RESEARCH REPORT SERIES
(*Statistics #2002-10)*

Maximizing Retention of Primary Sampling Units in a Two-Primary Sampling Unit Per Stratum Design

Jay Kim, Danielle Corteville, Patrick Flanagan

Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: December 4, 2002

Maximizing Retention of Primary Sampling Units in a Two-Primary Sampling Unit Per Stratum Design[1]

Jay J. Kim, Statistical Research Division, Bureau of the Census
Danielle Corteville, Demographic Statistical Methods Division
Patrick Flanagan, Bureau of Transportation Statistics

Abstract

In the 2000 redesign of the Survey of Income and Program Participation sample, a two-stage sample design is adopted. In the first stage, two geographic Primary Sampling Units are selected from each of the strata within each state. In selecting Primary Sampling Units, if a Primary Sampling Unit that was in the 1990 design is reselected in 2000, then the experienced field representative in that Primary Sampling Unit can remain working on the survey. If a new Primary Sampling Unit is selected, then a new field representative will have to be hired and trained. Therefore, reselecting as many of the Primary Sampling Units that were in the 1990 design as possible would minimize this turnover of field representatives. This will help reduce nonsampling errors caused by the inexperience of newly hired field representatives and the costs to train them. The Bureau of the Census employs the Ernst (1986) algorithm to select Primary Sampling Units while maximizing the Primary Sampling Unit "overlap" between the 1990 and 2000 designs. Ernst's approach (1986) is demonstrated on test data for selecting two Primary Sampling Units from each stratum. The test results are reported in this paper.

Key Words : Primary Sampling Units, Redesign, PSU Overlap

I. Introduction

The Bureau of the Census redesigns its surveys every ten years after its decennial population census to capture changes in the demographic, geographic and economic status of the population. The Bureau has been building the infrastructure for designing the 2000 surveys since the mid-90s. Its demographic surveys use multistage 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[1] are chosen from each sample PSU.

In selecting PSUs, there are advantages of retaining as many 1990 sample PSUs as possible in the 2000 redesign. They 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[2] for training each new hire. In order to maximally retain the

[1] There could be subsampling. In that case it becomes three-stage design.

[2] This is an estimate provided by Richard Bitzer, Field Division, Bureau of the Census.

1990 field reps, we have to maximize the overlap between the sample PSUs in the 1990 and 2000 redesigns. This means the 2000 PSU selection will be conditional on whether the PSU being considered was a 1990 sample PSU or not. Note that the (unconditional) selection probability for each PSU in a stratum is based on the estimated size of the PSU. Thus even if we try to maximize the retention level of the 1990 PSUs, we have to maintain the same 2000 unconditional selection probability (based on the 2005 estimated size) for each PSU.

The Census Bureau maximizes the PSU overlap between the 1990 and 2000 designs 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/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/stratum designs 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 designs," respectively) and only the sizes of the 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. The 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 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 cannot be used if there is no independence in selecting PSUs from stratum to stratum. Ernst (1986) developed an algorithm which does not require the above independence. This approach calculates the joint probability of selection of PSUs in the initial sample as if there is complete independence.

II Review of Ernst's Procedure (1986)

Let S denote a stratum in the current design. Note we select sample PSUs from S each time we face a new stratum. We assume there are n ($n \geq 2$) PSUs in S, which are denoted by $s_1, s_2, .. s_n$. Thus, we have $S = \{s_1, s_2, .. s_n\}$. Once we see the PSUs in S, we trace them back to the initial sample. Suppose they are from r strata. We denote the strata by $T_1, T_2, .., T_r$. Let y_i be the probability that T_i was selected in the initial sample. We will have at least one overlapping

(common) PSU[3] between S and T_i. Let I_{ij}[4], i = 1, 2, . . . , r; j = 1, 2, . . . , u_i be a PSU or pair of PSUs in $S \cap T_i$[5]. Note that the PSU(s) in $S \cap T_i$ does (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 actual outcome in the initial design. For the current design, we denote $N_1, N_2, . . , N_K$ for all possible pairs of PSUs[6] 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 π_k the probability that $N = N_k$, where N is the actual outcome in the current design.

We also define

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

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 (are) I_{ij} and the pair of PSUs being considered from S is N_k.

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

Maximize

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

which is the unconditional expected number of overlapping PSUs in the initial and current designs. Constraints associated with this linear programming problem are

$$\sum_{i=1}^{r} \sum_{j=1}^{u_i} x_{ijk} = \pi_k, \qquad k = 1, 2, 3, \quad . . . , K \tag{3}$$

---

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

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

[5] Ernst's method is a two step process. First select a T_i, then selecting N_{kh} depends on which I_{ij} was selected in T_i.

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

$$\sum_{k=1}^{K} x_{ijk} = y_i p_{ij}^{7}, \qquad i = 1, 2, 3, \ldots . r;\ j = 1, 2, 3, \ldots u_i \tag{4}$$

$$\sum_{i=1}^{r} y_i = 1 \tag{5}$$

Note

$$P(N = N_k \mid T = T_i, I_i = I_{ij}) \tag{6}$$

$$= \frac{P(T = T_i, I_i = I_{ij}, N = N_k)}{P(T = T_i, I_i = I_{ij})} = \frac{x_{ijk}}{y_i p_{ij}}. \tag{7}$$

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

$$P(N = N_k \mid I_1 = I_{1j_1}, .. I_r = I_{rj_r})^{8} \tag{8}$$

Suppose I_{ij_i} was selected from the stratum T_i in the initial design. Note in the two PSUs/stratum design, I_{ij_i} can be null set ($\varnothing$), a singleton or a pair of PSUs. By Laplace's rule of succession (Ross, 1994), the above expression becomes,

$$\sum_{i=1}^{r} y_i P(N = N_k \mid T = T_i, I_i = I_{ij_i}) = \sum_{i=1}^{r} \frac{x_{ij_ik}}{p_{ij_i}} \tag{9}$$

We define for a two PSUs/stratum design,

$$c_{ijk} = \sum_{h=1}^{2} p''_{ijkh}, \tag{10}$$

where

---

[7] This constraint comes from the fact that the sum of the probabilities over all k's in equations (6) and (7) is 1.

