Survey Methodology, December 1995
Vol. 21, No. 2, pp. 147–157
Statistics Canada

A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys

LAWRENCE R. ERNST and MICHAEL M. IKEDA[1]

ABSTRACT

When redesigning a sample with a stratified multi-stage design, it is sometimes considered desirable to maximize the number of primary sampling units retained in the new sample without altering unconditional selection probabilities. For this problem, an optimal solution which uses transportation theory exists for a very general class of designs. However, this procedure has never been used in the redesign of any survey (that the authors are aware of), in part because even for moderately-sized strata, the resulting transportation problem may be too large to solve in practice. In this paper, a modified reduced-size transportation algorithm is presented for maximizing the overlap, which substantially reduces the size of the problem. This reduced-size overlap procedure was used in the recent redesign of the Survey of Income and Program Participation (SIPP). The performance of the reduced-size algorithm is summarized, both for the actual production SIPP overlap and for earlier, artificial simulations of the SIPP overlap. Although the procedure is not optimal and theoretically can produce only negligible improvements in expected overlap compared to independent selection, in practice it gave substantial improvements in overlap over independent selection for SIPP, and generally provided an overlap that is close to optimal.

KEY WORDS: Linear programming; Sample redesign; Survey of Income and Program Participation.

1. INTRODUCTION

The problem of maximizing the expected number of primary sampling units (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). Typically, the motivation for maximizing the overlap of PSUs is to reduce additional costs, such as the training of a new interviewer for a household survey, incurred with each change of sample PSU. Procedures for maximizing overlap do not alter the unconditional probability of selection for a set of PSUs in a new stratum, but conditions its probability of selection in such a manner that the probability of a PSU being selected in the new sample is generally greater than its unconditional probability when the PSU was in the initial sample and less otherwise.

Overlap procedures are applicable when the redesign results in either a restratification of the PSUs or a change in their selection probabilities. Keyfitz (1951) presented an optimal procedure, but only for one-PSU-per-stratum designs in the special case when the initial and new strata are identical, with only the selection probabilities changing. Causey, Cox and Ernst (1985) obtained an optimal solution to the overlap problem under very general conditions by formulating it as a transportation problem, which is a special form of linear programming problem. This procedure imposes no restrictions on changes in strata definitions or number of PSUs per stratum. (A similar result had been independently obtained by Arthanari and Dodge (1981), although they did not discuss the issue of changes in strata definitions. Both sets of authors obtained their results by generalizing work of Raj (1968).) However, there are at least two other difficulties with the procedure of Causey, Cox and Ernst which can make it unusable in practice, one which is the focus of Ernst (1986), and the other the focus of the current paper.

The first difficulty is that, if the initial sample of PSUs was not selected independently from stratum to stratum, the information necessary to compute all the joint probabilities required by this method may not be available in practice. An alternative linear programming procedure, for use in such cases, was developed by Ernst (1986). The Bureau of the Census has used linear programming to overlap its demographic surveys on five occasions. On four of these occasions (the selection of the 1980s and 1990s Current Population Survey (CPS) designs, and the 1980s and 1990s National Crime Victimization Survey (NCVS) designs) the procedure in Ernst (1986) was used because the initial design was not selected independently from stratum to stratum. In particular, as explained in Ernst (1986), if the initial sample was itself selected by overlapping with a still earlier design then this independence assumption generally does not hold, which was the key reason why it did not hold for these four redesigns.

The second difficulty with the optimal procedure is that the transportation problem may be too large to solve in practice. The Bureau of the Census also used linear

[1] Lawrence R. Ernst, Chief, Research Group, Office of Compensation and Working Conditions, Bureau of Labor Statistics, Washington, DC 20212, U.S.A.; Michael M. Ikeda, Mathematical Statistician, Statistical Research Division, Bureau of the Census, Washington, DC 20233, U.S.A.

programming to overlap the 1990s Survey of Income and Program Participation (SIPP) design with the 1980s SIPP design, both two-PSUs-per-stratum designs. The initial sample for SIPP was selected independently from stratum to stratum. However, the transportation problem for the optimal procedure would have been too large to practically solve for many strata. This is because for each new stratum to be overlapped consisting of n PSUs, the number of variables in the transportation problem for the optimal procedure can be as large as $2^n \times \binom{n}{2}$. The largest value of n for which a transportation problem with that many variables can be solved with the computer facilities that we have used is approximately $n = 15$.

This paper presents a reduced-size formulation of the overlap procedure as a transportation problem which decreases the numbers of variables in the SIPP problem to $\left(\binom{n}{2} + n + 1\right) \times \binom{n}{2}$, a striking reduction for moderate to large values of n. The procedure assumes that the initial sample was selected independently from stratum to stratum, and hence could not have been used instead of the procedure of Ernst (1986) to overlap the CPS and NCVS designs. This reduced-size procedure has been successfully run for strata with as many as 68 PSUs. In contrast, for $n = 68$, the $2^{68} \times \binom{68}{2}$ possible number of variables for the unreduced formulation is far beyond the size of problem that can be solved by any current computer. Furthermore, though the reduced-size procedure sacrifices optimality in exchange for its size reduction, it does appear in practice to yield results fairly close to optimal, as we will show. The reduced-size procedure is the procedure that was used to overlap SIPP.

In Section 2 the procedure of Causey, Cox and Ernst (1985) is reviewed, to provide background for the presentation of the reduced-size procedure.

The reduced-size procedure is presented in Section 3. Although the approach has general applicability, for ease of presentation it is only described in detail for the case when both the initial and new designs are two-PSUs-per-stratum without replacement. A small, artificial example of the reduced-size procedure is also presented in Section 3. This example serves to illustrate the procedure and to demonstrate that the ordering of the pairs of PSUs in a new design stratum, a key step in the algorithm, affects the expected overlap. We also outline in this section some analytical results on the comparison between the reduced-size procedure and the optimal procedure. Upper bounds on the loss in expected overlap from using the reduced-size procedure instead of the optimal procedure are stated. It is also explained that in certain situations this loss can approach two PSUs for two-PSUs-per-stratum designs, the worst possible situation. Further details and proofs of the results in this section as well as some results in other sections are presented in Ernst and Ikeda 1994.

