situation in the initial design. That is,

$$P(N = N_k | T = T_i, I_i = I_{ij})$$
$$= \frac{P(T = T_i, I_i = I_{ij}, N = N_k)}{P(T = T_i, I_i = I_{ij})} = \frac{x_{ijk}}{y_i p_{ij}}. \qquad (3)$$

Suppose I_{ij_i} was selected from stratum T_i in the initial design. Note in the two PSU/stratum design, I_{ij_i} can be null set ($\emptyset$), a singleton or a pair of PSUs in terms of overlapped PSUs in S and T_i. The probabilities that we need for selecting PSUs in 2000 are

$$P(N = N_k | I = I_{1j_1}, \ldots, I_r = I_{rj_r})$$

By Laplace's rule of succession (see Ross, 1994), the above equation becomes,

$$\sum_{i=1}^{r} y_i P(N = N_k | T = T_i, I_i = I_{ij_i}) = \sum_{i=1}^{r} \frac{x_{ij_ik}}{p_{ij_i}} \qquad (4)$$

We define for two PSUs/stratum design,

$$c_{ijk} = \sum_{h=1}^{2} p''_{ijkh},$$

where

$$p''_{ijkh} = \begin{cases} 1, & \text{if } N_{kh} \in I_{ijt} \\ 0, & \text{if } N_{kh} \in T_i - I_{ijt} \end{cases} \qquad (5)$$

---

[2] Ernst says in his paper that it is "sample" PSUs, but it could be sample PSUs or non-sample PSUs.

[3] Since PSU j is nested within stratum i, we could denote it by j(i) following the notation commonly used in the experimental design.

[4] I_{ij} is a possible outcome, rather than an actual outcome, as we consider all possible pairs of PSUs in $S \cap T_i$ including the null set. As to be seen in the next section, PSUs in T_i which are not in $S \cap T_i$ are also included in forming I_{ij}. In other words, they are not completely ignored.

[5] Ernst says it is "sample" PSUs, but it could be sample PSUs or non-sample PSUs.

$$p^{*}_{kh} \qquad \text{otherwise.}$$

In the above p^{**}_{ijkh} is the conditional probability N_{kh} was in the initial sample given $T=T_i$ and $I_i = I_{ij}$, and p^{*}_{kh} is the unconditional selection probability of N_{kh} in the initial design.

In two PSUs/stratum design, there are two ways of computing c_{ijk}. One way is basically based on one PSU/stratum design, but as it is a two PSU/stratum design, the selection probability is doubled as shown in section III and the other way is based on two PSUs/stratum approach.

The PSU definitions can change over two censuses. Thus the PSUs can be partially overlapped between the censuses. Thus the component of the cost formula shold be revised. For the new formula, see page 198 of Ernst.

III Two Approaches of Computing p^{*}_{kh} in a Two PSUs/Stratum Design

Approach 1.

Let $T_1 = \{s_1,s_2,,,,s_J\}$. Let $p_j = P(s_j)$, which is the probability of selecting one PSU with the probability proportional to estimated size (PPES) in the initial stratum T_1. That is,

$$p_j = \frac{MOS(s_j)}{\sum_j MOS(s_j)},$$

where $MSO(s_j)$ is the measure of size of PSU s_j. Ernst's p^{*}_{kh} is the unconditional probability that N_{kh} is in the initial design. This probability is calculated in the initial stratum using the initial design's MOS. We assume the subscripts "kj" in p^{*}_{kj} point to the PSU "j" stratum T_1 in the initial design. Then as it is a two PSUs/stratum design, the selection probability will be $2p_j$, which is usually denoted by p_j.

Approach 2.

Before this approach is discussed, we will deal with all possible sampling situations when two PSUs are selected from a 1990 stratum.

Example 1

Suppose $S\_T_1 = \{s_1,s_2\}$ and $T_1 = \{s_1,s_2,s_5,s_6\}$. Ernst's procedure requires computing the probability of selecting in the initial design a pair (in this example, s_1 and s_2), a singleton (a singleton means s_1 and s_2 in this example, but actually selecting a singleton s_1 means selecting either s_1 and s_5 or s_1 and s_6, as pairs are candidates of selection) and null set (in this example, selecting null set means selecting the pair of s_5 and s_6). All the possible sampling situations in the initial design are

i j	PSUs	Prob
1 1	$\{s_1,s_2\}$	p_{12}
1 2	$\{s_1\}$	$p_1 - p_{12}$
1 3	$\{s_2\}$	$p_2 - p_{12}$
1 4	$\{\emptyset\}$	$1 - (p_1 + p_2 - p_{12})$
SUM		1