[8] In I_{ij_i}, j_i is the specific sample from stratum i.

$$p''_{ijkh} = \begin{cases} 1, & \text{if } N_{kh} \in I_{ijt} \\ 0, & \text{if } N_{kh} \in T_i \sim I_{ijt} \\ p'_{kh} & \text{otherwise.} \end{cases} \quad (11)$$

and I_{ijt}, t=1,2,.....,v_{ij} denotes a PSU in I_{ij}.

In the above p''_{ijkh} is the conditional probability that PSU 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} (Kim, 2000). One way is basically based on one PSU/stratum design (see equation 10) but as it is a two PSUs/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 in equation (11) should be revised accordingly as follows.

Let I_{ijt}, t = 1, 2, , , v_{ij} denotes PSU in I_{ij}.

f_{ijtkh} denotes the proportion of I_{ijt} that is in N_{kh} based on the new measure of size. (If $I_{ij}=\emptyset$ let v_{ij}=1 and $f_{ij1kh} = 0 \ \forall$ k, h. Note v_{ij}=1 if it is a one PSU/stratum design and two if a two PSUs per stratum design). In general,

$$f_{ijtkh} = \frac{\text{2000 MOS of 2000 PSU } h \in k^{th} \text{ pair within 1990 PSU } t \in j^{th} \text{ pair}}{\text{2000 MOS of 1990 PSU } t \in j^{th} \text{ pair}}$$

Using f_{ijtkh}, p''_{ijkh} can be reformulated as follows.

$$p''_{ijkh} = 1 - [\prod_{t=1}^{v_{ij}} (1 - f_{ijtkh})] \times \prod_{q=1, q\neq i}^{r} \{1 - \sum_{w=1}^{u_q} p_{qw}[1 - \prod_{t=1}^{v_{qw}} (1 - f_{qwtkh})]\} \quad (12)$$

In the above equation, $\prod_{t=1}^{v_{ij}} (1 - f_{ijtkh})$ is the conditional probability given $T=T_i$ and $I_i = I_{ij}$. For $q \neq i$, $p_{qw}[1 - \prod_{t=1}^{v_{ij}} (1 - f_{ijtkh})]$ is the unconditional probability that $I_q = I_{qw}$ and that at least some part of PSUs in I_{qw} is common with N_{kh}. $1 - \sum_{w=1}^{u_q} p_{qw}[1 - \prod_{t=1}^{v_{ij}} (1 - f_{ijtkh})]$ is the unconditional probability that no PSU in T_q is common with N_{kh} .

III Two Approaches of Computing p_{kh}' in a Two-PSU/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 size (PPS) in the initial stratum T_1. That is,

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

where $MOS(s_j)$ is the measure of size of PSU s_j. Ernst's p_{kh}' is the unconditional probability that N_{kh} is selected 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 π_j.

Approach 2.

Before this approach is discussed, we will show two examples dealing with the probabilities of selection for the initial design, which are needed in the PSU selection process, when two PSUs are selected from a stratum.

Example 1

Suppose $S \cap T_1 = \{s_1, s_2\}$ 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 or 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 neither s_1 nor s_2, thus it means selecting the pair of s_5 and s_6). All the possible sampling situations in the initial design in conjuction with the current design are as follows.

i j	PSUs	Prob
1 1	$\{s_1, s_2\}$	π_{12}
1 2	$\{s_1\}$	$\pi_1 - \pi_{12}$
1 3	$\{s_2\}$	$\pi_2 - \pi_{12}$
1 4	$\{\varnothing\}$	$1 - (\pi_1 + \pi_2 - \pi_{12})$
Sum		1

Example 2

Suppose $S \cap T_1 = \{s_1, s_2, s_3\}$ and $T_1 = \{s_1, s_2, s_3, s_5, s_6\}$. All the sampling situations in 1990 will be

i	j	PSUs	Prob
1	1	$\{s_1,s_2\}$	π_{12}
1	2	$\{s_1,s_3\}$	π_{13}
1	3	$\{s_2,s_3\}$	π_{23}
1	4	$\{s_1\}$	$\pi_1 - \pi_{12} - \pi_{13}$
1	5	$\{s_2\}$	$\pi_2 - \pi_{12} - \pi_{23}$
1	6	$\{s_3\}$	$\pi_3 - \pi_{13} - \pi_{23}$
1	7	$\{\varnothing\}$	$1 - (\pi_1 + \pi_2 + \pi_3 - \pi_{12} - \pi_{13} - \pi_{23})$
Sum			1

In this example $\{s_1\}$ means s_1 alone, ignoring all the areas shared with other PSUs in $S \cap T_1$ such as s_2 and s_3. Note π_{12} is the probability of having both s_1 and s_2. Similarly π_{13} is the probability of having both s_1 and s_3. Since $P(s_1)$ in this example is the probability of having only s_1, ignoring all the areas shared by other PSUs in $S \cap T_1$, it is $\pi_1 - \pi_{12} - \pi_{13}$. Now since we want the unconditional probability of having s_1 (whole s_1) in the sample, we have to add the probabilities of the three samples which include s_1. This means that we have to add π_{12} and π_{13} to $P(s_1)$ above. Then we have the same π_1 as observed for approach 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}' as

$$p_{kh}' = P(N_{kh}) + \sum_{h} \sum_{z_{z \neq h}} P(N_{kh,z}) \tag{13}$$

where π_j is the same probability as in approach 1, $P(s_j)$ is the probability of selecting s_j alone in the context of two PSU/stratum design and $P(s_j,s_z)$ is the joint probability of selection of PSUs s_j and s_z which are in $S \cap T_1$.

