BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-89/05

Further Applications of Linear Programming to Sampling Problems

by

Lawrence R. Ernst
Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

Recommended by: Lawrence R. Ernst

Report completed: 08/03/89

Report issued: 08/03/89

1. INTRODUCTION

Work by Cox and Ernst (1982), Causey, Cox and Ernst (1985) and Ernst (1986) has demonstrated the utility of linear programming in obtaining solutions to some statistical problems, particularly in sample design and estimation. In this paper some further developments in this area are presented.

In Section 2 the controlled rounding problem in three dimensions is considered. Controlled rounding is concerned with replacing nonintegers by integers in an additive array while preserving additivity. Cox and Ernst (1982) demonstrated that a controlled rounding exists for every two-dimensional additive array. It is established here, by means of a counterexample, that the natural generalization of their result to three dimensions does not hold, but that a rounding does always exist under less restrictive conditions.

Causey, Cox and Ernst (1985) presented an optimal solution under very general conditions to the problem of maximizing overlap between primary sampling units (PSUs) when redesigning sample surveys. Their solution modeled the problem as a transportation problem. In Section 3 two modifications of that procedure are presented. One modification very substantially reduces the size of the transportation problems used in the original procedure, which sometimes can be unmanageably large. The second modification results in an overlap procedure which preserves the independence of the selection of sample PSUs from stratum to stratum, an independence which is generally destroyed by overlap procedures if the initial and new designs do not have the same stratification.

In Section 4 linear programming is considered as an alternative to stratification as a method of reducing between PSU variances. The linear programming approach is conceptually very simple and flexible, and permits the optimal balancing of such often conflicting goals as the minimization of variances and the ability to estimate variances. Linear programming is also

applicable to the selection of PSUs for two or more dependent designs simultaneously, such as when the sample PSUs for one design are required to be a subset of the sample PSUs from a second design. As a result, this approach has possible applicability to the proposed expansion of the Current Population Survey (CPS) as will be explained. However, as noted in Section 4, the procedure also has the potentially fatal flaw for some design problems that the corresponding linear programming problem may be too large to solve practically.

2. THREE-DIMENSIONAL CONTROLLED ROUNDINGS

Cox and Ernst (1982) introduced the concept of controlled rounding in two dimensions and proved that there exists a controlled rounding for every two-dimensional additive array. The question of whether that result generalized to three dimensions had remained unanswered until now. In Section 2.2 a negative answer to this question is presented by means of a counterexample. Then in Section 2.3 it is proven that a rounding satisfying a less restrictive condition exists for each three-dimensional array. First, however, the notation and concepts of controlled rounding, and the results in Cox and Ernst (1982) are briefly summarized in Section 2.1.

2.1 Preliminaries

A $(m+1)\times(n+1)\times(\ell+1)$ array $A=(a_{ijk})$ is said to be a tabular array if

$$\sum_{i=1}^{m} a_{ijk} = a_{(m+1)jk}, \quad 1\le j\le n+1, \quad 1\le k\le \ell+1, \tag{2.1}$$

$$\sum_{j=1}^{n} a_{ijk} = a_{i(n+1)k}, \quad 1\le i\le m+1, \quad 1\le k\le \ell+1, \tag{2.2}$$

$$\sum_{k=1}^{\ell} a_{ijk} = a_{ij(\ell+1)}, \quad 1\le i\le m+1, \quad 1\le j\le n+1. \tag{2.3}$$

This is analogous to the definition of a tabular array in two dimensions for which the third subscript is omitted from (2.1) and (2.2), and there is no (2.3). A two-dimensional, (m+1)x(n+1) tabular array can be represented in the form

$$\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & a_{1(n+1)} \\ \vdots & \cdots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & a_{m(n+1)} \\ \hline a_{(m+1)1} & \cdots & a_{(m+1)n} & a_{(m+1)(n+1)} \end{array}$$

and a three-dimensional, (m+1)x(n+1)x(ℓ+1) tabular array can be represented by the following sequence of ℓ+1 two-dimensional tabular arrays, that will be referred to as levels

$$\begin{array}{ccc|c} a_{111} & \cdots & a_{1n1} & a_{1(n+1)1} \\ \vdots & \cdots & \vdots & \vdots \\ a_{m11} & \cdots & a_{mn1} & a_{m(n+1)1} \\ \hline a_{(m+1)11} & \cdots & a_{(m+1)n1} & a_{(m+1)(n+1)1} \end{array} \quad \cdots \quad \begin{array}{ccc|c} a_{11(\ell+1)} & \cdots & a_{1n(\ell+1)} & a_{1(n+1)(\ell+1)} \\ \vdots & \cdots & \vdots & \vdots \\ a_{m1(\ell+1)} & \cdots & a_{mn(\ell+1)} & a_{m(n+1)(\ell+1)} \\ \hline a_{(m+1)1(\ell+1)} & \cdots & a_{(m+1)n(\ell+1)} & a_{(m+1)(n+1)(\ell+1)} \end{array}$$

together with the additional condition that each cell entry in level (ℓ+1), that is the totals level, be the sum of the entries in the corresponding cells of levels 1 through ℓ. Observe also that in the three-dimensional case, the right hand side of (2.1) is a one-dimensional marginal if $j<n+1$ and $k<\ell+1$, a two-dimensional marginal if $j=n+1$ or $k=\ell+1$ but not both, and the grand total if $j=n+1$ and $k=\ell+1$. Similar statements hold for (2.2) and (2.3).

The conventional rounding of a tabular array $A=(a_{ijk})$ with respect to a positive integer base b is an array (r_{ijk}) for which $r_{ijk} = [a_{ijk}/b+.5]b$ for all i,j,k, where [ ] denotes the greatest integer function. Such a rounding is often not a tabular

array. The search for a less restrictive form of rounding which it was hoped might always yield a rounding of a tabular array that was itself tabular motivated the definition of controlled rounding. In the three-dimensional case a controlled rounding of a $(m+1)\times(n+1)\times(\ell+1)$ tabular array $A=(a_{ijk})$ with respect to a positive integer base b is a $(m+1)\times(n+1)\times(\ell+1)$ array $R(A)=(r_{ijk})$ for which

$$R(A) \text{ is a tabular array,} \tag{2.4}$$

$$r_{ijk} = [a_{ijk}/b]b \text{ or } r_{ijk}=[a_{ijk}/b]b + b \text{ for all } i,j,k. \tag{2.5}$$

The analogous definition in two dimensions is obvious. The definition of a slightly more restrictive form of rounding known as zero-restricted controlled rounding is obtained by replacing (2.5) by the condition,

$$r_{ijk} \text{ is an integral multiple of } b \text{ and } |r_{ijk}-[a_{ijk}/b]b|<b \text{ for all } i,j,k. \tag{2.6}$$

Note that for each i,j,k the set of r_{ijk} satisfying (2.5) and (2.6) only differ when a_{ijk} is a multiple of b, in which case (2.6) is satisfied only if $r_{ijk} = a_{ijk}$.