In Section 4 the performance of the reduced-size procedure is presented, both for the actual SIPP production overlap and for earlier, artificial simulations of the SIPP overlap. The expected overlap for this procedure is compared to that for independent selection of the new sample PSUs and to an upper bound on the optimal expected overlap. The results show that for this application, in contrast with some of the theoretical results described in Section 3, the expected overlap with the reduced-size procedure is much larger than if independent selection had been used to select the new sample PSUs, and nearly as large as the optimal expected overlap. Also presented are computer running times for the reduced-size procedure as a function of stratum size.

Finally, our conclusions are stated in Section 5.

2. REVIEW OF THE OVERLAP PROCEDURE OF CAUSEY, COX AND ERNST (1985)

The overlap procedure of Causey, Cox and Ernst (1985), like all overlap procedures, conditions the selection of sample PSUs in each new stratum in some way on which PSUs in the stratum were in the initial sample. This particular overlap procedure attains true optimality by making complete use of this information and formulating the procedure as a transportation problem. We proceed to present this procedure.

First, however, we introduce some notation that will be used throughout the paper. Let S denote a stratum in the new design. Each such stratum corresponds to a separate overlap problem. Let n denote the number of PSUs in S and let $A_1, \ldots, A_n$ denote the PSUs in S. Let I denote the random subset of $\{1, \ldots, n\}$ such that $k \in I$ if and only if A_k was in the initial sample, and let N denote the corresponding set with respect to the new sample. For example, if A_2 and A_3 were the PSUs in S that were in the initial sample and A_1 and A_3 are the PSUs in the new sample, then $I = \{2,3\}$ and $N = \{1,3\}$. Let m^*, n^* denote the number of possible values for I and N, respectively. Let J_i, $i = 1, \ldots, m^*$, denote the possible values for I and let S_j, $j = 1, \ldots, n^*$, denote the possible values for N. The goal of all overlap procedures is to maximize the expected number of PSUs in $N \cap I$, while preserving the values of the $P(S_j)$'s.

To illustrate some of these concepts further, consider an example for which $n = 3$. Then $n^* = 3$ if the new design is either 1 or 2 PSUs per stratum with the values for N, that is the S_j's, consisting of $\{1\},\{2\},\{3\}$ in the 1 PSU per stratum case and $\{1,2\},\{1,3\},\{2,3\}$ in the two PSUs per stratum case. Suppose PSUs A_1 and A_2 were in one initial stratum and PSU A_3 was in another initial stratum and there were three PSUs in each of these initial strata. If the initial design was 1 PSU per stratum, then $m^* = 6$, with the values of I, that is the J_i's, consisting of $\varnothing$, $\{1\},\{2\},\{3\},\{1,3\},\{2,3\}$; if the initial design was 2 PSUs per stratum then $m^* = 6$, with the J_i's consisting of $\{1\},\{2\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}$.

We now present the transportation problem for the overlap procedure of Causey, Cox and Ernst (1985). Abbreviate by $P(J_i)$ the probability that $I = J_i$ and by $P(S_j)$ the probability that $N = S_j$. In addition, let x_{ij} be the variable denoting the joint probability of these two events, and let c_{ij} denote the number of elements in $J_i \cap S_j$. The $P(J_i)$'s, $P(S_j)$'s and c_{ij}'s are known values, while the x_{ij}'s are variables for which the optimal values are to be determined. Then the transportation problem to solve is to determine $x_{ij} \geq 0$ which maximize

$$\sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij} \tag{2.1}$$

subject to

$$\sum_{j=1}^{n^*} x_{ij} = P(J_i), \quad i = 1, \ldots, m^*, \tag{2.2}$$

$$\sum_{i=1}^{m^*} x_{ij} = P(S_j), \quad j = 1, \ldots, n^*. \tag{2.3}$$

Note that in this transportation problem, the objective function (2.1) is the expected number of PSUs in S that are in $N \cap I$. Also note that the constraints (2.2) and (2.3) are required by the definitions of the $P(J_i)$'s, $P(S_j)$'s and the x_{ij}'s.

Once the optimal x_{ij}'s have been obtained, the conditional probability that $N = S_j$ given that $I = J_i$ is then $x_{ij}/P(J_i)$ for all i,j.

We present an example to illustrate the use of the formulation (2.1)-(2.3) in the case where both the initial and new designs are two-PSUs-per-stratum without replacement. In this example, and throughout the paper, p_i, π_i denote the predetermined probability that $i \in I$ and $i \in N$, respectively.

Consider a final stratum S with $n = 3$. All of the PSUs were in different initial strata. Let $p_1 = .6$, $p_2 = .75$, $p_3 = .7$, $\pi_1 = .5$, $\pi_2 = .8$, $\pi_3 = .7$. Since the PSUs were all in different initial strata, there are 8 different possibilities for I, with probabilities given in Table 1.

Table 1

Probabilities for Possible Sets of Initial Sample PSUs

i	1	2	3	4	5	6	7	8
J_i	{1,2,3}	{1,2}	{1,3}	{2,3}	{1}	{2}	{3}	∅
$P(J_i)$	.315	.135	.105	.21	.045	.09	.07	.03

Since the new design is two-PSUs-per-stratum without replacement, there are 3 different possibilities for N, namely the pairs $S_1 = \{1,2\}$, $S_2 = \{1,3\}$, $S_3 = \{2,3\}$, and hence $P(S_1) = .30$, $P(S_2) = .20$, $P(S_3) = .50$.

Furthermore, the values of c_{ij} are then as given in Table 2. Upon maximizing (2.1) subject to (2.2) and (2.3) with the given $P(J_i)$'s, $P(S_j)$'s and c_{ij}'s, an optimal set of x_{ij}'s, presented in Table 2, is obtained. Finally, by dividing each of the x_{ij} entries in row i of Table 2 by $P(J_i)$, an optimal set of conditional probabilities $P(S_j \mid J_i)$, is obtained. For example, since $x_{12} = .025$ and $P(J_1) = .315$, it follows that $P(S_2 \mid J_1) = 5/63$.