In general, the sampling situation in the initial design can be summarized as follows. Let $S \cap T_1$ has c PSUs, i.e., $S \cap T_1 = \{s_1,s_2,\ldots,s_c\}$ and let $T_1 = \{s_1,s_2,..,s_c,\ldots,s_m\}$. All the possible sampling situations in the initial design in conjunction with the current design include $\binom{c}{2}$ PSU pairs, c singletons and a null set which includes all $\binom{m-c}{2}$ PSU pairs of PSUs not in $S \cap T_1$.

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 $S \cap T_1$ or ii) we are dealing with the portion of the PSU excluding the portions which are shared by other PSUs in $S \cap T_1$. The first of the above means that

$$P(s_1) = \sum_{i=1}^{m-c} P(s_1, s_{c+i}) \tag{14}$$

The second of the above means that

$$P(s_1) = \pi_1 - \sum_{i=2}^{c} P(s_1, s_i) \tag{15}$$

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

$$\sum_{i=1}^{m-c} P(s_1, s_{c+i}) = \pi_1 - \sum_{i=2}^{c} P(s_1, s_i).$$

Proof. Since π_1 is the marginal probability of s_1 in the joint probability distribution $P(s_1, s_i)$, i = 2,3, . . . , m,

$$\sum_{i=2}^{m} P(s_1, s_i) = \pi_1.$$

The left-hand side of the equation can be broken down into following two terms.

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

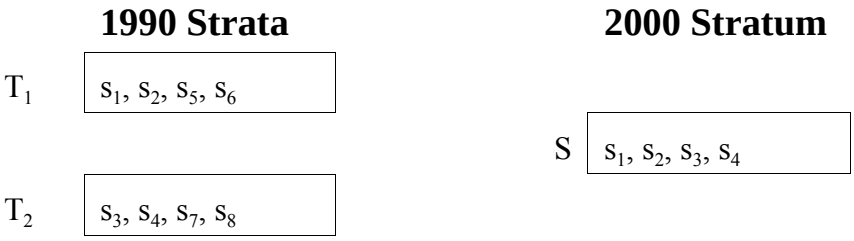
However, the second term on the right-hand side of the equation can be re-expressed as

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

This proves the theorem.

Following two tables show costs using two different approaches of computing the probability of selection.

Approach 1.



PSUs in S overlap PSUs in T_1 and T_2. We ignore prime and double primes.

We define, ignoring an initial stratum identifier but using unique PSU number, $\pi_i = 2P(s_i)$ for i=1,2,5,6 which are in S.

Table 1. Costs for stratum T1

		2000 PSU Selections					
		k=1	k=2	k=3	k=4	k=5	k=6
	1990	$\{s_1, s_2\}$	$\{s_1, s_3\}$	$\{s_1, s_4\}$	$\{s_2, s_3\}$	$\{s_2, s_4\}$	$\{s_3, s_4\}$
1	$\{s_1\}$	Cell 111	112	113	114	115	116
		$c_{111}=1$	$c_{112}=1+\pi_3$	$c_{113}=1+\pi_4$	$c_{114}=\pi_3$	$c_{115}=\pi_4$	$c_{116}=\pi_3+\pi_4$
2	$\{s_2\}$	Cell 121	122	123	124	125	126
		$c_{121}=1$	$c_{122}=\pi_3$	$c_{123}=\pi_4$	$c_{124}=1+\pi_3$	$c_{125}=1+\pi_4$	$c_{126}=\pi_3+\pi_4$
3	$\{s_1, s_2\}$	131	132	133	134	135	136
		$c_{131}=2$	$c_{132}=1+\pi_3$	$c_{133}=1+\pi_4$	$c_{134}=1+\pi_3$	$c_{135}=1+\pi_4$	$c_{136}=\pi_3+\pi_4$
4	$\{\varnothing\}$	141	142	143	144	145	146
		$c_{141}=0$	$c_{142}=\pi_3$	$c_{143}=\pi_4$	$c_{144}=\pi_3$	$c_{145}=\pi_4$	$c_{146}=\pi_3+\pi_4$

As mentioned before, the singletons are, actually, sums of pairs of PSUs, one of which is a non overlapping PSU between S and T_1. For example, in the above, $\{s_1\} = \{s_1 \cap s_5\} \cup \{s_1 \cap s_6\}$, where s_5 and s_6 are non overlapping PSUs between S and T_1, because they are in T_2. Similarly, the null set is the sum of all non-overlapping PSU pairs. In T_1, $\{\varnothing\}$ is equivalent with $\{s_5 \cap s_6\}$.

Approach 2. Two PSUs/stratum approach

We will ignore the initial stratum identifier but use PSU number for denoting probability. We will let $\pi_j' = \pi_j - \pi_{j,z_{z\neq j}}$, where s_z is in $S \cap T_1$.