In Cox and Ernst (1982) it was established that a controlled rounding and even a zero-restricted controlled rounding exists for every two-dimensional tabular array. In Causey, Cox and Ernst (1985), an example was presented of a three-dimensional tabular array which has no zero-restricted controlled roundings, but that array does have controlled roundings. In the next subsection it is shown that controlled roundings do not always exist in three dimensions, but then in Section 2.3 it is shown that there exists for every three-dimensional tabular array $A=(a_{ijk})$, a tabular array $R(A)=(r_{ijk})$ for which

$$r_{ijk} \text{ is an integral multiple of } b \text{ and } |r_{ijk} - [a_{ijk}/b]b|<2b \text{ for all } i,j,k. \tag{2.7}$$

2.2 A Three-Dimensional Tabular Array with No Controlled Roundings

For any $(m+1)x(n+1)x(\ell+1)$ tabular array $A=(a_{ijk})$, let $I(A)$ denote the $mxnx\ell$ matrix consisting of the internal elements of A, that is $I(A)=(a_{ijk})$, $1<i<m$, $1<j<n$, $1<k<\ell$.

The construction of a tabular array B for which no controlled rounding exists consists of two steps. First let $B=(b_{ijk})$ be the 5x5x5 tabular array, with the following representation as a set of five levels:

Level 1

.5	0	.5	0	1
0	.5	.5	0	1
.5	.5	0	0	1
0	0	0	0	0
1	1	1	0	3

Level 2

.5	0	0	.5	1
0	.5	0	.5	1
0	0	0	0	0
.5	.5	0	0	1
1	1	0	1	3

Level 3

0	0	0	0	0
0	0	.5	.5	1
0	.5	0	.5	1
0	.5	.5	0	1
0	1	1	1	3

Level 4

0	0	.5	.5	1
0	0	0	0	0
.5	0	0	.5	1
.5	0	.5	0	1
1	0	1	1	3

Level 5

1	0	1	1	3
0	1	1	1	3
1	1	0	1	3
1	1	1	0	3
3	3	3	3	12

Figure 1. The Tabular Array B

Then let $B' = (b'_{ijk})$ be the 13x13x5 tabular array with the set of internal elements $I(B')$ defined by

$$
\begin{aligned}
b'_{ijk} &= b_{ijk} && \text{if } 1<i<4, \quad 1<j<4, \\
&= b_{(i-4)(j-4)k} && \text{if } 5<i<8, \quad 5<j<8, \\
&= b_{(i-8)(j-8)k} && \text{if } 9<i<12, \quad 9<j<12, \\
&= 0 && \text{for all other } i,j,k.
\end{aligned}
$$

$I(B')$ can be represented as a set of 9 blocks, each 4x4x4, as shown in Figure 2.

I(B)	0	0
0	I(B)	0
0	0	I(B)

Figure 2. Block Representation of $I(B')$

Since the grand total for B, b_{555}, is 12, the grand total for a controlled rounding of B could be 12 or 13. However, a key step in the proof that there are no controlled roundings for B' is to establish the following result:

If $R(B) = (r_{ijk})$ is a controlled rounding of B
then $r_{555} = 13$. (2.8)

It will first be shown that there exists no controlled roundings of B' provided (2.8) holds, and then that (2.8) does indeed hold.

The proof that there exists no controlled roundings of B' given (2.8) is obtained by contradiction. Assume that $R(B')=(r'_{ijk})$ is a controlled rounding of B' and partition $I(R(B'))$ into 9 blocks $R_1,\ldots,R_9$, each 4x4x4 corresponding to the 9 blocks in Figure 2 as follows.

R_1	R_2	R_3
R_4	R_5	R_6
R_7	R_8	R_9

Figure 3. Block Representation of $I(R(B'))$

Since $b'_{13,13,5} = 36$,

$$r'_{13,13,5} = 36 \text{ or } r'_{13,13,5} = 37. \tag{2.9}$$

Furthermore, since all marginals of B' are integers, it follows that

$$r'_{ijk} = b'_{ijk} \text{ or } r'_{ijk} = b'_{ijk} + 1 \text{ for all marginal cells } i,j,k, \tag{2.10}$$

and, in order for (2.9) to be satisfied,

$r'_{ijk} = b'_{ijk} + 1$ for at most one marginal of dimension one in each of the three directions. (2.11)

It follows from (2.10) and (2.11) that either all elements in the off-diagonal blocks in Figure 3 are 0, or that there is a total of a single 1 in all of these six blocks combined. In the former case, R_1, R_5 and R_9 would then each constitute a set of internal elements for a controlled rounding of B since the marginal constraints that have to be satisfied for a controlled rounding of B would be met by each of these blocks by (2.10) and (2.11). However, then by (2.8), the sum of the elements in each of the

blocks R_1, R_5, and R_9 would be 13, contradicting (2.9).

In the latter case, when one off-diagonal block, say R_2, has a 1 cell, then R_9 would be a set of internal elements for a controlled rounding of B and, by (2.8), the sum of the elements in R_9 would be 13. The sum of the elements in each of R_1 and R_5 must be at least 12 by (2.10) and the sum of the elements in the union of R_1, R_2, R_5 and R_9 would be at least 38, contradicting (2.9) and thus establishing that there are no controlled roundings of B provided (2.8) holds.

To establish (2.8), consider a controlled rounding $R=(r_{ijk})$ of B and observe that if $r_{555}=b_{555}=12$ then

$$r_{ijk} = b_{ijk} \text{ for each marginal cell } (i,j,k), \tag{2.12}$$

since all marginals of B are integers. (2.8) will be established by assuming (2.12) and then showing that (2.12) implies that

$$\sum_{k=1}^{4} r_{ijk} \neq 1 \text{ for either } (i,j)$$

$$= (1,1),\ (2,2),\ (2,3) \text{ or } (3,2), \tag{2.13}$$

contradicting (2.12) for one of the corresponding four marginals cells (1,2,5), (2,2,5), (2,3,5), (3,2,5). (2.13) in turn will be proven by establishing the following four statements concerning the cell values for (r_{ijk}) in levels 1 through 4 respectively:

$$r_{111}=r_{231}=r_{321}=1,$$
$$\text{or } r_{111}=r_{231}=r_{321}=0, \tag{2.14}$$

$$r_{112}=1,\ r_{222}=r_{232}=r_{322}=0$$
$$\text{or } r_{112}=0,\ r_{222}=1,\ r_{232}=r_{322}=0, \tag{2.15}$$

$$r_{223}=1,\ r_{233}=r_{323}=0,\ \text{or}\ r_{223}=0,\ r_{233}=1,\ r_{323}=0,$$
$$\text{or}\ r_{223}=0,\ r_{233}=0,\ r_{323}=1, \tag{2.16}$$

$$r_{234}=r_{324}=0. \tag{2.17}$$

(2.14 - 2.17) establishes (2.13) since only the first possibility in (2.14) in combination with the first possibility in (2.16) satisfies

$$\sum_{k=1}^{4} r_{23k} = \sum_{k=1}^{4} r_{32k} = 1.$$

However, that combination together with either of the possibilities in (2.15) yield

$$\sum_{k=1}^{3} r_{11k} = 2 \quad \text{or} \quad \sum_{k=1}^{3} r_{22k} = 2.$$

Thus it remains only to establish (2.14 - 2.17). In proving these statements use is made of the fact that

$$\text{if } (i,j,k) \text{ is an internal cell of } B,\ (i,j,k) \neq (2,2,3)$$
$$\text{or } (1,1,4), \text{ and } b_{ijk}=0, \text{ then } r_{ijk}=0, \tag{2.18}$$

since at least one of the three, one-dimensional marginal cells corresponding to each such cell in B is 0.