Table 2

Values of c_{ij} and Values of x_{ij} that Maximize Overlap for Optimal Procedure

	c_{ij}			x_{ij}		
	j			j		
i	1	2	3	1	2	3
1	2	2	2	.000	.025	.290
2	2	1	1	.135	.000	.000
3	1	2	1	.000	.105	.000
4	1	1	2	.000	.000	.210
5	1	1	0	.045	.000	.000
6	1	0	1	.090	.000	.000
7	0	1	1	.000	.070	.000
8	0	0	0	.030	.000	.000

For this example, as can be computed from (2.1) and Table 2, the expected overlap under the optimal procedure is 1.735 PSUs. In comparison, the expected overlap if the initial and final designs are selected independently is $p_1\pi_1 + p_2\pi_2 + p_3\pi_3 = 1.39$ PSUs.

For two-PSU-per-stratum without replacement problems, the possible values for N are always the $\binom{n}{2}$ subsets of $\{1, \ldots, n\}$ of size 2, that is $n^* = \binom{n}{2}$. However m^* can vary widely. $m^* = \binom{n}{2}$ when the PSUs in S comprise a single initial stratum. The upper bound of 2^n on m^* is attained when all the PSUs in S were in different initial strata, as illustrated by the previous example, and in some other situations. A general, exact expression for m^* is presented in Ernst and Ikeda (1994).

For the two-PSUs-per-stratum without replacement overlap problem, the number of variables in the transportation problem for the optimal procedure is m^*n^* which can be as large as $2^n\binom{n}{2}$. For $n = 15$, $2^n\binom{n}{2} = 3{,}440{,}640$, which is about as large a transportation problem as can be solved with the computer facilities that we used. However, $n > 15$ for nearly half the nonselfrepresenting strata (that is strata consisting of noncertainty PSUs) in our SIPP application, and consequently it was necessary to develop a procedure, described in the next section, which reduces the size of the transportation problem, while still producing nearly maximal expected overlap in practice.

3. THE ALGORITHM FOR THE REDUCED-SIZE PROCEDURE

Previous work on reducing the size of the transportation problem (2.1)–(2.3) has focused on accomplishing the size reduction while retaining optimality. For example, the approach of Aragon and Pathak (1990) retains optimality and reduces the size of the problem by 75 percent when $m^* = n^*$. Unfortunately, when m^* is much larger than n^*, which is when size reduction is most needed, their method produces negligible size reduction in relative terms. A generalization of this approach is presented in Pathak and Fahimi (1992), but there is no indication that their procedure always yields a size reduction that is substantial in relative terms.

In this section a reduced-size procedure is presented which takes a different approach. We sacrifice optimality, at least in theory, in return for an assured size reduction down to a manageable size transportation problem. This size reduction is accomplished, in the case when the initial and new designs are both two PSUs per stratum for example, by ordering all pairs of PSUs in a new stratum and then conditioning the new selection probabilities for any initial set of sample PSUs of size greater than 2 on the first pair of PSUs in the ordering contained in the initial set, rather than conditioning on the entire initial set. That is, each possible initial set of sample PSUs which consists of more than 2 PSUs is combined with a set of size 2. As illustrated in Section 4, this procedure may yield a near optimal overlap in practice; particularly with an appropriate ordering of the pairs of PSUs, as described in Section 3.1.2.

The reduced-size procedure is applicable whenever PSUs in the initial and new designs are selected without replacement. However, the procedure will be described in detail, in Section 3.1, only for the case when both the initial and new designs are two-PSUs-per-stratum. Then, in Section 3.2, the changes necessary to apply this procedure for other initial and new designs will be sketched. Finally, in Section 3.3, some analytical results are outlined on the relationships among the expected overlap for the reduced-size procedure, the optimal procedure and independent selection. It is assumed throughout this section that PSUs in the initial sample were selected independently from stratum to stratum.

3.1 Reduced-Size Procedure When Both Designs Are Two-PSUs-Per-Stratum

The reduced-size procedure to be described includes the following key aspects: the specific ordering of the pairs of PSUs; the reformulation of the transportation problem (2.1)–(2.3) for the reduced size procedure; the computation of the probabilities for the initial outcomes for this formulation; and the computation of the cost coefficients (the c_{ij}'s) in the objective function. In Section 3.1.1 we present a detailed outline of the reduced-size procedure, including the reformulated transportation problem. The ordering of the pairs is described in Section 3.1.2. Finally, the computation of the probabilities for the initial outcomes and the cost coefficients are given in Section 3.1.3.

3.1.1 General Outline of the Procedure

The general outline of the procedure is as follows. First, the $\binom{n}{2}$ subsets of $\{1, \ldots, n\}$ of size 2 are ordered in a manner to be described later. (For now, we simply note that any ordering can be used to reduce the size of the transportation problem. The specific one used is for the purpose of accomplishing the size reduction while also attempting to give up as little as possible of the gains in overlap that the optimal procedure yields.) We let I_i, $i = 1, \ldots, \binom{n}{2}$, denote the i-th element in the ordering; let $I_{\binom{n}{2}+1}, \ldots, I_{\binom{n}{2}+n}$ be the n singleton subsets; and set $I_{\binom{n}{2}+n+1} = \emptyset$. Thus, the I_i's constitute all subsets of $\{1, \ldots, n\}$ of 2 or fewer elements. For each possibility for I, a unique set I^* is associated among these $\binom{n}{2} + n + 1$ subsets and the new selection probabilities conditioned on the associated I^*, rather than on I itself. Therefore, the new selection probabilities are conditioned on $\binom{n}{2} + n + 1$ events instead of a possible 2^n events, which is the reason for the size reduction. The associated I^* is the first I_i for which $I_i \subset I$. That is, if I consists of at least two integers, the associated I^* is the first pair in the ordering contained in I, while if I is a singleton set or empty then $I^* = I$.