Table 2. Costs for stratum T_1

		2000 PSU Selections					
		k=1	k=2	k=3	k=4	k=5	k=6
	1990	$\{s_1, s_2\}$	$\{s_1, s_3\}$	$\{s_1, s_4\}$	$\{s_2, s_3\}$	$\{s_2, s_4\}$	$\{s_3, s_4\}$
1	$\{s_1\}$	Cell 111	Cell 112	Cell 113	Cell 114	Cell 115	Cell 116
		$c_{111}=1$	$c_{112}= 1+\pi_3'+\pi_{34}$	$c_{113}= 1+ \pi_4'+\pi_{34}$	$c_{114}= \pi_3'+ \pi_{34}$	$c_{115}=\pi_4'+ \pi_{34}$	$c_{116}=\pi_3'+\pi_4'+2 \pi_{34}$
2	$\{s_2\}$	Cell 121	122	123	124	125	126
		$c_{121}=1$	$c_{122}=\pi_3' + \pi_{34}$	$c_{123}=\pi_4' + \pi_{34}$	$c_{124}= 1+ \pi_3'+\pi_{34}$	$c_{125}= 1+\pi_4'+\pi_{34}$	$c_{126}=\pi_3'+\pi_4'+2 \pi_{34}$
3	$\{s_1, s_2\}$	131	132	133	134	135	136
		$c_{131}=2$	$c_{132}=1+\pi_3' + \pi_{34}$	$c_{133}=1+\pi_4' + \pi_{34}$	$c_{134}=1+\pi_3' + \pi_{34}$	$c_{135}=1+\pi_4' + \pi_{34}$	$c_{136}=\pi_3'+\pi_4'+2 \pi_{34}$
4	$\{\varnothing\}$	141	142	143	144	145	146
		$c_{141}=0$	$c_{142}=\pi_3'+ \pi_{34}$	$c_{143}=\pi_4'+ \pi_{34}$	$c_{144}=\pi_3'+ \pi_{34}$	$c_{145}=\pi_4'+\pi_{34}$	$c_{146}=\pi_3'+\pi_4'+2 \pi_{34}$

Note in the above table, $\pi_3' = \pi_3 - \pi_{34}$. In cell 112, the cost thus will become $1 + \pi_3 - \pi_{34} + \pi_{34} = 1 + \pi_3$. In cell 126, $\pi_4' = \pi_4 - \pi_{34}$. Thus $c_{126} = \pi_3' + \pi_4' + 2\pi_{34} = \pi_3 + \pi_4$, which is the same as the one in Table 1. Note 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 preferred.

IV An Example

In this example, we will illustrate the method presented in the previous sections, specifically setting up a linear programming problem and solving it. We assume the initial and current designs are two PSUs per stratum design[9]. S is composed of five PSUs, s_1, s_2, s_3, s_4, s_5, with new selection probabilities of .30, .20, .15, .30, .05. Then there are 10 combinations of pairs of PSUs, k = 1,2, , , ,10, one of which will be selected. We assume $T_1 = \{s_1,s_2,s_3,s_6,s_7\}$ and $T_2 = \{s_4,s_5,s_8,s_9\}$ and none of the PSUs was an SR PSU in 1990. Then $S \cap T_1 = \{s_1,s_2,s_3\}$ and

[9] The stratum we get from the initial design could be self-representative PSU. In this case, it will be one PSU/stratum case.

$S \cap T_2 = \{s_4, s_5\}$. We assume the initial selection probabilities of PSUs in T_1 were .40, .30, .15, .10 .05 and those in T_2 are .40, .30, .20, 10.

In S, we let k = 1 correspond to $\{s_1, s_2\}$ k = 2 to $\{s_1, s_3\}$, k = 3 to $\{s_1, s_4\}$, k=4 to $\{s_1, s_5\}$, k = 5 to $\{s_2, s_3\}$, k = 6 to $\{s_2, s_4\}$, k=7 to $\{s_2, s_5\}$, k=8 to $\{s_3, s_4\}$, k=9 to $\{s_3, s_5\}$ and k=10 to $\{s_4, s_5\}$. Corresponding probabilities of selection are .16113, .11394, .29, .03491, .05985, .16113, .0179, .11394, .01228 and .03491.

In T_1, $Pr(s_1, s_2) = .43427$, $Pr(s_1, s_3) = .18612$, $Pr(s_2, s_3) = .0853$, $Pr(s_1) = .17961$, $Pr(s_2) = .08043$, $Pr(s_3) = .02858$ and $Pr(\varnothing) = .0057$.

In T_2, $Pr(s_4, s_5) = .4277$, $Pr(s_4) = .3723$, $Pr(s_5) = .1723$ and $Pr(\varnothing) = .0277$.

The unknowns, $x_{i.j.k}$'s, to be solved and the associated constraints for the linear programming problem are shown in Table 3. The associated costs are in Table 4. Note that each subscript in x is delimited by ".", as the third subscript k of $x_{i\,j\,k}$ can reach 10, two digit number. Linear programming software CPLEX allows this convention. The cost table is constructed assuming the PSU configurations did not change over the 10 year period.

Using CPLEX we obtained the primal solutions in Table 5.

Table 3. x_{ijk}'s and Constraints Due to Column and Row Totals.