To prove (2.14), observe first that if $r_{111}=1$, then since $r_{151}=1$ by (2.12), it follows that $r_{131}=0$. Next $r_{131}=0$ combined with $r_{531}=1$ and (2.18) yield $r_{231}=1$, which in turn implies that $r_{221}=0$, which finally implies that $r_{321}=1$. Thus if $r_{111}=1$, the first possibility in (2.14) holds. Similarly $r_{111}=0$ yields the second possibility in (2.14).

To establish (2.15) first note that $r_{232}=r_{322}=0$ by (2.18). Furthermore, if $r_{112}=1$ then, proceeding as was done to establish (2.14), it follows that $r_{142}=0$, $r_{242}=1$ and finally $r_{222}=0$. Similarly, it is established that if $r_{112}=0$ then $r_{222}=1$.

To obtain (2.16), first note that if $r_{223}=1$ then $r_{233}=r_{323}=0$ by (2.12). If $r_{223}=0$ and $r_{233}=1$ then from (2.12) and (2.18) it is successively obtained that $r_{243}=0$, $r_{343}=1$ and finally $r_{323}=0$. Similarly if $r_{223}=0$ and $r_{233}=0$ then $r_{323}=1$.

(2.17) immediately follows from (2.18).

Remark 2.1: All the marginals of B' are integers, a fact that was used extensively in proving that there are no controlled roundings of B'. However, B' can be easily modified to obtain a 5x5x13 tabular array, $B''=(b''_{ijk})$ which has no controlled roundings and for which none of the cells, internal or marginal, are integers. Simply define $I(B'')$ by choosing any ε with $0<\varepsilon<1/576$ and letting $b''_{ijk}=b'_{ijk}+\varepsilon$ for each internal cell (i,j,k). Since there are 576 internal cells in B'', no cells of B'', including marginals, are integers and $[b''_{ijk}]=[b'_{ijk}]$ for all cells in B''. Therefore, the set of controlled roundings of B'' is identical to the set of controlled roundings of B', namely the empty set.

2.3 An Additive Rounding in Three Dimensions Which Always Exists

It will be shown that for every $(m+1)x(n+1)x(\ell+1)$ tabular array $A=(a_{ijk})$ there exists a $(m+1)x(n+1)x(\ell+1)$ array $R(A)=(r_{ijk})$ satisfying (2.4) and (2.7). Such an array is obtained by successively defining a sequence of two-dimensional, base b, zero-restricted controlled roundings. First let (r_{ij1}) be a zero-restricted controlled rounding of the $(m+1)x(n+1)$ array (a_{ij1}). Then for $k=2,\ldots,\ell$, let

$$c_{ijk} = \sum_{t=1}^{k} a_{ijt} - \sum_{t=1}^{k-1} r_{ijt}, \quad 1\le i\le m+1, \ 1\le j\le n+1, \qquad (2.19)$$

and take (r_{ijk}) to be a two-dimensional zero-restricted controlled rounding of the $(m+1)x(n+1)$ array (c_{ijk}) with k fixed. Finally, let

$$r_{ij(\ell+1)} = \sum_{k=1}^{\ell} r_{ijk}, \quad 1 \le i \le m+1, \; 1 \le j \le n+1. \tag{2.20}$$

Observe that the three-dimensional array (r_{ijk}) is clearly tabular. To establish (2.7), first note that it is obviously true for any cell for which k=1. For $k=2,\ldots,\ell$, it follows from (2.19) that

$$\begin{aligned} |r_{ijk} - a_{ijk}| &\le |r_{ijk} - c_{ijk}| + |c_{ijk} - a_{ijk}| \\ &= |r_{ijk} - c_{ijk}| + |c_{ij(k-1)} - r_{ij(k-1)}| < 2b, \end{aligned} \tag{2.21}$$

except that a_{ij1} replaces c_{ij1} in (2.21) when k=2.

Finally, by (2.19) and (2.20),

$$\begin{aligned} |r_{ij(\ell+1)} - a_{ij(\ell+1)}| &= \left| \sum_{k=1}^{\ell} r_{ijk} - \sum_{k=1}^{\ell} a_{ijk} \right| \\ &= |\, r_{ij\ell} - c_{ij\ell} \,| < b. \end{aligned}$$

3. FURTHER RESULTS ON MAXIMIZING THE OVERLAP BETWEEN SURVEYS

The problem of maximizing the expected number of PSUs retained in sample when redesigning a survey with a stratified design for which the PSUs are selected with probability proportional to size was introduced to the literature by Keyfitz (1951). Causey, Cox and Ernst (1985) were able to obtain an optimal solution to this problem under very general conditions by formulating it as a transportation problem. Unlike previous approaches, this procedure imposed no restrictions on changes in strata definitions or number of PSUs per stratum. The reader of this section is urged to read that paper to facilitate understanding of the work to be presented here.

There are several difficulties associated with the use of the procedure of Causey, Cox and Ernst. One of these can occur when the initial sample of PSUs was not selected independently from stratum to stratum, in which case the information necessary to compute all the joint probabilities needed to use this method may not be known in practice. An alternative linear programming procedure for use in that situation was developed by Ernst (1986) and was used by the Bureau of the Census in the last redesign of the demographic surveys that they conduct, in the selection of the sample PSUs for the CPS and the National Crime Survey. In this section approaches are presented for handling two other difficulties.

The first problem is that in the procedure of Causey, Cox and Ernst the transportation problem used in the selection of the sample PSUs for the new design in each stratum can be unmanageably large. To see this, note that each possibility for the set of PSUs in a new stratum S that were in the sample for the initial design corresponds to a row in the transportation problem, and each possibility for the set of PSUs in S in sample in the new design corresponds to a column. If S consists of n PSUs from which m are to be selected without replacement in the new design, then the number of columns is $\binom{n}{m}$, which is a reasonably-sized number for m=1 or 2 say, if n is moderately sized. However, for any m the number of rows can be as large as 2^n, resulting in a transportation problem too large to practically solve even for moderately-sized n. For example, if n=40, m=2, the $\binom{n}{m} = 780$, while 2^n exceeds one trillion. A situation where the upper bound of 2^n is attained occurs if each of the PSUs in S were in a different initial nonself-representing stratum and the sampling in the initial design was one PSU per stratum selected independently from stratum to stratum. In that case any of the 2^n subsets of S can be the set of PSUs in S that were initial sample PSUs.

In Section 3.1 a modified procedure is presented for which the number of initial outcomes used in the transportation problem

is vastly reduced, resulting in a transportation problem that should be manageable for typical values of n and m. The expected number of PSUs retained when applying this modified procedure is, not surprisingly, generally less than for the original procedure, but it is believed that in practice the loss in overlap usually would be small.

The second problem considered in this section, unlike the first, applies not only to the procedure of Causey, Cox and Ernst, but to all previous overlap procedures that this author is aware of, whenever the initial and new designs have different stratifications. Overlap procedures in this case destroy the independence of the selection of sample PSUs from stratum to stratum in the new design (Ernst 1986). Among the consequences of this loss of independence are changes in variances which are almost never accounted for in the variance estimates. In Section 3.2 another modification of the procedure of Causey, Cox and Ernst is presented which preserves the independence of the selection of sample PSUs from stratum to stratum in the new design. The procedure also generally reduces expected overlap in comparison with the original procedure, in some cases drastically.