The reduced-size transportation problem attempts to retain the PSUs corresponding to elements in the associated set I^* in the new sample, but does not use information on elements in $I \sim I^*$. The form of this reduced-sized transportation problem based on the set of I_i's is as follows. Let p_i^* be the probability that $I^* = I_i$, $i = 1, \ldots, \binom{n}{2} + n + 1$, and abbreviate $\pi_j^* = P(S_j)$, $j = 1, \ldots, \binom{n}{2}$. For each i,j, the variable x_{ij} is the joint probability that $I^* = I_i$ and that $N = S_j$, while c_{ij} is the expected number of elements in $I \cap S_j$ given $I^* = I_i$. The problem to solve is to determine $x_{ij} \geq 0$ that maximize

$$\sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij} x_{ij}, \tag{3.1}$$

subject to

$$\sum_{j=1}^{\binom{n}{2}} x_{ij} = p_i^*, \quad i = 1, \ldots, \binom{n}{2} + n + 1, \tag{3.2}$$

$$\sum_{i=1}^{\binom{n}{2}+n+1} x_{ij} = \pi_j^*, \quad j = 1, \ldots, \binom{n}{2}. \tag{3.3}$$

Once the optimal x_{ij}'s have been obtained, then the conditional new selection probabilities for $S_j, j = 1, \ldots, \binom{n}{2}$, given $I^* = I_i$, are x_{ij}/p_i^*. Note that the number of variables, x_{ij}, in the formulation (3.1)-(3.3) is $\left(\binom{n}{2} + n + 1\right) \times \binom{n}{2}$, in comparison with a maximum of $2^n \times \binom{n}{2}$ in the formulation (2.1)-(2.3).

It remains to explain the general method for obtaining the ordering of the $\binom{n}{2}$ pairs and the procedures for computing the p_i^*'s and c_{ij}'s. Before doing this, we present an example of the reduced-size procedure, namely the two-PSUs-per-stratum example used in Section 2 to illustrate the transportation problem formulation for the optimal procedure.

The ordering of the pairs for this example, as will be shown later, is $\{2,3\}$, $\{1,2\}$, $\{1,3\}$. Consequently, the I_i's, are as given in Table 3. Note that if $I = \{1,2,3\}$ or $I = \{2,3\}$, then the associated set is $I_1 = \{2,3\}$. For the other six possibilities for I the associated set is I itself.

Consequently, from Table 1 we obtain that

$$p_1^* = P(I = \{1,2,3\}) + P(I = \{2,3\}) = .525, \quad (3.4)$$

$p_i^* = P(J_i), i = 2,3$, and $p_i^* = P(J_{i+1}), i = 4, \ldots, 7$, yielding the values in Table 3. Since $\pi_j^* = P(S_j)$, we have $\pi_1^* = .30$, $\pi_2^* = .20$, $\pi_3^* = .50$.

Table 3

Probabilities of Associated Sets: Reduced-Size Procedure

	i						
	1	2	3	4	5	6	7
I_i	{2,3}	{1,2}	{1,3}	{1}	{2}	{3}	∅
p_i^*	.525	.135	.105	.045	.09	.07	.03

The c_{ij} values for this example are given in Table 4. In order to obtain these values, we simplified the computation by letting

$$b_{it} = P(t \in I \mid I^* = I_i),$$

$$i = 1, \ldots, \binom{n}{2} + n + 1, \quad t = 1, \ldots, n, \quad (3.5)$$

and noting that if $S_j = \{s,t\}$ then

$$c_{ij} = b_{is} + b_{it}. \quad (3.6)$$

That is, the expected number of elements in $I \cap S_j$ given $I^* = I_i$ is simply the sum of the probabilities that each of the two elements in S_j was in I given $I^* = I_i$. Also observe that while the transportation problem for the optimal procedure knows the exact value for I and hence knows with certainty whether each element in S_j was in I, this is not the case for the reduced-size procedure, since only the associated set I_i is known. To illustrate, consider the first row of Table 4. Since $I_1 = \{2,3\}$, we know that $2 \in I$ and $3 \in I$, and hence $b_{12} = b_{13} = 1$. However, we do not with certainty whether $1 \in I$ since I_1 is the associated set for both $I = \{1,2,3\}$ and $I = \{2,3\}$. In fact, from Table 1,

$$b_{11} = \frac{P(I = \{1,2,3\})}{P(I = \{1,2,3\}) + P(I = \{2,3\})} = .6.$$

Then $c_{11} = b_{11} + b_{12} = 1.6$, with c_{12}, c_{13} computed similarly. For the remaining six rows in Table 4, $I_i = I$ and hence it is known with certainty which integers were in I. Consequently, the c_{ij}'s for these six rows are easily computed.

Finally, we maximize the expected overlap (3.1) subject to (3.2) and (3.3), obtaining the x_{ij} values in Table 4. The conditional probabilities $P(N = S_j \mid I^* = I_i)$ in Table 5 are then obtained by dividing each of the x_{ij} entries in the i-th row of Table 4 by p_i^*.

Table 4

Values of c_{ij} and Values of x_{ij} that Maximize Overlap for the Reduced-Size Procedure

		c_{ij}			x_{ij}		
		j			j		
i	I_i	1	2	3	1	2	3
1	{2,3}	1.6	1.6	2.0	0.000	0.025	0.500
2	{1,2}	2.0	1.0	1.0	0.135	0.000	0.000
3	{1,3}	1.0	2.0	1.0	0.000	0.105	0.000
4	{1}	1.0	1.0	0.0	0.045	0.000	0.000
5	{2}	1.0	0.0	1.0	0.090	0.000	0.000
6	{3}	0.0	1.0	1.0	0.000	0.070	0.000
7	∅	0.0	0.0	0.0	0.030	0.000	0.000

Table 5

Conditional Probabilities for the Reduced-Size Procedure

		j		
i	I_i	1	2	3
1	{2,3}	0	1/21	20/21
2	{1,2}	1	0	0
3	{1,3}	0	1	0
4	{1}	1	0	0
5	{2}	1	0	0
6	{3}	0	1	0
7	∅	1	0	0

The expected overlap for the reduced-size procedure is .01 less than optimal, that is 1.725 PSUs. The deviation from optimality arises solely because the expected overlap is 1.6 for the joint event that $I^* = \{2,3\}$ and $N = \{1,3\}$. Since the probability of this joint event is .025, and the optimal procedure for this example always produces an overlap of 2 when at least 2 of the PSUs were in the initial sample, the deviation from optimality is $.025(2 - 1.6) = .01$.