T_i	I_{ij}	k = 1 $\{s_1, s_2\}$	k=2 $\{s_1, s_3\}$	k=3 $\{s_1, s_4\}$	k=4 $\{s_1, s_5\}$	k=5 $\{s_2, s_3\}$	k=6 $\{s_2, s_4\}$	k=7 $\{s_2, s_5\}$	k=8 $\{s_3, s_4\}$	k=9 $\{s_3, s_5\}$	k=10 $\{s_4, s_5\}$	SUM
T_1	$I_{11} = \{s_1, s_2\}$	$x_{1.1.1}$	$x_{1.1.2}$	$x_{1.1.3}$	$x_{1.1.4}$	$x_{1.1.5}$	$x_{1.1.6}$	$x_{1.1.7}$	$x_{1.1.8}$	$x_{1.1.9}$	$x_{1.1.10}$	$.43427y_1$
T_1	$I_{12} = \{s_1, s_3\}$	$x_{1.2.1}$	$x_{1.2.2}$	$x_{1.2.3}$	$x_{1.2.4}$	$x_{1.2.5}$	$x_{1.2.6}$	$x_{1.2.7}$	$x_{1.2.8}$	$x_{1.2.9}$	$x_{1.2.10}$	$.18612y_1$
T_1	$I_{13} = \{s_2, s_3\}$	$x_{1.3.1}$	$x_{1.3.2}$	$x_{1.3.3}$	$x_{1.3.4}$	$x_{1.3.5}$	$x_{1.3.6}$	$x_{1.3.7}$	$x_{1.3.8}$	$x_{1.3.9}$	$x_{1.3.10}$	$.0853\, y_1$
T_1	$I_{14} = \{s_1\}$	$x_{1.4.1}$	$x_{1.4.2}$	$x_{1.4.3}$	$x_{1.4.4}$	$x_{1.4.5}$	$x_{1.4.6}$	$x_{1.4.7}$	$x_{1.4.8}$	$x_{1.4.9}$	$x_{1.4.10}$	$.17961y_1$
T_1	$I_{15} = \{s_2\}$	$x_{1.5.1}$	$x_{1.5.2}$	$x_{1.5.3}$	$x_{1.5.4}$	$x_{1.5.5}$	$x_{1.5.6}$	$x_{1.5.7}$	$x_{1.5.8}$	$x_{1.5.9}$	$x_{1.5.10}$	$.08043\, y_1$
T_1	$I_{16} = \{s_3\}$	$x_{1.6.1}$	$x_{1.6.2}$	$x_{1.6.3}$	$x_{1.6.4}$	$x_{1.6.5}$	$x_{1.6.6}$	$x_{1.6.7}$	$x_{1.6.8}$	$x_{1.6.9}$	$x_{1.6.10}$	$.02858y_1$
T_1	$I_{17} = \{\varnothing\}$	$x_{1.7.1}$	$x_{1.7.2}$	$x_{1.7.3}$	$x_{1.7.4}$	$x_{1.7.5}$	$x_{1.7.6}$	$x_{1.7.7}$	$x_{1.7.8}$	$x_{1.7.9}$	$x_{1.7.10}$	$.0057y_1$
T_2	$I_{21} = \{s_4, s_5\}$	$x_{2.1.1}$	$x_{2.1.2}$	$x_{2.1.3}$	$x_{2.1.4}$	$x_{2.1.5}$	$x_{2.1.6}$	$x_{2.1.7}$	$x_{2.1.8}$	$x_{2.1.9}$	$x_{2.1.10}$	$.4277\, y_2$
T_2	$I_{22} = \{s_4\}$	$x_{2.2.1}$	$x_{2.2.2}$	$x_{2.2.3}$	$x_{2.2.4}$	$x_{2.2.5}$	$x_{2.2.6}$	$x_{2.2.7}$	$x_{2.2.8}$	$x_{2.2.9}$	$x_{2.2.10}$	$.3723y_2$
T_2	$I_{23} = \{s_5\}$	$x_{2.3.1}$	$x_{2.3.2}$	$x_{2.3.3}$	$x_{2.3.4}$	$x_{2.3.5}$	$x_{2.3.6}$	$x_{2.3.7}$	$x_{2.3.8}$	$x_{2.3.9}$	$x_{2.3.10}$	$.1723y_2$
T_2	$I_{24} = \{\varnothing\}$	$x_{2.4.1}$	$x_{2.4.2}$	$x_{2.4.3}$	$x_{2.4.4}$	$x_{2.4.5}$	$x_{2.4.6}$	$x_{2.4.7}$	$x_{2.4.8}$	$x_{2.4.9}$	$x_{2.4.10}$	$.0277y_2$
SUM		.16113	.11394	.290	.03491	.05985	.16113	.0179	.11394	.01228	.03491	1.000

Table 4. Costs (Conditional Expected Number of Overlapped PSUs between 1990 and 2000) Table

T_i	I_{ij}	k = 1 $\{s_1, s_2\}$	k=2 $\{s_1, s_3\}$	k=3 $\{s_1, s_4\}$	k=4 $\{s_1, s_5\}$	k=5 $\{s_2, s_3\}$	k=6 $\{s_2, s_4\}$	k=7 $\{s_2, s_5\}$	k=8 $\{s_3, s_4\}$	k=9 $\{s_3, s_5\}$	k=10 $\{s_4, s_5\}$
T_1	$I_{11} = \{s_1, s_2\}$	2	1	1.8	1.6	1	1.8	1.6	0.8	0.6	1.4
T_1	$I_{12} = \{s_1, s_3\}$	1	2	1.8	1.6	1	0.8	0.6	1.8	1.6	1.4
T_1	$I_{13} = \{s_2, s_3\}$	1	1	0.8	0.6	2	1.8	1.6	1.8	1.6	1.4
T_1	$I_{14} = \{s_1\}$	1	1	1.8	1.6	0	0.8	0.6	0.8	0.6	1.4
T_1	$I_{15} = \{s_2\}$	1	0	0.8	0.6	1	1.8	1.6	0.8	0.6	1.4
T_1	$I_{16} = \{s_3\}$	0	1	0.8	0.6	1	0.8	0.6	1.8	1.6	1.4
T_1	$I_{17} = \{\varnothing\}$	0	0	0.8	0.6	0	0.8	0.6	0.8	0.6	1.4
T_2	$I_{21} = \{s_4, s_5\}$	1.4	1.1	1.8	1.8	0.9	1.6	1.6	1.3	1.3	2
T_2	$I_{22} = \{s_4\}$	1.4	1.1	1.8	0.8	0.9	1.6	0.6	1.3	0.3	1
T_2	$I_{23} = \{s_5\}$	1.4	1.1	0.8	1.8	0.9	0.6	1.6	0.3	1.3	1
T_2	$I_{24} = \{\varnothing\}$	1.4	1.1	0.8	0.8	0.9	0.6	0.6	0.3	0.3	0