The two procedures to be presented in this section can also be combined to accomplish the goals of both procedures, as described in Section 3.2.

3.1 A Reduced-Size Transportation Problem for Maximizing Overlap

The reduced-size procedure will, for ease of presentation, be described in detail only for the case when both the initial and new designs are two PSUs per stratum without replacement. Then the changes necessary to apply this procedure for other initial and new designs will be sketched. It is assumed throughout this subsection that PSUs in the initial sample were selected independently from stratum to stratum.

The general outline of the procedure for the particular case to be detailed is as follows. Let $A_1,\ldots,A_n$ denote the set of

PSUs in a new stratum S. Let the random set I denote the set of integers i for which A_i was in the initial sample and let N be the corresponding random set with respect to the new sample. The set of all distinct pairs of integers $i,j \in \{1,\ldots,n\}$ will be ordered in a manner that the pairs i,j listed earlier correspond to pairs of PSUs A_i, A_j that have a better chance of being retained in sample in the new design if they were in sample in the initial design. If I consists of at least two integers then the new selection probabilities are conditioned only on the first listed pair in the ordering contained in I. If I consists of exactly one integer or is empty then the new selection probabilities are conditioned on the actual initial outcome, that is I itself. Thus the new selection probabilities would be conditioned on exactly $\binom{n}{2}+n+1$ events instead of a possible 2^n events.

To obtain the desired ordering of the pairs of integers, an ordering $f(1),\ldots,f(n)$ of $\{1,\ldots,n\}$ will first be obtained. Then corresponding to each $k=1,\ldots,n-1$, an ordering $g_k(1),\ldots,g_k(n-k)$ of $\{1,\ldots,n\}\sim\{f(1),\ldots,f(k)\}$ will be constructed. A linear ordering of the distinct pairs in $\{1,\ldots,n\}$ would then be determined as follows. Each such pair can be represented uniquely as an ordered pair $(f(k), g_k(\ell))$ for some $k \in \{1,\ldots,n-1\}$ $\ell \in \{1,\ldots,n\}\sim\{f(1),\ldots,f(k)\}$. A second pair representable in the form $(f(k'), g_{k'}(\ell'))$ precedes $(f(k), g_k(\ell))$ if and only if either $k'<k$, or $k'=k$ and $\ell'<\ell$.

To obtain the ordering $f(1),\ldots,f(n)$, first let p_i, π_i denote the probability that $i \in I$ and $i \in N$ respectively, and p_{ij}, π_{ij}, $i\neq j$, be the joint probability that $i,j \in I$ and $i,j \in N$ respectively. Then successively define $f(k)$, $k=1,\ldots,n$, by choosing $f(k) \in T_k$ satisfying

$$\pi_{f(k)}/p_{f(k)} = \max\{\pi_i/p_i^{(k)} : i \in T_k\},$$

where

$$T_1 = \{1,\ldots,n\},\ T_k = T_{k-1} \sim \{f(k-1)\},\ k=2,\ldots,n,$$

$$p_i^{(k)} = P(i \in I \text{ and } I \subset T_k),\ k=1,\ldots,n,\ i \in T_k.$$

Since $p_i^{(1)} = p_i$, the ordering just defined corresponds to placing first a PSU A_i with the highest ratio for π_i/p_i. This is appropriate since a high value for this ratio makes it more likely that a PSU can be retained in the new sample when it was in the initial sample. For $k>2$, the denominator of this ratio, $p_i^{(k)}$, is the probability that if $f(k)=i$ then an attempt is made to retain A_i in the new sample either as the first member of an ordered pair of initial sample PSUs or as the only initial sample PSU in S.

It remains to explain how to compute $p_i^{(k)}$ for $k>2$. To this end, let r denote the number of initial strata with PSUs in common with S and let F_α, $\alpha=1,\ldots,r$ denote a partition of $\{1,\ldots,n\}$ such that i and j are in the same F_α if and only if A_i and A_j were in the same initial stratum. Then let

$$p'_\alpha(T) = P(I \cap F_\alpha \subset T) \text{ for } T \subset \{1,\ldots,n\},\ \alpha=1,\ldots,r,$$

$$p''_{i\alpha}(T) = P(i \in I \text{ and } I \cap F_\alpha \subset T) \text{ for } T \subset \{1,\ldots,n\},\ \alpha=1,\ldots,r,\ i \in F_\alpha \cap T,$$

and observe that

$$p'_\alpha(T) = 1 - \sum_{i \in F_\alpha \sim T} p_i + \sum_{\substack{i,j \in F_\alpha \sim T \\ i<j}} p_{ij},$$

$$p''_{i\alpha}(T) = p_i - \sum_{j \in F_\alpha \sim T} p_{ij},$$

and finally that

$$p_i^{(k)} = p'_{i\alpha}(T_k) \prod_{\substack{\ell=1 \\ \ell\neq\alpha}}^{r} p'_\ell(T_k), \quad k=1,\ldots,n, \; i \in F_\alpha \cap T_k.$$

Next, for each $k=1,\ldots,n-1$ the ordering $g_k(\ell)$, $\ell=1,\ldots,n-k$, is recursively defined by choosing $g_k(\ell) \in T_{k\ell}$ satisfying

$$\pi_{f(k),g_k(\ell)}/p^{(\ell)}_{f(k),g_k(\ell)} = \max\{\pi_{f(k),j}/p^{(\ell)}_{f(k),j} : j \in T_{k\ell}\},$$

where

$$T_{k1} = \{1,\ldots,n\}\sim\{f(1),\ldots,f(k)\},$$

$$T_{k\ell} = T_{k(\ell-1)}\sim\{g_k(\ell-1)\}, \; \ell=2,\ldots,n-k,$$

$$T^*_{k\ell} = T_{k\ell}\cup\{f(k)\}, \quad \ell=1,\ldots,n-k,$$

$$p^{(\ell)}_{f(k),j} = P(f(k), \; j \in I \text{ and } I \subset T^*_{k\ell}), \; \ell=1,\ldots,n-k, \; j \in T_{k\ell}.$$

The rationale for this ordering for the second PSU in the pair is analogous to the rationale for the ordering of the first PSU.

To compute $p^{(\ell)}_{f(k),j}$, observe that if $f(k) \in F_\alpha$, $j \in F_\beta$, then

$$p^{(\ell)}_{f(k),j} = p_{f(k),j} \prod_{\substack{t=1 \\ t\neq\alpha}}^{r} p'_t(T^*_{k\ell}) \quad \text{if } \alpha=\beta,$$

$$= p'_{f(k),\alpha}(T^*_{k\ell}) p'_{j\beta}(T^*_{k\ell}) \prod_{\substack{t=1 \\ t\neq\alpha,\beta}}^{r} p'_t(T^*_{k\ell}) \quad \text{if } \alpha\neq\beta.$$

Having defined the ordering of the distinct pairs of integers in $\{1,\ldots,n\}$, it will next be stated how for each

possibility for I a unique I_i is associated among the subsets I_i, $i=1,\ldots,\binom{n}{2}+n+1$, of $\{1,\ldots,n\}$ of two or fewer elements, together with formulas for computing p_i^*, the probability that I_i is the associated set. For each I, it is the associated I_i on which the new selection probabilities are conditioned. If I consists of two or more integers then $I_i = \{f(k), g_k(\ell)\}$ where $(f(k), g_k(\ell))$ is the first pair in the ordering just defined for which $\{f(k), g_k(\ell)\} \subset I$, and $p_i^* = p^{(\ell)}_{f(k),g_k(\ell)}$. If $I=\{t\}$, for some $t \in F_\alpha$, then $I_i=\{t\}$ and

$$p_i^* = p''_{t\alpha}(\{t\}) \prod_{\substack{u=1 \\ u\neq\alpha}}^{r} p'_u(\{t\}) .$$

Finally if $I=\emptyset$, then $I_i = \emptyset$ and

$$p_i^* = \prod_{u=1}^{r} p'_u(\emptyset) .$$

As for the new sample, there are $\binom{n}{2}$ possibilities denoted S_j, $j=1,\ldots,\binom{n}{2}$, for N. If $S_j = \{s,t\}$, then π_j^*, the probability that $N=S_j$, is simply π_{st}.