The reason that the reduced-size procedure is not able to obtain optimality is that the pair $\{2,3\}$ has a smaller probability of selection in the new sample than in the initial sample. As a result, both the optimal procedure and the reduced-size procedure must sometimes select another pair (always $\{1,3\}$ for both procedures in this example) when $\{2,3\}$ was in the initial sample. The distinction between the two procedures is that the optimal procedure only selects $\{1,3\}$ when $1 \in I$. The reduced-size procedure is unable to use the information about whether $1 \in I$. As a result, when $\{2,3\} \subset I$, $1 \in N$ independently of whether $1 \in I$. This results in a deviation from the optimal overlap.

3.1.2 The Ordering of the Pairs

We now proceed to show in general how the ordering of the pairs is obtained. We use the additional notation here that p_{st}, π_{st}, $s, t = 1, \ldots, n$, $s \neq t$, is the joint probability that $s, t \in I$ and $s, t \in N$, respectively.

The motivation for the ordering of the pairs is as follows. If the i-th pair in the ordering is $\{s,t\}$ then it would be possible for the transportation problem to retain this pair in the new sample when $I^* = I_i$ with conditional probability $\min\{1, \pi_{st}/p_i^*\}$. (The conditional retention probability cannot be any higher than this, since a higher value would result in an unconditional selection probability for the pair in the new design exceeding π_{st}.) Therefore, roughly the goal in the ordering is to make these conditional probabilities as large as possible on average over all pairs.

To illustrate how the ordering of the pairs affects the expected overlap we consider the example of Table 3. Our ordering procedure, as will be shown later, produces the indicated ordering and yields an expected overlap of 1.725 PSUs. Next consider the following alternative ordering for this example. Let the first pair in the ordering be $\{1,3\}$, the second pair be $\{1,2\}$ and the last pair be $\{2,3\}$. With this alternative ordering, $I^* = \{1,3\}$ whenever either $I = \{1,2,3\}$ or $I = \{1,3\}$. Therefore, for this ordering p_1^* is the probability that $I^* = \{1,3\}$, which is now .42. Furthermore, for this alternative ordering, $p_3^* = P(I^* = \{2,3\}) = P(I = \{2,3\}) = .21$, while the other 5 columns in Table 3 remain unchanged. The alternative ordering results in a table of conditional probabilities similar to Table 5, except that in row 1 the I_i, $j = 2$ and $j = 3$ columns now become $\{1,3\}$, 10/21 and 11/21, respectively, and in row 3 the corresponding columns are now $\{2,3\}$, 0 and 1, respectively.

It can be calculated, using the same approach used for the original ordering that the expected overlap for the alternative ordering is 0.055 less than optimal, that is 1.68 PSUs. The reason that this alternative ordering results in a lower expected overlap is as follows. In general a later placement of a pair in the ordering, results in a lower value for the corresponding p_i^*, and hence a higher conditional retention probability when $I^* = I_i$. That is, with $\{1,3\}$ first in the ordering, $\pi_{13}/p_1^* = 10/21$, which is the conditional retention probability for this pair when $I^* = \{1,3\}$; while when $\{1,3\}$ is third in the ordering, $\pi_{13}/p_3^* > 1$ and this pair is retained with certainty. Now the conditional retention probability for the pair $\{2,3\}$ when $I^* = \{2,3\}$ also increases to 1 when $\{2,3\}$ is moved from first to third in the ordering, but the increase is only from 20/21, and hence the original ordering in Table 3 produces a higher expected overlap than the alternative ordering.

Thus, as this example illustrates, the goal of the ordering is to place pairs earlier in the ordering that have a relatively high conditional retention probability even with an early placement. 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 by recursion. 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 by recursion. 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 $k \in \{1, \ldots, n - 1\}$, $\ell \in \{1, \ldots, n - k\}$. 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 illustrate, for the example just considered it will be shown later that $f(1) = 2$, $f(2) = 3$, $f(3) = 1$, $g_1(1) = 3$, $g_1(2) = 1$, $g_2(1) = 1$, and hence the ordering of the pairs is $\{2,3\}$, $\{2,1\}$, $\{3,1\}$. Both the f ordering and the g_k ordering will be constructed to meet the goal stated at the beginning of this paragraph.

To obtain the ordering $f(1), \ldots, f(n)$, recursively define $f(k)$, $k = 1, \ldots, n$, by choosing $f(k) \in T_k$ satisfying

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

where

$$T_1 = \{1, \ldots, n\}, \quad T_k = T_{k-1} \sim \{f(k - 1)\},$$
$$k = 2, \ldots, n, \quad p_i^{(k)} = P(i \in I \text{ and } I \subset T_k),$$
$$k = 1, \ldots, n, \quad i \in T_k. \qquad (3.7)$$

Since $p_i^{(1)} = p_i$, the ordering just defined corresponds to placing first a PSU with the greatest value of π_i/p_i^*. For all k, $p_{f(k)}^{(k)}$ is the probability that $f(k)$ was in I and none of the $k - 1$ elements preceeding $f(k)$ in the f ordering were in I, and hence $p_{f(k)}^{(k)}$ is the probability that an attempt is made to retain $A_{f(k)}$ 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. Generally, the larger $\pi_{f(k)}/p^{(k)}_{f(k)}$ is, the greater the probability that this attempt would be successful. Thus, the motivation for the f ordering of the individual PSUs is the analog of the motivation for the ordering of the pairs of PSUs that we previously discussed.

