OPTIMAL INTEGRATION OF SURVEYS

P. K. Pathak, Department of Mathematics and
Statistics, University of New Mexico, Albuquerque

and

M. Fahimi, Department of Mathematics and
Statistics, University of New Mexico, Albuquerque

The problem of integration of surveys is known to be of considerable practical as well as theoretical interest in the design of multi-purpose and continuing surveys. The object of this paper is to present a brief review of current developments in this area and to furnish a unified framework within which integration of surveys can be studied from various angles.

Introduction

The problem of integration of surveys, i.e., the problem of designing a sampling program for two or more surveys which maximizes the overlap between observed samples is known to be of considerable practical as well as theoretical interest in the design of multi-purpose and continuing surveys (Keyfitz, 1951). Development of cost-efficient sampling programs of this kind is a problem which agencies such as the National Sample Survey of India, Statistics Canada, the U.S. Bureau of the Census, the U.S.D.A., and others worldwide, have been continuing to tackle on an ad hoc basis. And although the literature on it is now over 40 years old, basic research in it has reached a modest level of maturity only recent (cf. Arthanari and Dodge, 1981; Causey et al. 1985; Krishnamoorthy and Mitra, 1987; Maczynski and Pathak, 1980; and others). Nevertheless despite these recent gains, there remains a pressing need for a unified framework within which integration of surveys and other similar problems of this nature, such as controlled selection and controlled rounding (Goodman and Kish, 1950, and Causey et al., 1985), can be studied with all their ramifications. In due course, such an approach is bound to provide a powerful guide to the cost-efficient design of survey programs commonly encountered in practice. The primary object of this paper is to briefly review the contemporary work in this area from the theoretical as well as computational viewpoints.

In broad terms, integration of surveys can be referred to as the sampling program for two or more surveys. It has its origin in multipurpose surveys and sampling over successive occasions. In multipurpose surveys, the population characteristics under study are often subdivided into two or more groups of

This research has been supported by the National Science Foundation Grants DMS-8703798 and DMS-8844575.

positively correlated characteristics and different sampling schemes are employed for data collection for the different groups of characteristics. For example in multipurpose surveys, traditionally population size is made the basis for socio-economic surveys and geographical area for agricultural surveys. This gives rise to the multivariate problem of designing an overall sampling program which imbeds the different sampling schemes into a single multi-variate sampling scheme in a cost-effective manner. A problem of similar kind arises in sampling over successive occasions (Keyfitz, 1951) in which a given population is sampled by probabilities proportional to size over two or more successive time periods. Over time, sizes of population units change and this necessitates sampling the given population according to a new set of probabilities at each new time period in a cost-effective manner. In this case, it makes sense to design an overall joint sampling program which in some sense maximizes the overlap between samples over different occasions, or equivalently minimizes the number of distinct population units sampled over different occasions. In applications of this nature, population units to be sampled are typically primary sampling units (psu's) and their selection represents a considerable financial investment. A new independent selection amounts to selecting an almost new set of psu's each time and is not cost-effective. On the other hand, the use of the same psu's on succeeding occasions, much like in the paired t-test, leads to significant reductions in errors of comparisons between periodic surveys (Kish and Scott, 1971). Thus in applications of this nature, an integrated joint sampling program which minimizes the number of distinct psu's selected over successive time periods is cost-effective and highly desirable.

The problem of integration of surveys for surveys involving two with replacement sampling schemes was originally formulated and solved by Keyfitz (1951). Lahiri (1954) proposed a serpentine arrangement of geographically contiguous psu's for optimal integration of two surveys. Raj (1957) studied the problem of integration of two surveys as a transportation problem and established the optimality of Lahiri's algorithm under a one-dimensional metric. Felligi (1966) studied the problem of integration of two without replacement sampling schemes and noted that even the simplest case of the sample size $n = 2$ causes added complications. In the context of k (≥ 2) with replacement sampling schemes, Maczynski and Pathak (1980) presented a general solution in a closed form under certain assumptions. More recently Krishnamoorthy and Mitra (1986), and Mitra and Pathak (1984) have presented sequential algorithms for 'optimal' integration of two or three surveys in the context of with replacement sampling (cf. Arthanari and Dodge, 1981; Keyfitz, 1951; Lahiri, 1954; Raj, 1957; Kish and Scott, 1971; and others). The recent upsurge of research in this area is both practically useful and theoretically interesting.