The transportation problem to solve for this procedure can at last be stated. For $i=1,\ldots,\binom{n}{2}+n+1$, $j=1,\ldots,\binom{n}{2}$, x_{ij} is the joint probability that I_i is the set associated to I and $N=S_j$, while c_{ij} is the expected number of PSUs in $I \cap S_j$ given I_i. The x_{ij}'s are the variables and the transportation problem to solve is to determine $x_{ij}>0$ that maximize

$$\sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij}x_{ij},$$

subject to

$$\sum_{j=1}^{\binom{n}{2}} x_{ij} = p_i^*, \qquad i=1,\ldots,\binom{n}{2}+n+1,$$

$$\sum_{i=1}^{\binom{n}{2}+n+1} x_{ij} = \pi_j^*, \qquad j=1,\ldots,\binom{n}{2}.$$

Once the optimal x_{ij}'s have been obtained, the conditional new selection probabilities for S_j, $j=1,\ldots,\binom{n}{2}$, given I_i, are x_{ij}/p_i^*.

It remains only to explain how to compute c_{ij}. Let

$$b_{it} = P(t \in I | I_i), \quad i=1,\ldots,\binom{n}{2}+n+1, \quad t=1,\ldots,n,$$

and note that if $S_j = \{s,t\}$ then $c_{ij} = b_{is} + b_{it}$.

To compute b_{it}, observe that

$$\begin{aligned} b_{it} &= 0 \quad \text{if} \quad I_i = \emptyset, \\ &= 1 \quad \text{if} \quad I_i = \{v\} \text{ and } t=v, \\ &= 0 \quad \text{if} \quad I_i = \{v\} \text{ and } t \neq v, \end{aligned}$$

while if $I_i = \{f(k), g_k(\ell)\}$ and $f(k) \in F_\alpha$, $g_k(\ell) \in F_\beta$, $t \in F_\gamma$, then

$$\begin{aligned} b_{it} &= 1 && \text{if } t=f(k) \text{ or } t=g_k(\ell), \\ &= 0 && \text{if } t \notin T_{k\ell}^*, \\ &= 0 && \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \text{ and } \alpha=\beta=\gamma, \\ &= \frac{p_{f(k),t}}{p_{f(k),\alpha}(T_{k\ell}^*)} && \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \text{ and } \alpha=\gamma \neq \beta, \end{aligned}$$

$$= \frac{p_{g_k(\ell),t}}{p''_{g_k(\ell),\beta}(T^*_{k\ell})} \qquad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \text{ and } \beta=\gamma\neq\alpha,$$

$$= \frac{p''_{t\gamma}(T^*_{k\ell})}{p'_{\gamma}(T^*_{k\ell})} \qquad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \text{ and } \gamma\neq\alpha,\ \gamma\neq\beta.$$

Modifications of the procedure just described when either the initial or new designs are not two PSUs per stratum will now be sketched.

A different number of PSUs per stratum in the initial design only requires modification of some of the computations. For example if m=2, but the initial design was not two PSUs per stratum, then the computations for $p_i^{(k)}$, $p_{f(k),j}^{(\ell)}$ and c_{ij} would be different but their definitions would not change.

If m=3, then the set of all distinct triples, instead of pairs, of integers in {1,...,n}, is ordered. If I consists of at least three integers then the new selection probabilities are conditioned only on the first listed triple in the ordering. Otherwise, the new selection probabilities are conditioned on I itself. Thus the new selection probabilities would be conditioned on $\binom{n}{3}+\binom{n}{2}+n+1$ events.

To obtain the desired ordering of the triples of integers, first the orderings $f(1),\ldots,f(n)$ and $g_k(1),\ldots,g_k(n-k)$ are constructed exactly as in the case m=2. Then corresponding to each $k=1,\ldots,n-2$, $\ell=1,\ldots,n-k-1$, an ordering $h_{k\ell}(1),\ldots,h_{k\ell}(n-k-\ell)$ of $\{1,\ldots,n\}\sim\{f(1),\ldots,f(k), g_k(1),\ldots,g_k(\ell)\}$ is constructed similarly to the construction of $g_k(1),\ldots,g_k(n-k)$. For example, in defining $h_{k\ell}(v)$ for $v>2$, $p_{f(k),j}^{(\ell)}$ in the definition of $g_k(\ell)$ is replaced by

$$P(f(k),\ g_k(\ell),\ j \in I \text{ and } I \subset T^*_{k\ell}\sim\{h_{k\ell}(1),\ldots,h_{k\ell}(v-1)\}).$$

A linear ordering of the distinct triples in $\{1,\ldots,n\}$ is then determined by representing each triple uniquely as an ordered triple of the form $(f(k), g_k(\ell), h_{k\ell}(v))$. A second triple $(f(k'), g_{k'}(\ell'), h_{k'\ell'}(v'))$ precedes the first if and only if either $k'<k$, or $k'=k$ and $\ell'<\ell$, or $k'=k$ and $\ell'=\ell$ and $v'<v$.

For $m>4$, ordered m-tuples would be defined in a similar manner and the new selection probabilities conditioned on $\binom{n}{m}+\binom{n}{m-1}\ldots+n+1$ events.

For $m=1$, the new selection probabilities are conditioned on the first member of the ordering $f(1),\ldots,f(n)$ in I if $I \neq \emptyset$, or on $\emptyset$ if $I = \emptyset$.

Note that if m exceeds the number of PSUs per stratum in the initial design it is possible that at least some ordered m-tuples cannot be subsets of I, in which case all such subsets should be excluded from the ordering and the set of events on which the new selection probabilities are conditioned. If no m-tuples can be a subset of I then the new selection probabilities are conditioned on I itself.

It is not necessary to limit the initial events used in the transportation problem to subsets of I of size m or less. For example, if $m=2$ and $\binom{n}{3}+\binom{n}{2}+n+1$ is sufficiently small then a procedure conditioned on subsets of three or less could be used resulting in a generally higher expected overlap. Conversely, if $\binom{n}{m}+\binom{n}{m-1}+\ldots+n+1$ is too large, the new selection probabilities could be conditioned on subsets of I of size m' or less where $m'<m$, although with resulting a smaller expected overlap.