Example 2

Suppose $S\_T_1 = \{s_1,s_2,s_3\}$ and $T_1 = \{s_1,s_2,s_3,s_5,s_6\}$. All the possible sampling situations in 1990 are

i j	PSUs	Prob
1 1	$\{s_1,s_2\}$	p_{12}
1 2	$\{s_1,s_3\}$	p_{13}
1 3	$\{s_2,s_3\}$	p_{23}
1 4	$\{s_1\}$	$p_1 - p_{12} - p_{13}$
1 5	$\{s_2\}$	$p_2 - p_{12} - p_{23}$
1 6	$\{s_3\}$	$p_3 - p_{13} - p_{23}$
1 7	$\{\emptyset\}$	$1 - (p_1 + p_2 + p_3 - p_{12} - p_{13} - p_{23})$
Sum		1

In the above example, p_1 is $2p_1$, as shown in approach 1. By adding p_{12} and p_{13} to the above probability of selecting s_1 in the two PSUs/stratum design case, we can have the same p_1 as observed for approach 1. p_{12} and p_{13} are all the joint probabilities involving PSU s_1 in S_T_1. We can make similar observations concerning probabilities of selecting s_2 and s_3. This can be generalized as follows. Assuming the subscripts "kh" in p^{*}_{kh} point to the PSU N_{kh} in stratum T_1 in the initial design, we can express p^{*}_{kh}

$$p^{*}_{kh} = P(N_{kh}) + \sum_{h} \sum_{l \ldots h} P(N_{kh,l})$$

where p_j is the same probability as in approach 1, $P(s_j)$ is the probability of selecting s_1 in the context of 2 PSU/stratum design and $P(s_j,s_l)$ is the joint probability of selection of PSUs s_j and s_l which were in S_T_1.

In general, the sampling situation in the initial design can be summarized as follows. Let S_T_1 has c PSUs, i.e., $S\_T_1 = \{s_1,s_2,\ldots,s_c\}$ and let $T_1 = \{s_1,s_2,..,s_c,\ldots,s_m\}$. All the possible sampling situations in the initial design include $\binom{c}{2}$ PSU pairs, c singletons and a null set which combines all $\binom{m-c}{2}$ PSU pairs. That only one PSU (e.g. s_1) is selected in the sample in the initial design from among the PSUs in $S \cap T_1$ means that i) one of the PSUs of the pair is from the PSUs which are not in $S \cap T_1$ but in T_1 or ii) we are dealing with the portion of the PSU excluding the portions which are shared with other PSUs in $S \cap T_1$. The first of the above means that

$$P(s_1) = \sum_{i=1}^{m-c} P(s_1, s_{c+i})$$

The second of the above means that

$$P(s_1) = p_1 - \sum_{i=2}^{c} P(s_1, s_i)$$

Theorem. Probability of s_1 excluding the portions which are shared with other PSUs in $S \cap T_1$ is equivalent with the probability of s_1 sharing with other PSUs which are not in $S \cap T_1$. That is,

$$\sum_{i=1}^{m-c} P(s_1, s_{c+i}) = p_1 - \sum_{i=2}^{c} P(s_1, s_i).$$

Proof. Since p_1 is the marginal probability of s_1 in the joint probability distribution of $(s_1, s_i)_{i \neq 1}$

$$\sum_{i=2}^{m} P(s_1, s_i) = p_1 .$$

The left hand side of the equation can be broken down into two terms, that is

$$\sum_{i=2}^{m} P(s_1, s_i) = \sum_{i=2}^{c} P(s_1, s_i) + \sum_{i=c+1}^{m} P(s_1, s_i)$$

However, the second term on the right-hand side of the equation can be re-expressed as

$$\sum_{i=c+1}^{m} P(s_1, s_i) = \sum_{i=1}^{m-c} P(s_1, s_{c+i}).$$

This proves the theorem.

Durbin-Brewer formula is used for computing the joint probability.

Even if the two approaches provides the same cost, two PSU/stratum approach involves more terms, and requires more calculations. Thus approach 1 is preferable in actual use.

IV Test Runs

In order to test the methodologies set in place for the 2000 redesign, a test data set was created for SIPP. As the 2000 census data were not available at the time of test runs, the 1990 census data was used. However, in order to be more realistic, the 2000 geography was used. PSUs based on the this data set were stratified and stratum data were used for the PSU maximum overlap test runs.