It remains to explain how to compute $p_i^{(k)}$ for $k \geq 2$. 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

$$p'_\alpha(T) = P(I \cap F_\alpha \subset T), \quad \alpha = 1, \ldots, r, \qquad T \subset \{1, \ldots, n\}, \tag{3.8}$$

$$p''_{i\alpha}(T) = P(i \in I \text{ and } I \cap F_\alpha \subset T), \quad \alpha = 1, \ldots, r, \qquad T \subset \{1, \ldots, n\}, \quad i \in F_\alpha \cap T, \tag{3.9}$$

and observe that

$$p'_\alpha(T) = 1 - \sum_{i \in F_\alpha \sim T} p_i + \sum_{\substack{i,j \in F_\alpha \sim T \\ i<j}} p_{ij}, \tag{3.10}$$

$$p''_{i\alpha}(T) = p_i - \sum_{j \in F_\alpha \sim T} p_{ij}, \tag{3.11}$$

and finally, as established in Ernst and Ikeda (1994),

$$p_i^{(k)} = p''_{i\alpha}(T_k) \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^{r} p'_\ell(T_k), \quad k = 1, \ldots, n, \qquad i \in F_\alpha \cap T_k. \tag{3.12}$$

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

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

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

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

$$p^{(\ell)}_{f(k),j} = P(f(k), \ j \in I \text{ and } I \subset T^*_{k\ell}), \qquad \ell = 1, \ldots, n-k, \quad j \in T_{k\ell}. \tag{3.13}$$

Note that $p^{(\ell)}_{f(k),j}$ is thus the joint probability that $f(k)$ is the first integer in the f ordering in I, that none of the first $\ell - 1$ integers in the g_k ordering are in I, and that $j \in I$. Consequently, $p^{(\ell)}_{f(k),\,g_k(\ell)}$ is the probability that $I^* = \{f(k), g_k(\ell)\}$. Furthermore, if $I_i = \{f(k), g_k(\ell)\}$ then $p_i^* = p^{(\ell)}_{f(k),g_k(\ell)}$, and hence the choice of $g_k(\ell)$ results in the largest value of $\pi_{f(k),g_k(\ell)}/p_i^*$ among the elements in $T_{k\ell}$ in accordance with the previously stated goal for the ordering of the pairs of PSUs.

To compute $p^{(\ell)}_{f(k),j}$, it is established in Ernst and Ikeda (1994) that if $f(k) \in F_\alpha$, $j \in F_\beta$, then

$$\begin{aligned} p^{(\ell)}_{f(k),j} &= p_{f(k),j} \prod_{\substack{t=1 \\ t \neq \alpha}}^{r} p'_t(T^*_{k\ell}) \ \text{ if } \alpha = \beta, \\ &= p''_{f(k),\alpha}(T^*_{k\ell}) p''_{j\beta}(T^*_{k\ell}) \prod_{\substack{t=1 \\ t \neq \alpha,\beta}}^{r} p'_t(T^*_{k\ell}) \ \text{ if } \alpha \neq \beta. \end{aligned} \tag{3.14}$$

We illustrate the computations used in obtaining the ordering for the example that we have been considering. First note that $f(1) = 2$ since the largest value of π_i/p_i occurs for $i = 2$. Next we find $g_1(1)$ which, since $f(1) = 2$, is the $j \in \{1,3\}$ with the maximum value of $\pi_{2j}/p^{(1)}_{2j}$. To find this j, first let $F_\alpha = \{\alpha\}$, $\alpha = 1,2,3$, and note that $T^*_{11} = \{1,2,3\}$. From (3.14) with $\alpha = 2$, $\beta = 1$, it then follows that

$$p^{(1)}_{21} = p''_{22}\{1,2,3\} p''_{11}\{1,2,3\} p'_3\{1,2,3\} = p_2 p_1 \cdot 1 = .45,$$

and similarly it can be obtained that $p^{(1)}_{23} = .525$. Hence $g_1(1) = 3$, since $.5/.525 > .3/.45$. Therefore, the first pair in the ordering is $\{f(1), g_1(1)\} = \{2,3\}$. Then $g_1(2) = 1$, since 1 is the only integer remaining to be used in the g_1 ordering, and consequently the second pair in the ordering is $\{f(1), g_1(2)\} = \{2,1\}$. It is not really necessary to determine $f(2)$, since $\{1,3\}$ is the only remaining pair, and hence the last pair, but to further illustrate the computations, observe that $T_2 = \{1,3\}$, $p_1^{(2)} = p''_{11}\{1,3\} p'_2\{1,3\} p'_3\{1,3\} = p_1(1 - p_2) \cdot 1 = .15$ by (3.12), and similarly $p_3^{(2)} = p_3(1 - p_2) \cdot 1 = .175$. Hence $f(2) = 3$, since $.7/.175 > .5/.15$. Consequently, $g_2(1) = 1$, $f(3) = 1$.

3.1.3 Computation of p_i^* and c_{ij}

Next we explain the computation of the p_i^*'s. If I_i consists of the pair of integers $I_i = \{f(k), g_k(\ell)\}$ then, as previously noted, $p_i^* = p^{(\ell)}_{f(k),g_k(\ell)}$. Consequently, p_i^* can be computed from (3.14) with $j = g_k(\ell)$.

If I_i is a singleton set $\{t\}$ for some $t \in F_\alpha$, then, as established in Ernst and Ikeda (1994),

$$p_i^* = p''_{t\alpha}(\{t\}) \prod_{\substack{u=1 \\ u \neq \alpha}}^{r} p'_u(\emptyset). \tag{3.15}$$

Finally, if $I_i = \emptyset$, then

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

It remains only to explain how to compute the c_{ij}'s which, by (3.5) and (3.6), reduces to computing b_{it}, $i = 1, \ldots, \binom{n}{2} + n + 1$, $t = 1, \ldots, n$.

To compute b_{it}, observe that

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

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

$$b_{it} = 1 \quad \text{if} \quad t = f(k) \quad \text{or} \quad t = g_k(\ell), \tag{3.16}$$

$$= 0 \quad \text{if} \quad t \notin T^*_{k\ell}, \tag{3.17}$$

$$= 0 \quad \text{if} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{and} \quad \gamma = \alpha = \beta, \tag{3.18}$$

$$= \frac{p_{f(k),t}}{p''_{f(k),\alpha}(T^*_{k\ell})} \quad \text{if} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{and} \quad \gamma = \alpha \neq \beta, \tag{3.19}$$

$$= \frac{p_{g_k(\ell),t}}{p''_{g_k(\ell),\beta}(T^*_{k\ell})} \quad \text{if} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{and} \quad \gamma = \beta \neq \alpha, \tag{3.20}$$

$$= \frac{p''_{t\gamma}(T^*_{k\ell})}{p'_\gamma(T^*_{k\ell})} \quad \text{if} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{and} \quad \gamma \neq \alpha, \gamma \neq \beta. \tag{3.21}$$

In Ernst and Ikeda (1994) it is demonstrated how (3.16)-(3.21) were obtained.