3.2 An Overlap Procedure That Preserves Independence from Stratum to Stratum

The key to a modified overlap procedure that preserves the independence of the selection of sample PSUs from stratum to stratum in the new design if such independence existed in the selection of sample PSUs in the initial design is as follows.

Let $F_1,\ldots,F_r$ and $S_1,\ldots,S_t$ denote the set of strata in the initial and new designs respectively, and let I denote the set of initial sample PSUs across all initial design strata. With each S_j, $j=1,\ldots,t$, a subset S_j' of S_j is associated such that each distinct pair S_j', S_k' of such sets have no initial stratum in common, that is for each $i=1,\ldots,r$ either $S_j' \cap F_i = \emptyset$ or $S_k' \cap F_i = \emptyset$. Therefore, the set of PSUs in $I \cap S_j'$ and $I \cap S_k'$, were selected independently into the initial sample, even though this is not necessarily true for $I \cap S_j$ and $I \cap S_k$. Consequently, a modified overlap procedure which conditions the selection of new design sample PSUs for S_j on $I \cap S_j'$ instead of $I \cap S_j$, as in the original procedure of Causey, Cox and Ernst, would result in an independent selection from stratum to stratum of the new design sample PSUs.

A simple method of obtaining S_j', $j=1,\ldots,t$, satisfying the required condition is to associate to each initial stratum F_i a unique new stratum $S_{f(i)}$, by means of a mapping $f: \{1,\ldots,r\} \to \{1,\ldots,t\}$, and let

$$S_j' = S_j \cap \bigcup_{i \in f^{-1}(\{j\})} F_i, \qquad j=1,\ldots,t.$$

Appropriate choices for f will be discussed latter in this subsection.

The transportation problem to be solved for this modified overlap procedure can now be stated. As in the procedure presented in Causey, Cox and Ernst, each stratum in the new design requires the solution of a separate transportation problem. Dropping the subscript j, let S be a stratum in the new design with S' the corresponding subset as described above. Let $I_1,\ldots,I_m$ denote all possibilities for the subset of S' consisting of all PSUs in S' that were in the initial sample and let $N_1,\ldots,N_n$ denote all possibilities for the subset of S consisting of all new sample PSUs in S. For $i=1,\ldots,m$, $j=1,\ldots,n$, let p_i denote the probability that I_i was the set of initial sample PSUs in S', π_j the probability that N_j is the set

of new sample PSUs in S, x_{ij} the joint probability that both of these events occur, and c_{ij} the expected number of PSUs in $I \cap N_j$ given I_i. Again it is the x_{ij}'s that are the variables whose optimal values are to be determined.

Now proceed exactly as in Causey, Cox and Ernst, that is determine $x_{ij} \geq 0$ that maximize

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

subject to

$$\sum_{j=1}^{n} x_{ij} = p_i, \qquad i=1,\ldots,m,$$

$$\sum_{i=1}^{m} x_{ij} = \pi_i, \qquad j=1,\ldots,n.$$

Then, once the optimal x_{ij}'s have been obtained, the conditional new selection probabilities for N_j, $j=1,\ldots,n$, given I_i, are x_{ij}/p_i.

It remains to explain how to compute c_{ij}. Let $N_{j1},\ldots,N_{jk}$ denote the PSUs in N_j, and for $\ell=1,\ldots,k$ let

$$\begin{aligned} c'_{ij\ell} &= 1 \text{ if } N_{j\ell} \in I_i \cap S', \\ &= 0 \text{ if } N_{j\ell} \in S' \sim I_i, \\ &= P(N_{j\ell} \in I) \text{ if } N_{j\ell} \notin S'. \end{aligned}$$

Then

$$c_{ij} = \sum_{\ell=1}^{k} c'_{ij\ell}.$$

Although the procedure just described can be used with any mapping f, some mappings result in a larger expected number of PSUs retained in sample for the entire new design than other

mappings. A mapping that generally yields a relatively high overlap is obtained by choosing for each i=1,...,r an f(i) satisfying

$$\sum_{A \in F_i \cap S_{f(i)}} P(A \in I) = \max \left\{ \sum_{A_i \in F_i \cap S_j} P(A \in I) : j=1,\ldots,t \right\}, \qquad (3.1)$$

that is, $S_{f(i)}$ is a stratum in the new design which contains a largest piece of F_i as measured by the initial probabilities of selection. This approach maximizes

$$\sum_{j=1}^{t} \sum_{A \in S_j} P(A \in I).$$

An alternative f is obtained by replacing $P(A \in I)$ in (3.1) by the minimum of the probabilities of A being in sample for the initial and new designs, since the minimum of these two probabilities is the maximum probability with which A can be retained in the new sample.

Neither of these mappings is necessarily optimal in terms of maximizing the expected number of PSUs retained. A drawback to both of them is that they can associate several of the F_i with the same S_j, while other S_j may not be associated with any F_i. In theory, the selection of an optimal f could itself be made part of an optimization problem, but one which would involve all initial and new strata in a single problem, and thus tend to be unmanageably large.

Remark 3.1 The relative effectiveness of this modified procedure in retaining PSUs in the new design in comparison with the original procedure of Causey, Cox and Ernst depends heavily on how much the stratification for the new design differs from that of the initial design. In general, when the stratification does not differ much, the S_j's are a larger proportion of the the S_j's, which results in a better overlap.

Remark 3.2 If there are only a few initial strata that map onto S', then the set of initial outcomes may be reasonably small. If not, the procedure of this subsection can itself be modified using the procedure of Section 3.2. For example if both the initial and new designs are two PSUs per stratum and S' consists of s PSUs, then applying the procedure of Section 3.2 would reduce the number of possible initial outcomes used in the transportation problem to $\binom{s}{2}+s+1$ from a maximum of 2^s.

4. LINEAR PROGRAMMING AS AN ALTERNATIVE TO STRATIFICATION IN SELECTING SAMPLE PSUs

Consider a survey with a multistage design for which the PSUs are contiguous geographic areas. A common design technique to reduce between PSU variance is to partition the sets of PSUs into a collection of strata of approximately equal measures of size, with the PSUs in each stratum homogenous with respect to a key characteristic or characteristics of interest. The sample PSUs are then selected independently in each stratum with probability proportional to size. Stratification is generally effective in reducing between PSU variances. However, there are some disadvantages to this procedure. A key problem is that the process of forming strata, which fits into the general category of clustering problems, is often not an easy task. Furthermore, sometimes the deviations from the goal of equal-sized strata are nontrivial which tends to increase variances. If two or more surveys are to be designed together from stratified designs with the sample PSUs for one survey required to be a subset of the sample PSUs for the other, then techniques such as collapsing of strata may be necessary, which may not be highly efficient.

Linear programming is considered in this section as an alternative to stratification. This approach, as will be demonstrated, is conceptually very simple and extremely flexible, and software is readily available to solve linear programming problems. Unfortunately, there is a serious and, in many situations, fatal difficulty associated with the use of linear

programming in this context, namely that the size of the linear programming problem can readily get so large that it cannot be solved in practice even with powerful modern computers. However, as will be discussed, there are important situations where either this difficulty does not arise, or where some hybrid combination of linear programming and stratification may be feasible.

To state the problem to be considered more specifically, consider a multistage sample design for which there are N PSUs from which n are to be selected without replacement with probability proportional to size.