Table 5. Primal Solutions of x_{ijk}'s

T_i	I_{ij}	k = 1 $\{s_1, s_2\}$	k=2 $\{s_1, s_3\}$	k=3 $\{s_1, s_4\}$	k=4 $\{s_1, s_5\}$	k=5 $\{s_2, s_3\}$	k=6 $\{s_2, s_4\}$	k=7 $\{s_2, s_5\}$	k=8 $\{s_3, s_4\}$	k=9 $\{s_3, s_5\}$	k=10 $\{s_4, s_5\}$	$p_{ij}\, y_i$
T_1	$I_{11} = \{s_1, s_2\}$	.16113	0	.03888	0	0	.1047	0	0	0	0	.30470*
T_1	$I_{12} = \{s_1, s_3\}$	0	0	.02443	0	0	0	0	.09389	.01228	0	.13059*
T_1	$I_{13} = \{s_2, s_3\}$	0	0	0	0	.05985	0	0	0	0	0	.05985
T_1	$I_{14} = \{s_1\}$	0	0	.12602	0	0	0	0	0	0	0	.12602
T_1	$I_{15} = \{s_2\}$	0	0	0	0	0	.05643	0	0	0	0	.05643
T_1	$I_{16} = \{s_3\}$	0	0	0	0	0	0	0	.02005	0	0	.02005
T_1	$I_{17} = \{\varnothing\}$	0	0	0	0	0	0	0	0	0	.00400	.00400
T_2	$I_{21} = \{s_4, s_5\}$	0	.04389	0	.03491	0	0	.0179	0	0	.03091	.12761
T_2	$I_{22} = \{s_4\}$	0	.01038	.10068	0	.00002	0	0	0	0	0	.11108
T_2	$I_{23} = \{s_5\}$	0	.05141	0	0	0	0	0	0	0	0	.05141
T_2	$I_{24} = \{\varnothing\}$	0	.00826	0	0	0	0	0	0	0	0	.00826
SUM		.16113	.11394	.290	.03491	.05985	.16113	.0179	.11394	.01228	.03491	1.0000

$y_1 = .701641$
$y_2 = .298359$

* Due to rounding error, the sum of the entries in rows 1 and 2, respectively, is different from the value given by $p_{ij}\, y_i$ by 1 in the fifth decimal place.

The dual procedure was also tried on the data. However, its solutions are exactly the same as the primal solutions.

If s_1 from T_1 and $\{s_4, s_5\}$ from T_2 were selected in the 1990 sample, the conditional probability of selection (CPOS) needed for the selection of the 2000 sample PSU pairs is

$$\text{CPOS} = \frac{x_{14_1k}}{.17961} + \frac{x_{21_2k}}{.4277}.$$

Note $\{s_1\}$ in T_1 corresponds to j = 4 and $\{s_4, s_5\}$ from T_2 to j = 1.

Table 6. Conditional Probability of Selection for Each Pair of PSUs

Candidate Pair of PSUs (k)	Conditional Probability	Cum Cond. Prob
1	0	0
2	.10262	.10262
3	.70164	.80426
4	.08162	.88588
5	0	.88588
6	0	.88588
7	.04185	.92773
8	0	.92773
9	0	.92773
10	.07227	1.00000
Sum	1.00000	1.00000

As sampling with probability proportional to estimated size (PPES) is used, a random number is needed to pick one of the k's in the table. It is generated using the "rannui" routine in SAS. The random seed used is 100,002 and the resulting random number was .62691. Thus, k=3, the pair of s_1 and s_4 was picked.

V 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, it was modified mimicking the 2000

geography. PSUs based on this data set were stratified and stratum data were transmitted to us 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}. Table 7 shows the number and percent of strata for which SUNSET Software successfully ran at different parameter settings. SUNSET Software initially set tolerance at 9. At that level, the software ran successfully for only 28 percent of a total of 107 strata. Tolerance level = 9 means that any number which lies between $-e^{-9}$ (-.000123409) and e^{-9} (.000123409) is considered zero. This is needed because rounding is involved in various calculations. This tolerance level turns out to be too stringent. When it was relaxed to 6, that is, any number that lies between $-e^{-6}$ (-.00247875) and e^{-6} (.00247875), is regarded as zero, the percentage of the strata that the software ran successfully more than doubled (61.68 percent). Relaxing further to 5 did not help. Devex pricing which is used for avoiding near-zero pivots (degeneracy problems) helped improve the success rate to 81.31 percent. As the original parameter settings did not allow us to read in large problems (we sometimes had problems having more than 277,100 variables in the objective function), when the settings were changed, it was able to handle all problems.

Table 7. Number of Strata Having Feasible Solution for Different Tolerance Level

	Tolerance le-9	Tolerance le-6	Tolerance le-5	Tolerance le-6 Devex Pricing
# of Strata Ran	30	66	64	87
Percent	28.04	61.68	59.81	81.31

In the beginning, another LP software CPLEX maintained by Statistical Research Division (SRD) was to be used to verify SUNSET Software's solutions of x_{ijk}'s. However, since DSMD was not able to run SUNSET Software on many strata, CPLEX was tried on the same strata. Note CPLEX has tolerance of 6 as a default. It could run on any stratum except two 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 programs. They were identical for stratum 421005 (Kim, 2001). However, for many strata, they were different (for this, see Wright and Tsao) . For example, solutions provided by SUNSET Software and CPLEX for stratum 531002, which are quite different, are shown below.

