FURTHER APPLICATIONS OF LINEAR PROGRAMMING TO SAMPLING PROBLEMS

Lawrence R. Ernst, Bureau of the Census*
Washington, DC 20233

KEY WORDS: optimizing designs, maximizing overlap, controlled rounding

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. 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.

Due to space limitations the full paper is not presented here. All proofs, the list of references and some exposition have been omitted. The complete paper is available from the author.

2. THREE-DIMENSIONAL CONTROLLED ROUNDINGS

Cox and Ernst (1982) 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

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

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

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

Cell (i,j,k) is an internal cell if $i<m+1$, $j<n+1$ and $k<\ell+1$. If equality replaces strict inequality in any of these relations then the cell is a marginal of dimension equal to the number of indices for which equality holds. This definition 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).

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

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

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

where [ ] denotes the greatest integer function. 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 adding the requirement that $r_{ijk} = a_{ijk}$ if a_{ijk} is a multiple of b.

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 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} is an integral multiple of b and

$$|r_{ijk} - a_{ijk}| < 2b \text{ for all } i,j,k. \qquad (2.7)$$

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 $I(A)=(a_{ijk})$, $1\leq i\leq m$, $1\leq j\leq n$, $1\leq k\leq \ell$.

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, that is as a set of two-dimensional tabular arrays corresponding to k=1,2,3,4,5.

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

$$b'_{ijk} = b_{ijk} \quad \text{if } 1\leq i\leq 4, \quad 1\leq j\leq 4,$$

$$= b_{(i-4)(j-4)k} \quad \text{if } 5\leq i\leq 8, \quad 5\leq j\leq 8,$$

$$= b_{(i-8)(j-8)k} \quad \text{if } 9\leq i\leq 12, \quad 9\leq j\leq 12,$$

$$= 0 \quad \text{for all other } i,j,k.$$

The proof that B' has no controlled roundings is presented in the complete paper.

Remark 2.1: All the marginals of B' are integers. 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

$$c_{ijk} = \sum_{t=1}^{k} a_{ijt} - \sum_{t=1}^{k-1} r_{ijt},$$

$$1\leq i\leq m+1, \quad 1\leq j\leq n+1, \qquad (2.19)$$

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

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

The proof that the array just defined satisfies the required properties is presented in the complete paper.

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. 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. The description and solution to one of these difficulties is presented in Ernst

(1986). In this section approaches are presented for handling two other problems.

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.

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.

3.1 A Reduced-Size Transportation Problem for Maximizing Overlap

The reduced-size procedure will, for ease of presentation, be described here only for the case when both the initial and new designs are two PSUs per stratum without replacement. Many of the details even for that case have been omitted here but are available in the full paper, where, in addition, the changes necessary to apply this procedure for other initial and new designs are 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.

The details of the ordering of the pairs are presented in the full paper. From this ordering, the following ordering, $I_1,I_2,\ldots,I_{\nu+n+1}$ of all subsets of $\{1,\ldots,n\}$ of two or fewer elements is constructed, where the abbreviation $\nu=\binom{n}{2}$ is used. $I_1,\ldots,I_\nu$ consists of the pairs of integers determined by their ordering. $I_{\nu+1}, \ldots, I_{\nu+n}$ consists of the n singleton sets in any ordering, and $I_{\nu+n+1} = \emptyset$. For each I, a unique I_i is associated, namely the first I_i in the sequence for which $I_i \subset I$. For each I it is the associated I_i rather than I itself on which the new selection probabilities are conditioned, and thus the new selection probabilities are conditioned on only $\nu+n+1$ events. For each $i=1,\ldots,\nu+n+1$, p_i^* denote the probability that I_i is the subset associated with I. Formulas for computing p_i^* are presented in the complete paper.

As for the new sample, there are ν possibilities, denoted S_j, $j=1,\ldots,\nu$, for N. If $S_j = \{s,t\}$, then π_j^*, the probability that $N=S_j$, is simply the predetermined probability that both s and t are in the new sample.

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

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

subject to

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

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

Once the optimal x_{ij}'s have been obtained, the conditional new selection probabilities for

S_j, $j=1,\ldots,\nu$, given I_i, are x_{ij}/p_i^*.

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 with 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 are presented in the complete paper.

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

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

Then

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

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.

Although stratification tends to lower (4.2), as explained in the complete paper, linear programming can attack the problem of minimizing (4.2) more directly. (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:

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

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):

$$\pi_{12} = \pi_{34} = 0, \qquad (4.4)$$

$$\pi_{13} = \pi_{14} = \pi_{23} = \pi_{24} = 3/4. \qquad (4.5)$$

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 for general n, let S denote the set of distinct n-tuples of PSUs and for each $s\in S$, π_s' denote 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

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

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

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

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. This is explained in the complete paper.

*This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.