The Demographic Statistical Methods Division (DSMD), which designs the surveys, purchased SUNSET Software for the linear programming work. More specifically, SUNSET Software was used to solve x_{ijk}. As the constraints of the linear programming problem are equalities and rounding errors are involved in calculation of probabilities, room for rounding error should be allowed. Thus a tolerance level should be set. Table 1 shows the number and percent of strata for which SUNSET Software successfully ran at different parameter settings. SUNSET Software initially set the tolerance at 9. At that level, the software ran successfully for only 28 percent of a total of 107 strata. Tolerance level of 9 means that any number which lies between $-e^{-9}$ (-.000123409) and e^{-9} (.000123409) is considered zero. This turns out to be too stringent. When it was relaxed to e^{-6}, that is, any number that lies between $-e^{-6}$ (-.00247875) and e^{-6} (.00247875), is regarded as zero, the percentage of the software running successfully more than doubled (61.68 percent). Relaxing further to e^{-5} did not help. Use of Devex pricing for avoiding near-zero pivots helped improve the success rate to 81.31 percent. As the original parameter settings did not allow us to read in large problems, when the settings were changed, it was able to handle all problems.

Table 1. Number of Strata Having Feasible Solution for Different Tolerance Level

Tolerance e^{-9}	Tolerance e^{-6}	Tolerance e^{-6} + Devex Pricing
30 strata	66 strata	87 strata
28 %	62 %	81 %

In the beginning, another LP software CPLEX maintained by Statistical Research Division (SRD) was to be used to verify SUNSET Software's solutions of the x_{ijk}'s. However, since DSMD was not able to run SUNSET Software on many strata, CPLEX was used to check whether it can run on the same strata. Note CPLEX has tolerance of e^{-6} as a default. It could run on any stratum except 2 extremely large strata without modifying parameter settings. Experience with CPLEX helped us modify the parameter settings for SUNSET Software. We also compared the solutions provided by those two softwares. They were identical for stratum 421005 (see Kim, 2001). However, for many strata, they were different. For example, solutions provided by SUNSET Software and CPLEX for stratum 531002, are quite different. SUNSET Software provided 49 nonzero solutions for the x_{ijk}'s, but CPLEX 61 ones. Only five solutions are exactly same for the same variables, three variables have different solutions and, for the rest, different variables or combinations of different variables have the same solutions.

Note all the LP problems we faced had more unknowns in the objective function than the number of equations in the constraints. Thus we were in a multiple solutions situation.

Once the x's are solved, we pick a sample of PSUs, N_k, conditional on the 1990 sampling situation (using equation 4). The sampling situation can mean that, for example, one PSU (PSU 10021) was selected in 1990 sample from T_1, 2 PSUs (PSUs 10032 and 10034) were selected in 1990 sample from T_2 and none (i) were selected in the 1990 sample from T_3.

Four steps are needed to pick the sample PSUs.

Step 1. Compute the conditional probability in equation 4 for each k;
Step 2. Compute cumulative probability over k using the probabilities in step 1;
Step 3. Generate a random number; For SIPP we used SAS ranuni(seed) routine and seed was obtained by stratum number x 1,000 + 2.
step 4. Pick a k, that is, a pair of PSUs.

Note this is sampling with probability proportional to estimated size (PPES). This is PPES because p's, 1990 selection probabilities, are based on the projected 1995 MOS' and p 's are based on the projected 2005 MOS'.

It is well-known that every linear programming problem, called the *primal* problem, has associated with it another linear programming problem, called the *dual* problem (Hillier-Lieberman, 1972). We ran CPLEX using both options on stratum 421005. Most of the Primal and Dual solutions of x's were the same, except for three cases. The Primal solution for one variable is shared by two variables in Dual. The Dual solution for one variable is divided into solutions for two variables. However, the conditional probabilities of selection in equation 4 for the stratum were same for all k's and thus the same PSUs were selected disregarding whether we use primal or dual solutions or which random number we use for picking the PSUs.

As projected 1995 MOSs were used for 1990 design and the projected 2005 MOSs are used for 2000, the parameters of the LP problems are subject to errors. Thus it is instructive to perform some sensitivity analyses to determine the effect of errors in the population size on the optimal solution of revised parameter values. Minor sensitivity analyses on the solutions for stratum 421005 were performed. Originally the coefficient of x_{121} was .76814. When it was raised to .76825 (an increase of .00011), no changes were observed in optimal solution and thus the optimal value of the objective function. When it was raised to .78000 (an increase of .01186), solution values for three x's (out of 228 x's) were changed and the optimal value of the objective function changed slightly. As we do not know how good the estimates are, we may have to do sensitivity analysis by changing all p values in the objective function and all p values in the constraints.