In the actual implementation for the SIPP application, modifications of the reduced-size procedure were needed to overlap the 1990s SIPP design with the 1980s SIPP design. The modifications were necessary because the PSU definitions in the 1980s and 1990s designs were not identical. As a result, some PSUs in the 1990s design could intersect more than one 1980s design PSU. These modifications are detailed in Ernst and Ikeda (1994).

3.2 Modifications of Reduced-Sized Procedure for Other Designs

In general, consider any m'-PSUs-per-stratum without replacement initial design and any m-PSUs-per-stratum without replacement final design, where m', m are any positive integers. Although the reduced-size procedure in Section 3.1 was only presented for the case $m = m' = 2$, it is actually applicable for any m, m'. We will sketch the modifications necessary when $m \neq 2$ or $m' \neq 2$.

A different value of m' only requires modification of some of the computations. For example, if $m = 2$, but $m' \neq 2$, 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, regardless of the value of m', the set of all distinct triples, instead of pairs, of integers in $\{1, \ldots, n\}$, 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 contained in I. Otherwise, the new selection probabilities are conditioned on I itself. Thus the new selection probabilities are 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 in a manner similar to the construction of $g_k(1), \ldots, g_k(n - k)$. For example, in defining $h_{k\ell}(v)$ for $v \geq 2$, $p_{f(k),j}^{(\ell)}$ in the definition of $g_k(\ell)$ is replaced by

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

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 $m \geq 4$, ordered m-tuples would be defined in a similar manner and the new selection probabilities conditioned on $\binom{n}{m} + \binom{n}{m-1} \cdots + n + 1$ 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 $m > m'$, 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-tuple 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 can 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 can be conditioned on subsets of I of size m'' or less, where $m'' < m$, although with a generally smaller expected overlap.

3.3 Relationship Between Expected Overlap for the Reduced-Size Procedure, the Optimal Procedure and Independent Selection

Let Ω_I, Ω_R, Ω_O denote the expected overlap for the independent selection, the reduced-size procedure, and the optimal procedure, respectively. In Ernst and Ikeda (1994) the relationship between these quantities is explored. We briefly summarize here some of the results.

It is established that $\Omega_I \leq \Omega_R \leq \Omega_O$ for any m, m' where m, m' are as in Section 3.2. In addition, for the case that we have been focusing on, $m = m' = 2$, lower bounds are established on Ω_R and upper bounds are established on Ω_O and $\Omega_O - \Omega_R$.

For example, let μ_2 denote the probability that there are at least two elements in I, μ_1 denote the probability that I is a singleton set, and

$$\lambda = \min\{\min\{\pi_i/p_i : i = 1, \ldots, n\},\ \min\{\pi_{ij}/p_{ij} : i,j = 1, \ldots, n, i \neq j\}, 1\}.$$

Then $\Omega_O \leq 2\mu_2 + \mu_1$, $\Omega_R \geq \lambda(2\mu_2 + \mu_1/2)$, and $\Omega_O - \Omega_R \leq 2(1 - \lambda)\mu_2 + (1 - \lambda/2)\mu_1$.

Unfortunately these bounds are not always very tight. However, in certain circumstances they are useful. For example, if $\pi_{ij} \geq p_{ij}$ for all i,j and the probability is 1 that there is at least two elements in I, then it follows from these bounds that $\Omega_R = \Omega_O = 2$.

Finally, an example is presented to illustrate a worst case situation for Ω_R in relation to Ω_O for the case m, $m' = 2$. It shows that Ω_O may equal 2, while Ω_R is arbitrarily close to 0. Thus, at least in theory, the reduced-size procedure can be ineffective. However, in practice, as will be shown in the next section, Ω_R is much closer to Ω_O than to Ω_I, at least for the SIPP application.

4. APPLICATION OF REDUCED-SIZE PROCEDURE TO SIPP

Results from simulations of the SIPP overlap, done prior to production for research and testing purposes, are presented, as well as results from the actual SIPP production overlap. Further details are given in Ernst and Ikeda (1992b, 1994).

In the implementation of the reduced-size overlap procedure, minimum cost flow (MCF) optimization software, written by Darwin Kingman and John Mote at the University of Texas at Austin, was used to solve the required transportation problem. A FORTRAN program was written to produce input to and process output from the MCF software.

To test the software prior to production, the program was used to overlap two stratifications, based on 1970 census data, of the SIPP Midwest region with the actual 1980s design stratification for the SIPP Midwest region. (At the time of this test, 1990 census data was not yet available.) The 1970-based stratifications were produced by stratifying the 1980s SIPP noncertainty PSUs in the Midwest region using 1970 data. Both of the 1970-based stratifications partitioned the noncertainty PSUs into 31 strata, using different sets of stratification variables. The stratifications based on 1980 and 1970 data were treated as "initial" and "final" stratifications for the purposes of the overlap algorithm.

In the actual implementation, as noted in Section 3.1 and detailed in Ernst and Ikeda (1994), a modification of the reduced-size procedure was used to overlap the 1990s SIPP design with the 1980s SIPP design, because the PSU definitions in the 1980s and 1990s designs were not identical. The modified reduced-size procedure was used to overlap 103 final (1990s design) nonselfrepresenting strata in SIPP.

The expected overlap was calculated for the reduced-size maximum overlap algorithm, for independent selection of final PSUs, and for an upper bound to the expected overlap for the optimal procedure. An upper bound was calculated instead of the actual optimal overlap, since the optimal overlap cannot be calculated for the larger strata. For the simulation, the upper bound used is the one stated in Section 3.3, $\mu_2 + 2\mu_1$, while for the production SIPP, a different upper bound, described in Ernst and Ikeda (1994), was required because the PSU definitions in the 1980s and 1990s were not identical.