Let π_{ij} be the probability that the i-th PSU is in the sample of n PSUs and let π_{ij} be the probability that both the i-th and j-th PSUs are in sample. Let $\hat{Y}_i$ be an unbiased estimator of the i-th PSU total, Y_i, based on sampling at the second and subsequent stages. Then (Raj 1968) an unbiased estimator, $\hat{Y}$, of the population total Y is given by

$$\hat{Y} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i},$$

with variance

$$V(\hat{Y}) = \sum_{\substack{i,j \\ i<j}}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i} . \qquad (4.1)$$

Typically, in determining the sample design, the values of the π_i's and Y_i's are fixed beforehand from census data, for example. Then the between PSU variance component of $V(\hat{Y})$, which is

$$\sum_{\substack{i,j \\ i<j}}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 , \qquad (4.2)$$

would be minimized by the optimal choice for the π_{ij}'s

independently of the only other variables in (4.1), the σ_i^2's. This will be the focus of the work in this section, the minimization of (4.2) by optimal choice of the π_{ij}'s.

For an optimal set of π_{ij}'s, because of the minus sign preceding π_{ij} in (4.2), the quantity $(Y_i/\pi_i - Y_j/\pi_j)^2$ would tend to be large for those i,j for which π_{ij} is large and, likewise, these two quantities would also tend to be small together.

The π_{ij}'s resulting from a stratification typically yield such a relationship. For example, if the n sample PSUs are selected by partitioning the N PSUs into n equal-sized strata, with PSUs in the same stratum having values of Y_i/π_i as close together as possible, and selecting one PSU per stratum with probability proportional to size, then $\pi_{ij}=0$ if the i-th and j-th PSUs are in the same stratum and $\pi_{ij}=\pi_i\pi_j$ otherwise. Likewise, if two PSUs per stratum are selected using the Brewer-Durbin procedure, for example, or three or more using its generalization by Sampford (Cochran, 1977), then $\pi_{ij}<\pi_i\pi_j$ if the pair of PSUs are in the same stratum and $\pi_{ij}=\pi_i\pi_j$ otherwise.

However, linear programming can attack the problem of minimizing (4.2) more directly than stratification. (4.2) is linear in the only variables, the π_{ij}'s, so it is only necessary to minimize this objective function with respect to these variables subject to appropriate linear constraints on the π_{ij}'s. In order to insure that the i-th PSU is selected with the required probability, π_i, for each i, the following set of constraints must be satisfied:

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \pi_{ij}=(n-1)\pi_i, \quad i=1,\ldots,N. \tag{4.3}$$

If selecting PSUs with predetermined probabilities is the only design requirement, then this would be the only set of constraints needed. However, other requirements, such as the ability to obtain variance estimates with desirable properties would lead to additional constraints as will be described later.

A set of π_{ij}'s satisfying (4.3) always exists, since the π_{ij}'s arising from the use of Sampford's method yields one solution. Unfortunately, for n>2, there does not necessarily exist a set of selection probabilities attached to the set of distinct n-tuples of PSUs which satisfies an optimal solution to the problem of minimizing (4.2) subject to (4.3), that is there may be no sampling procedure which actually yields the optimal π_{ij}'s. For example, if N=4, n=3, $Y_1/\pi_1 = Y_2/\pi_2$ and $Y_3/\pi_3 = Y_4/\pi_4$, then the following set of π_{ij}'s minimize (4.2) subject to (4.3).

$$\pi_{12} = \pi_{34} = 0, \tag{4.4}$$

$$\pi_{13} = \pi_{14} = \pi_{23} = \pi_{24} = 3/4. \tag{4.5}$$

However, if π_{ijk} denotes the probability that the sample consists of the i-th, j-th and k-th PSUs, then π_{ijk} must be 0 for all four distinct triples in order for (4.4) to be satisfied, in which case (4.5) is not satisfied and thus there is no set of π_{ijk}'s satisfying (4.4) and (4.5) simultaneously.

To avoid this problem in the case n=3, the π_{ijk}'s could be used as variables in (4.2) and (4.3) in place of the π_{ij}'s by replacing π_{ij} by $\sum_{\substack{k \\ k\neq i,j}}^{N} \pi_{ijk}$ in (4.2) and (4.3). This substitution in (4.3) would reduce to

$$\sum_{\substack{j,k \\ j<k}}^{N} \pi_{ijk} = \pi_i, \quad i=1,\ldots,N.$$

Similarly for general n, if S denotes the set of distinct n-tuples of PSUs and for each $s\in S$, π_s' denotes the probability that s is selected, then if $\sum_{\substack{s\in S \\ i,j\in s}} \pi_s'$ is substituted for π_{ij} in (4.2) and (4.3), these expressions become respectively

$$\sum_{\substack{i,j \\ i<j}}^{N} [\pi_i \pi_j - \sum_{\substack{s \in S \\ i,j \in s}} \pi_s'] \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 , \tag{4.6}$$

and

$$\sum_{\substack{s \in S \\ i \in s}} \pi_s' = \pi_i, \quad i=1,\ldots,N. \tag{4.7}$$

Since a solution to the optimization problem (4.6), (4.7) immediately yields selection probabilities for each possible n-tuple of PSUs, the difficulty described with the formulation (4.2) and (4.3) cannot occur. Furthermore, Sampford's method always provides a feasible solution to (4.7). However, in practice, a possibly insurmountable operational problem can occur. The number of variables in (4.6) and (4.7) is $\binom{N}{n}$, which can be impractically large. Thus the use of this procedure appears to be limited to cases where $\binom{N}{n}$ does not exceed the software and hardware limitations of the available equipment.

This method could be potentially applicable to the Current Population Survey, which has a state based design, and hence a separate linear programming problem for each state. For the smaller states at least, $\binom{N}{n}$ may be sufficiently small.

If $\binom{N}{n}$ is too large to use the linear programming formulation directly, a hybrid of statification and linear programming could be used. With this approach, stratification would first be used to partition the population of PSUs into a number of super-strata and linear programming then used to select the sample PSUs from each super-stratum. The number of super-strata would be smaller than if stratification were used alone but there would have to be enough super-strata to insure that the linear programming problem corresponding to each super-stratum was sufficiently small.

When the problem of minimizing (4.6) subject to (4.7) is sufficiently small to solve, there are at least two additional set of constraints that might be added to the problem in order to be able to produce variance estimates with desirable properties. They are

$$\sum_{\substack{s\in S \\ i,j\in s}} \pi_s' \leq \pi_i\pi_j, \quad i,j=1,\ldots,N, \quad i\neq j, \tag{4.8}$$

$$\sum_{\substack{s\in S \\ i,j\in s}} \pi_s' \geq c\pi_i\pi_j, \quad i,j=1,\ldots,N, \quad i\neq j, \tag{4.9}$$

where $c<1$ is a constant. (4.8) and (4.9) are equivalent to $\pi_{ij}\leq\pi_i\pi_j$ and $\pi_{ij}\geq c\pi_i\pi_j$ respectively. The reasons for requiring these sets of constraints are as follows. If (Raj 1968) $\hat{Y}_i$ and $\hat{\sigma}_i^2$ are unbiased estimators of Y_i and σ_i^2 respectively, $i=1,\ldots,N$, then provided $\pi_{ij}>0$ for all $i,j=1,\ldots,N$, $i\neq j$, an unbiased estimator of (4.1) is

$$v(\hat{Y}) = \sum_{\substack{i,j \\ i<j}}^{n} \left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{\hat{Y}_i}{\pi_i}-\frac{\hat{Y}_j}{\pi_j}\right)^2 + \sum_{i=1}^{n}\frac{\hat{\sigma}_i^2}{\pi_i}. \tag{4.10}$$