Although the connection between integration of surveys and the transportation problem is well-known (Aragon and Pathak, 1990; Arthanari and Dodge, 1981; Causey et al., 1985; Maczynski and Pathak, 1980, p. 137; and Raj, 1957), it is only in recent years that serious attempts have been made to solve the problem of optimal integration of surveys as a transportation problem. In general, integration of surveys is a transportation problem with an exponentially large number of variables, e.g. a simple problem of integration of two samples of size $n = 5$ each from a population of size $N = 50$ is equivalent to a transportation problem with approximately 4.5×10^{12} variables. At the present time solvers of the transportation problems of this size are unavailable in the public domain. An despite the unique sparse structure of the underlying tableaus of integration of surveys, there are very few results in the literature on size reduction techniques as an alternative to solving these very large transportation problems. Much remains to be done in the area of size reduction techniques for integration of surveys.

Formulation of the Problem

For clarity in the exposition, we adopt the following terminology:

Z : the set of the first N natural numbers. We use the artifice of identifying the population under study by Z.

k : number of surveys to be carried out, $k \geq 2$.

S : collection of all possible samples from which a sample is to be drawn for each of the k surveys; S being a subset of the power set of Z.

n_i : sample size for the i^{th} survey, $1 \leq i \leq k$.

x_i : the outcome of the i^{th} survey.

X : the joint outcome of k surveys, i.e., $X = (x_1,\ldots,x_k)$.

P_{ij} : the probability of selecting the j^{th} sample s_j on the i^{th} survey, i.e., $P_{ij} = P(x_i = s_j)$, $x_i \in X$, $s_j \in S$, $1 \leq i \leq k$.

d : a cost function defined on S^k, i.e., it is non-negative and sub-additive on S^k.

A survey or a sampling scheme on Z is a given, but otherwise quite arbitrary, collection S of samples from Z endowed with a given probability distribution P. Thus a survey is expressed by the pair $(S, P) = \{(s, P(s)) \colon s \in S\}$ in which $P(s)$ denotes the probability of selection of the sample s.

The problem of optimal integration of k surveys can now be stated as follows:

Given k individual surveys, (S, P_i), $1 \leq i \leq k$, and a cost function d on S^k, find a joint probability distribution $\mathbb{P}$ for X on S^k, which for each x_i realizes the preassigned marginal probabilities P_i determined by the i^{th} survey, i.e.,

$\mathbb{P}(x_i = s_j) = P_{ij}$, and at the same time minimizes the expectation of the cost function over the class of all surveys of this kind.

Raj (1957) was perhaps the first to paraphrase the problem of integration of surveys as a transportation problem. In terms of our terminology, it is as follows:

Problem 1

Given k surveys $\{(S, P_i): 1 \leq i \leq k\}$ and a cost function d on S^k,

$$\text{minimize } \phi(X) = \sum_X P(X) \cdot d(X)$$

$$\text{subject to } \sum_{X_{ij}} P(X) = P_{ij},\ X_{ij} = \{X \in S^k: x_i = s_j\},$$

$$P(X) \geq 0, \forall X \in S^k.$$

Example 1

Consider a population of four psu's and suppose that on two occasions a sample of size two (*wor*) is to be drawn from the population according to the sampling schemes given in Table 1.

Table 1: Sampling Schemes for Example 1

Sample	1st Survey	2nd Survey
s_j	P_{1j}	P_{2j}
(1, 2)	0.04	0.22
(1, 3)	0.16	0.15
(1, 4)	0.21	0.29
(2, 3)	0.13	0.07
(2, 4)	0.32	0.10
(3, 4)	0.14	0.17

To maximize the expectation of the overlap between the two samples selected on the two occasions, the cost function d is taken to be the number of distinct psu's in the two samples, i.e., $d(s_m \cup s_n) = \#(s_m \cup s_n)$. The transportation problem representation of this problem is as follows:

$$\text{minimize} \quad \sum_{m=1}^{6} \sum_{n=1}^{6} P(x_1 = s_m, x_2 = s_n) d(s_m, s_n)$$

$$\text{subject to} \quad \sum_{n=1}^{6} P(x_1 = s_m, x_2 = s_n) = P_{1m},$$

$$\sum_{m=1}^{6} P(x_1 = s_m, x_2 = s_n) = P_{2n},$$

$$P(x_1 = s_m, x_2 = s_n) \geq 0,$$

$$\forall\ m,\ n = 1,\ldots,6.$$

To solve this problem, one can use a standard linear programming package, e.g., the use of the SPLO-program (1981) yields the results summarized in Table 2.