Table 8. Comparison of SUNSET Software and CPLEX Solutions

x_{ijk}-SUNSET Soft	x_{ijk} - CPLEX	Solution	Comments
$x_{2,1,14}$	$x_{1,1,14} + x_{1,2,14} + x_{1,3,14} + x_{1,7,14}$	.02505	
$x_{2,1,21} + x_{3,2,21}$	$x_{1,11,21} + x_{3,2,21}$	.22301	Different solutions for $x_{3,2,21}$.
$x_{2,1,41} + x_{3,4,41}$	$x_{1,4,41} + x_{1,9,41}$	.06664	
$x_{2,1,42}$	$x_{1,5,42} + x_{1,10,42}$	.01801	
$x_{2,2,2}$	$x_{1,11,2}$	.00506	
$x_{2,2,4}$	$x_{3,4,4}$	.00028	
$x_{2,2,7}$	$x_{1,5,7}$	.00154	
$x_{2,2,8}$	$x_{1,4,8}$	.00138	
$x_{2,2,11}$	$x_{1,7,11}$	.00077	
$x_{2,2,12}$	$x_{1,7,12}$	.00227	
$x_{2,2,17}$	$x_{1,3,17}$	.00291	
$x_{2,2,23}$	$x_{1,9,23} + x_{3,2,23}$	.00951	
$x_{2,2,25}$	$x_{3,4,25}$	.00043	
$x_{2,2,28}$	$x_{1,8,28}$	.00236	
$x_{2,2,33}$	$x_{1,8,33}$	.00695	
$x_{2,2,34}$	$x_{1,4,34}$	.00620	
$x_{2,2,43}$	$x_{1,4,43}$	.03335	
$x_{2,2,44}$	$x_{1,5,44}$	.00887	
$x_{3,1,5}$	$x_{1,11,5}$	.00072	
$x_{3,1,20}$	$x_{1,11,20} + x_{3,1,20}$	.05650	Different solutions for $x_{3,1,20}$.
$x_{3,1,31}$	$x_{3,4,31}$	.00323	
$x_{3,1,38}$	$x_{1,6,38} + x_{1,9,38}$	.00722	
$x_{3,1,39}$	$x_{1,10,39}$	.00413	
$x_{3,2,10}$	$x_{1,7,10}$	.04010	

$x_{3,2,18}$	$x_{1,11,18}$	.00776	
$x_{3,2,19}$	$x_{1,11,19}$	.02267	
$x_{3,2,22}$	$x_{1,8,22}$	.11712	
$x_{3,2,24}$	$x_{1,6,24} + x_{1,10,24}$	.02890	
$x_{3,3,13}$	$x_{1,7,13}$	.00577	
$x_{3,3,26}$	$x_{3,4,26}$	.00110	
$x_{3,3,36}$	$x_{3,3,36} + x_{3,4,36} + x_{1,11,36}$	.03540	Different solutions for $x_{3,3,36}$
$x_{3,3,37}$	$x_{1,5,37} + x_{1,8,37}$	.01754	
$x_{3,4,1}$	$x_{1,7,1}$	.00050	
$x_{3,4,3}$	$x_{3,4,3}$	.00010	Same
$x_{3,4,6}$	$x_{3,4,6}$	.00314	Same
$x_{3,4,9}$	$x_{1,3,9} + x_{1,6,9}$	.00360	
$x_{3,4,15}$	$x_{1,1,15}$	.01237	
$x_{3,4,16}$	$x_{1,2,16}$	.01105	
$x_{3,4,27}$	$x_{3,4,27}$	.00482	Same
$x_{3,4,29}$	$x_{1,9,29}$	.00211	
$x_{3,4,30}$	$x_{1,5,30}$	.00055	
$x_{3,4,32}$	$x_{3,4,32}$	.01412	Same
$x_{3,4,35}$	$x_{1,6,35}$	.00163	
$x_{3,4,40}$	$x_{1,8,40}$	.07428	
$x_{3,4,45}$	$x_{1,6,45}$	.00793	

SUNSET Software in the above table provided 49 nonzero solutions for x_{ijk}'s, compared to 61 for CPLEX. Note that we used 68 constraints on x's and one constraint on y's. Only five solutions are exactly the 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.

Using equation (10) we can compute the conditional probability of selection (CPOS) given the 1990 sampling situation for this stratum, which is,

$$CPOS = \sum_{i=1}^{r} \frac{x_{i j_i k}}{p_{i j_i}}$$

Since $p_{17_1} = .0667$ and $p_{32_3} = .62261$ (the specific subscripts here are due to the sampling situation in 1990 and p's are obtained from two 1990 strata),

$$CPOS = \frac{x_{17_1 k}}{.0667} + \frac{x_{32_3 k}}{.62221}$$

Note we use the sampling with probability proportional to size (PPES). In this case CPOS is used as estimated size.

Table 9. Conditional Probability of Selection for Each Pair of 2000 PSUs
- Based on CPLEX Solutions

Candidate Pair of PSUs (k)	Conditional Probability	Cum Cond. Prob
1	.00754	.00754
2	0	.00754
3	0	.00754
4	0	.00754
5	0	.00754
6	0	.00754
7	0	.00754
8	0	.00754
9	0	.00754
10	.60117	.60871
11	.01157	.62028
13	.03408	.65436
14	.08646	.74082
15	.12612	.86694
16	0	.86694
17	0	.86694

18	0	.86694
19	0	.86694
20	0	.86694
21	.13310	1.00004
Sum	1.00004	1.00004

As the random number used was .055549, using the cumulative conditional probability based on the CPLEX solutions, the 10th pair was picked from the above. If we repeat the above using SUNSET Software solutions, the CPOS and cumulative CPOS for k = 10 are .06445 and thus the 10th pair is picked. Thus even if the solutions for most x's are different between the softwares, we end up picking up the same pair of PSUs, disregarding whether we use SUNSET Software or CPLEX. However, if a different random number is generated, a different pair could be picked.