(4.8) is needed to insure that $v(\hat{Y})$ is always nonnegative. Without (4.9), π_{ij} could be 0 for some i,j, in which case $v(\hat{Y})$ is not unbiased. Furthermore, (4.9) forces an upper bound of $1/c-1$ on $(\pi_i\pi_j-\pi_{ij})/\pi_{ij}$. The variance of $v(\hat{Y})$, for a solution to the optimization problem that includes (4.9), generally decreases as c increases, since $1/c-1$ decreases with increasing c. On the other hand $V(\hat{Y})$ increases with increasing c since the set of feasible solutions to (4.9) becomes smaller with increasing c. If c becomes too large there are no feasible solutions to the optimization problem. Thus the selection of a value for c in (4.9) involves a tradeoff between decreasing $V(\hat{Y})$ and the variance of $v(\hat{Y})$. The determination of a c which optimally

balances these two goals would have to be obtained by trial and error or through the solution of a nonlinear programming problem.

Until now the problem of minimizing between PSU variance using linear programming has been considered with respect to only a single characteristic. However, a virtually identical approach can be used to minimize certain types of averages of the between PSU variances for several characteristics. For example, to minimize an average of the variances for r characteristics, $(Y_i/\pi_i - Y_j/\pi_j)^2$ in (4.2) might be replaced by

$$\sum_{k=1}^{r} W_k \, (Y_{ik}/\pi_i - Y_{ik}/\pi_j)^2, \qquad (4.11)$$

where Y_{ik} is the total for the k-th characteristic in the i-th PSU. W_k would be either a scaling factor or a preference factor or some combination of the two types of factors (see Kostanich et al. 1981). Since all the quantities in (4.11) are assumed known, substitution of (4.11) into (4.2) as described does not change the form of the optimization problem.

Linear programming is also applicable to the selection of sample PSUs for two or more designs when the samples are not selected independently from design to design, again assuming that the resulting problem is not unmanageably large. For example, suppose two samples s_1 and s_2 are to be chosen for designs 1 and 2 respectively with the requirement that $s_2 \subset s_1$. One approach to this task, not involving linear programming, is to select the sample PSUs for design 1 first from a stratification and then to collapse the design 1 strata together to form the design 2 strata, from each of which a subsample of the design 1 sample PSUs is chosen. The design 2 strata formed in this way may neither be as homogenous nor as nearly equal-sized as they would be if designs 1 and 2 were independent, resulting in larger between PSU variances. This may be particularly true when the number of design 1 strata are small. Similar problems would

arise if the design 2 sample was selected first and the design 1 sample obtained by adding PSUs.

A linear programming approach can avoid these difficulties. Consider the following formulation. For designs 1 and 2, n_1 and n_2 PSUs are to be selected respectively without replacement, from the same population of N PSUs, with $\pi_i^{(1)}$, $\pi_i^{(2)}$ the probability that the i-th PSU is in the sample for designs 1 and 2 respectively. Note that if the same measure of size is used for both designs then $\pi_i^{(1)}/\pi_i^{(2)} = n_1/n_2$. Let S' denote the set of all possible joint samples of PSUs for the two designs, that is all ordered pairs (s_1, s_2) of distinct n_1-tuples and n_2-tuples respectively with $s_2 \subset s_1$, and let π'_{s_1,s_2} denote the probability that s_1 and s_2 are the set of sample PSUs for designs 1 and 2 respectively. The π'_{s_1,s_2}'s are the variables in the linear programming problem. A two design analogue of objective function (4.6) is then

$$\sum_{k=1}^{2} W_k \sum_{\substack{i,j \\ i<j}}^{N} \left(\pi_i^{(k)}\pi_j^{(k)} - \sum_{\substack{(s_1,s_2)\in S' \\ i,j\in s_k}} \pi'_{s_1,s_2}\right)\left(\frac{Y_{ik}}{\pi_i^{(k)}} - \frac{Y_{jk}}{\pi_{jk}^{(k)}}\right)^2, \qquad (4.12)$$

where Y_{ik} is the population total for the design k characteristic and W_k is a weighting factor that can serve as a combination of scaling factor and preference factor for design k. (4.12) is easily generalized to situations where there is more than one characteristic in the optimization problem for each design.

The two design analogue of constraints (4.7), (4.8) and (4.9) are respectively

$$\sum_{\substack{(s_1,s_2)\in S' \\ i\in s_k}} \pi'_{s_1,s_2} = \pi_i^{(k)}, \quad i=1,\ldots,N, \quad k=1,2,$$

$$\sum_{\substack{(s_1,s_2)\in S' \\ i,j\in s_k}} \pi'_{s_1,s_2} \leq \pi_i^{(k)}\pi_j^{(k)}, \quad i,j=1,\ldots,N,\ i\neq j,\ k=1,2,$$

$$\sum_{\substack{(s_1,s_2)\in S' \\ i,j\in s_k}} \pi'_{s_1,s_2} \geq c\pi_i^{(k)}\pi_j^{(k)}, \quad i,j=1,\ldots,N,\ i\neq j,\ k=1,2.$$

In this formulation the selection probabilities for each PSU in each design are exactly proportional to the measure of size, avoiding the problems arising from collapsing of strata. Furthermore, the weights W_1, W_2 can be used to balance the optimality of the two design in a simple manner. The effect on the problem of modification of requirements can also be easily ascertained with the linear programming formulation. For example, the effect on variances of the requirement that $s_2 \subset s_1$ can be determined by computing the minimal values of the objective function with and without this requirement.

The planned CPS expansion is a potential application of the two design, linear programming formulation. Under the current proposal sometime after the new CPS design is phased in, beginning in 1994, an expansion will take place that will enable monthly estimates to be published for all 50 states and the District of Columbia. Currently only annual average estimates are produced for all but the eleven largest states. For each state, s_2 and s_1 can be considered the design before and after the expansion respectively.

REFERENCES

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

Cochran, William G. (1977), Sampling Techniques, Third Edition, New York: John Wiley and Sons.

Cox, Lawrence H., and Ernst, Lawrence R. (1982), "Controlled Rounding," INFOR, 20, 423-432.

Ernst, Lawrence R. (1986), "Maximizing the Overlap Between Surveys When Information Is Incomplete," European Journal of Operational Research, 27, 192-200.

Fagan, J.T., Greenberg, B.V., and Hemming, B. (1988), "Controlled Rounding of Three Dimensional Tables," Report Census/SRD/RR-88/02, U.S. Bureau of the Census, Statistical Research Division.

Keyfitz, Nathan (1951), "Sampling With Probabilities Proportional to Size: Adjustment for Changes in Probabilities," Journal of the American Statistical Association, 66, 461-470.

Kostanich, D., Judkins, D., Singh, R., and Schautz, M. (1981), "Modification of Freedman-Rubin's Clustering Algorithm for Use in Stratified PPS Sampling," Proceedings of the Section on Survey Research Methods, American Statistical Association, 285-290.

Raj, Des (1968), Sampling Theory, New York: McGraw Hill.