V. Percentage of Overlapped PSUs

The percentage of the 1990 SIPP sample PSUs selected again in the 2000 sample redesign test runs is 49.21. Since there are multi-county PSUs in both 1990 and 2000 redesigns, PSU configurations can be different between the 1990 and 2000 designs and the percentage of the 1990 SIPP sample counties which are retained in 2000 was computed, which is 52.06. Comparing with Ernst’s 56 percent for one PSU/stratum design, we find them slightly lower. The reasons for this can be two-fold. First, the 1990 SIPP design was region-based design, but the 2000 SIPP employs State-based design. In the region-based design, PSUs or counties in different states can be in the same stratum. Thus there could be states, especially small states, from which no sample PSUs or sample counties were selected in 1990. However, in a state-based design, each state will have at least two PSUs selected in the sample. In those states, we end up picking PSUs which were not the 1990 sample PSUs or counties. Second, there could be some peculiarity of the strata. Visual inspection shows that in some strata, there are six1990 sample PSUs in a stratum. This means we will miss at least four 1990 sample PSUs in the 2000 redesign.

VI. Concluding Remarks

This is the first time at the Bureau we implemented Ernst’s 1986 approach for maximally overlapping PSUs in two designs in two PSUs/stratum design. We created a test data set and tested this approach on the data set. Two ways of computing the cost associated with this problem are shown. SUNSET Software, which was purchased by DSMD for running linear programming as this procedure involves linear programming, was tried on the data. In the beginning, it did not run on more than 70 percent of the strata. By running CPLEX separately on the same data set for which SUNSET Software did not run, we got clues why SUNSET Software did not run on the strata. Thus by changing the parameter settings for SUNSET Software, We were able to run SUNSET Software on all strata. In rare cases, SUNSET Software and CPLEX provided the identical solutions for x's, but in most cases different solutions. We investigated whether these different solutions led to picking different pairs of PSUs or not. We also investigated whether or not dual procedure provided the same solutions and the same pairs of PSUs. We also did a sensitivity analysis of the solutions by changing a coefficient of the objective function, as the coefficient could represent a probability which is subject to error because projected counts are used as MOS. It should be noted that even if the solutions were different for some or many variables, we ended up picking the same pair of PSUs given the random numbers used. However, depending on the selected random numbers used for picking PSUs, we could end up with different results.

From our test runs, only around 50 percent of the 1990 sample PSUs were retained in the 2000 redesign, which is lower than 56 percent, the rate of retaining the 1980 CPS PSUs in the 1990 CPS design, which is one PSU/stratum design. The reason for this might be that in the 1990 design, SIPP selected PSUs from strata which could cross state boundaries (but not region boundaries), but in the 2000 design, PSUs and strata are defined within the state boundaries. That is, in the 1990 design, there could be some states which did not have any sample PSUs, but in the 2000 design, they will have at least two sample PSUs. Thus in those states, no maximum overlap could occur.

For more detailed version of this paper, see Kim et al (2002).

VII. References

1. Brewer, K.W.R. (1963) A Model of Systematic Sampling with Unequal Probabilities. Australian Journal of Statistics Vol 5, pp 5-13.
2. Cochran, W.G. (1977) Sampling Techniques, Third Edition, John Wiley and Sons.
3. Durbin, J. (1967) Design of Multi-Stage Surveys for Estimation of Sampling Errors. Applied Statistics, Vol. 16, pp 152-164.
4. Ernst, L.R. (1986) Maximizing the Overlap between Surveys When Information Is Incomplete. European Journal of Operational Research Vol 27, No. 2, pp. 192-200.
5. Ernst, L.R. and Ikeda, M.M. (1995) A Reduced-Size Transportation Algorithm for Maximizing the Overlap between Surveys. Survey Methodology. Vol. 21, No. 2, pp. 147-157.
6. Hillier, FS and Lieberman, G.J. (1972) Introduction to Operations Research. Holden-Day, Inc
7. ILOG (1999) ILOG CPLEX 6.5 User's Manual, ILOG.
8. Kim, J.J. (2000) Equivalency of Two Cost Formulae. Internal Census Bureau Memorandum.
9. Kim, J.J. (2001) Some Observation About Solutions of LP and Selection of PSUs - Stratum 421005. Internal Census Bureau Memorandum.
10. Kim, J.J. (2001) Some Observations About Solutions of LP and Selection of PSUs - NCVS Strata 11004 and 81007. Internal Census Bureau Memorandum.
11. Kim, J.J. (2001) Some Observations Concerning Primal and Dual Solutions of LP and Selection of PSUs - Stratum 531002. Internal Census Bureau Memorandum.
12. Kim, J.J., Corteville, D.R. and Flanagan, P.E. (2002) Maximizing Retention of Primary Sampling Units (PSUs) in a Two-PSU Per Stratum Design. Statistical Research Division Report Series
13. Ross, S. (1994) A First Course in Probability. Fourth Edition. McMillan College Publishing Company.