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). CPLEX was run using both options for stratum 421005 (Kim, 2001), whose results are as follows.

Table 10. Comparison of Primal and Dual Solutions - Stratum 421005

x_{ijk} - Primal	x_{ijk} - Dual	Solution	Comments
x_{211}	$x_{211} + x_{121}$	.10406	Different solution for x_{211}.
x_{112}	x_{112}	.03726	Same
x_{212}	x_{212}	.07296	Same
x_{213}	x_{123}	.08077	Different x's have the same solution.
$x_{124} + x_{214}$	x_{214}	.11644	Different solution for x_{214}.
x_{215}	x_{215}	.17613	Same
x_{226}	x_{226}	.03438	Same
x_{227}	x_{227}	.02489	Same
x_{228}	x_{228}	.03642	Same
x_{229}	x_{229}	.05653	Same
$x_{2,2,10}$	$x_{2,2,10}$	.02644	Same

$x_{2,2,11}$	$x_{2,2,11}$	.03867	Same
$x_{2,2,12}$	$x_{2,2,12}$	.05997	Same
$x_{2,2,13}$	$x_{2,2,13}$	.02801	Same
$x_{2,2,14}$	$x_{2,2,14}$	.04361	Same
$x_{2,2,15}$	$x_{2,2,15}$	.06347	Same

Most of the Primal and Dual solutions are the same, except for three cases. The Primal solution for x_{211} is shared by x_{211} and x_{121} in Dual. The Dual solution for x_{214} is divided into solutions for x_{214} and x_{124}. Thus individually the Primal and Dual solutions are not the same for x_{211} and x_{214}. The Primal solution for x_{213} is the same as the Dual solution for x_{123}.

For this stratum,

$$\text{CPOS} = \frac{x_{11_1k}}{.29053} + \frac{x_{22_2k}}{.47305} + \frac{x_{37_3k}}{.13253}$$

Since there are no differences between primal and dual solutions for x_{11_1k}, x_{22_2k} and x_{37_3k}, CPOS' for all k's are the same and thus the same PSUs will be selected disregarding whether we use primal or dual solutions or which random number we use for picking the PSUs.

Remember that most of the times the solutions of SUNSET Software were different from those of CPLEX. This indicates multiple solutions exist for the same problem.

The parameters of the LP problems we are dealing with are either the 1990 selection probabilities of PSUs or expected numbers of overlapped PSUs between the1990 and 2000 redesigns. As projected 1995 MOSs were used for the 1990 design and projected 2005 MOSs will be used for the 2000 design, the parameters of the LP problems are subject to errors. Thus it is instructive to perform some sensitivity analyses to determine the effect of error 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 x1.2.1 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 remained almost the same. As we do not know how good the estimates are, we may have to do sensitivity analysis by changing many more values in the objective function and constraints.

V. Percentage of The Retained 1990 PSUs in 2000 Test Runs

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 was computed, which is 52.06. Comparing with Ernst's 56 percent for one PSU/stratum design, we find them slightly low. The reasons for this can be twofold. 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 2000 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 we implemented Ernst's 1986 approach for maximally overlapping PSUs 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, DSMD purchased 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 for changing the parameter settings for SUNSET Software. In the end we were able to run SUNSET Software on all strata. In some cases, SUNSET Software and CPLEX provided the identical solution values for x's, but in others different solutions. We investigated whether these different solutions led to picking different pairs of PSUs or not. We also investigated whether or not the 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 represents a probability which is subject to error because projected counts are used as measure of size. 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. However, depending on the selected random number 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 states which did not have any sample PSUs, but in the 2000 design, every state has at least two sample PSUs. Thus in those states, the 2000 sample PSUs are not the sample PSUs in the 1990 design.

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. Causey, B.D., Cox, L.H., and Ernst, L.R. (1985) Applications of Transportation Theory to Statistical Problems. Journal of American Statistical Association, Vol. 80, pp 903-909.

3. Cochran, W.G. (1977) Sampling Techniques, Third Edition, John Wiley and Sons.

4. Durbin, J. (1967) Design of Multi-Stage Surveys for Estimation of Sampling Errors. Applied Statistics, Vol. 16, pp 152-164.

5. 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.

6. 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.

7. Hillier, F.S and Lieberman, G.J. (1972) Introduction to Operations Research. Holden-Day, Inc.

8. ILOG (1999) ILOG CPLEX 6.5 User's Manual, ILOG.

9. Keyfitz, N. (1954) Sampling with Probabilities, Journal of American Statistical Association, Vol. 46, pp 105-109.

10. Kim, J.J. (2000) Equivalency of Two Cost Formulae. Internal Census Bureau Memorandum.

11. Kim, J.J. (2001) Some Observations about Solutions of LP and Selection of PSUs - Stratum 421005. Internal Census Bureau Memorandum.

12. Kim, J.J. (2001) Some Observations About Solutions of LP and Selection of PSUs - NCVS Strata 11004 and 81007. Internal Census Bureau Memorandum.

13. Kim, J.J. (2001) Some Observations Concerning Primal and Dual Solutions of LP and Selection of PSUs - Stratum 531002. Internal Census Bureau Memorandum.

14. Raj, D. (1968) Sampling Theory, McGraw-Hill, New York.

15. Ross, S. (1994) A First Course in Probability. Fourth Edition. Macmillan College Publishing Company.

16. Wright, T., and Tsao, H. (1985) On an Optimal Solution for Maximizing the Probability of Retention in PPS Sampling. Linear Algebra and Its Applications. Vol 67, pp 67-80.

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.