The results from the two final stratifications in the simulation were generally similar to each other. Combining the results from both stratifications, the mean expected overlap for this set of 62 strata was 1.552, 1.569 and 0.480 PSUs/stratum for the reduced-size procedure, the upper bound to the optimal overlap and independent selection respectively. For the actual SIPP implementation, the corresponding number was 1.523, 1.647 and 0.582, respectively, while the corresponding expected number of PSUs overlapped for the 103 strata was 156.9, 169.6 and 59.9, respectively. Thus, in both the simulations and the production SIPP, the reduced-size procedure yielded results reasonably close to the upper bound for the optimal procedure.

The reduced-size algorithm took a fairly short time to run on most strata. The CPU times in the simulation for final strata with different numbers of PSUs are given below. The reduced-size program was run on a Solbourne 5/605 computer. The median number of PSUs in a stratum, for the entire group of 62 strata, was 17 PSUs. The 68 PSUs stratum was the largest stratum.

Table 6

CPU Times for Reduced-Size Procedure

Number of PSUs	CPU Time (hrs:min:sec)
18	0:36
37	5:44
49	24:05
68	2:23:43

We also calculated for the actual SIPP implementation, that of the 103 final strata overlapped by the modified reduced-size procedure, 41 would not have run under the optimal procedure. This calculation was based on our estimate that the maximum size transportation problem, in terms of number of variables, that could have run in production was 4 × 10^6. The number of variables for the optimal procedure was less than 4 × 10^6 for all 56 strata for which $n \leq 14$, but exceeded this limit for all but 6 of the 47 strata with $n \geq 15$, including two with $n = 15$. The maximal size of the transportation for the optimal procedure among the 103 strata occurred for a stratum with $n = 46$, for which there were 3.61 × 10^{12} variables. In contrast, there were 1.03 × 10^6 variables for the modified reduced-size procedure for this stratum.

Another question of interest is the overlap effectiveness of the reduced-size procedure in comparison with the overlap procedure of Ernst (1986). In general it is believed that the reduced-size procedure should produce a higher overlap in situations when both are usable, since the reduced-size procedure makes use of the stratum-to-stratum independence in the initial design. However, although the procedure in Ernst (1986) is applicable to two-PSU-per-stratum designs, no computer program has ever been written at the Census Bureau (or anywhere else that the authors are aware of) to implement this procedure for such designs, since there has not yet been a production application for this program. Consequently, we cannot make a direct comparison of these two methods on the same data. However, a crude comparison can be made from the results of the reduced-size overlap procedure for SIPP data and the results of the overlap using the procedure in Ernst (1986) for the overlap of 1990s CPS and NCVS designs with their respective 1980s designs. (Both the 1980s and 1990s designs for CPS and NCVS are one-PSU-per-stratum designs.)

For CPS, the overlap procedure resulted in an average increase in expected overlap, in comparison with independent selection, of .26 PSUs/stratum, and for NCVS the overlap procedure resulted in an average increase in expected overlap of .30 PSUs/stratum. This compares with an increase of .94 PSUs/stratum for the reduced-size procedure over independent selection for SIPP. If the two overlap procedures are equally effective, then one might expect that the increase in overlap per stratum for SIPP would be roughly twice as large as for CPS and NCVS, since SIPP has a two-PSUs-per-stratum design. By this standard, the reduced-size procedure program performs better than the procedure in Ernst (1986). However, since the stratifications were quite different for these three surveys, the validity of this comparison is open to question.

For the example considered in Sections 2 and 3, a valid comparison of the different overlap procedures can be made, since the expected overlap values for the procedure in Ernst (1986)), 1.625, was easily calculated by hand. For the reduced-size procedure the corresponding overlap value is 1.725, and for the optimal procedure it is 1.735.

CONCLUSIONS

The reduced-size overlap procedure presented in this paper meets its two key objectives in practice. It reduces the size of the transportation problems to a usable size, as evidenced both by the size of the transportation problem in the formulation (3.1)-(3.3), and the fact that it has actually been implemented in the redesign of a major survey. In addition, the procedure accomplishes the size reduction while yielding nearly optimal overlap, at least for the SIPP application. It can only be used when the PSUs in the initial design are selected independently from stratum to stratum, but when this condition is met we believe it is the overlap procedure of choice for large strata.

ACKNOWLEDGEMENTS

The programming assistance of Todd Williams is gratefully acknowledged. The authors would also like to thank the referees and the editors for their constructive comments. The views expressed in this paper are attributable to the authors and do not necessarily reflect those of the Bureau of Labor Statistics and the Census Bureau.

REFERENCES

ARAGON, J., and PATHAK, P.K. (1990). An algorithm for optimal integration of two surveys. *Sankhyā: The Indian Journal of Statistics*, 52, 198-203.

ARTHANARI, T.S., and DODGE, Y. (1981). *Mathematical Programming in Statistics*. New York: John Wiley and Sons.

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.

ERNST, L.R. (1986). Maximizing the overlap between surveys when information is incomplete. *European Journal of Operational Research*, 27, 192-200.

ERNST, L.R. (1989). Further Applications of Linear Programming to Sampling Problems. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-89/05.

ERNST, L.R., and IKEDA, M. (1992a). Modification of the Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units Are Redefined in the New Design. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-91/01.

ERNST, L.R., and IKEDA, M. (1992b). Summary of the Performance of the Maximum Overlap Algorithms for the 1990's Redesign of the Demographic Surveys. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-92/01.

ERNST, L.R., and IKEDA, M. (1994). A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-93/02.

GLOVER, F., KARNEY, D., KLINGMAN, D., and NAPIER, A. (1974). A computation study on start procedures, basic change criteria and solution algorithms for transportation problems. *Management Sciences*, 20, 793-813.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in probabilities. *Journal of the American Statistical Association*, 46, 105-109.

KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

PATHAK, P.K., and FAHIMI, M. (1992). Optimal integration of surveys. In *Essays in Honor of D. Basu*. Eds. M. Ghosh, and P.K. Pathak. Hayward, California: Institute of Mathematical Statistics, 208-224.

PERKINS, W.M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Stata. Memorandum to Joseph Waksberg, Bureau of the Census.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw Hill.