Table 2: A Solution to Example 1
$P(x_1 = s_m, x_2 = s_n),\ m,\ n = 1,\ldots,6$

Survey II/I

	1, 2	1, 3	1, 4	2, 3	2, 4	3, 4	
1, 2	0.04						0.04
1, 3	0.01	0.15					0.16
1, 4			0.21				0.21
2, 3	0.06			0.07			0.13
2, 4	0.11		0.08		0.10	0.03	0.32
3, 4						0.14	0.14
	0.22	0.15	0.29	0.07	0.10	0.17	1.00

Based on this solution, we find that the minimum value of the objective function as defined in Problem 1 is 2.29. It is worth noting that this represents the maximum expected overlap between the two sampling schemes.

Integration of Surveys with Replacement Sampling Schemes

Unfortunately, most realistic problems of integration of surveys are not as tractable as the example in the preceding section seems to indicate. For example, it is easily seen that a straight forward 3-dimensional integration of surveys problem for a population of 20 psu's for without replacement samples of size $n = 3$ amounts to a transportation problem with over a billion variables. At the present time, hardware or software which can handle a problem of this magnitude is unavailable in the public domain. So it is not surprising at all that earlier attempts at solutions of these problems were largely directed towards finding closed form solutions under special circumstances. The first general result of this nature was obtained by Maczynski and Pathak (1980). Based on the following lemma, they provided closed form solutions for the special case of the sample size $n_i = 1$ and k surveys, $k \geq 2$.

Lemma 1

Consider the general problem of integration of k surveys with $n_i = 1$ and suppose that there exists a joint probability distribution $\mathbb{P}$ on S^k such that for each h, $1 \leq h \leq k$, and $1 \leq i_1 < \ldots < i_h \leq k$,

$$\mathbb{P}(x_{i_1} = x_{i_2} = \ldots = x_{i_h} = j) = min(P_{i_1 j}, \ldots, P_{i_h j}).$$

Then

$$\mathbb{P}(\bigcup_i \{x_i = j\}) = max_i P_{ij}.$$

Moreover such a $\mathbb{P}$ minimizes the expected number of distinct psu's selected in all the k samples.

The case $k = 2$. An immediate corollary of Lemma 1 is that for $k = 2$ surveys with $n_i = 1$, the following closed-form solution is optimal:

$$\mathbb{P}(x_1 = h, x_2 = j) = f(h), \qquad h = j$$

$$\mathbb{P}(x_1 = h, x_2 = j) = f_{12}(h, j), \qquad h < > j$$

where

$$f(h) = min(P_{1h}, P_{2h}),$$

$$f_{12}(h, j) = (P_{1h} - f(h))(P_{2j} - f(h))(1 - \Sigma_r f(r))^{-1}.$$

Algorithmic representations of the above solution have been provided by Keyfitz (1951) and Mitra and Pathak (1984).

The case $k = 3$. The problem of optimal integration of three surveys with $n_i = 1$ is essentially solved now. In this case Lemma 1 forms the basis of all closed-form solutions which minimize the expected number of distinct sample units over the three occasions. In a series of fundamental papers Krishnamoorthy and Mitra (1986, 1987) have established the optimality of the Mitra-Pathak type algorithmic approach (1984) for the integration of three with replacement sampling schemes. For further details in this connection we refer the reader to the elegant work of Krishnamoorthy and Mitra (1987). For completeness, it would be worthwhile indeed to investigate extensions of these results to the general case of k surveys with $k > 3$.

Size Reduction Techniques

In this section we present a brief review of a technique which has the potential of significantly reducing the size of the induced transportation problem in the context of integration of two surveys. To illustrate this technique, consider the induced transportation problem of Example 1 and observe that the cost function d of this example is in fact a metric. This simple observation allows one to establish the following interesting result (Aragon and Pathak, 1990):

Theorem 1

Consider the problem of integration of two surveys of equal size in which the cost function d is induced by a metric. Then there is an optimal feasible solution $\mathbb{P}$ such that for each sample s_m

$$\mathbb{P}(x_1 = s_m, x_2 = s_m) = min(P_{1m}, P_{2m}).$$

The theorem implies that by setting these diagonal probabilities equal to their largest admissible values, namely the minimum of the corresponding row and column marginal probabilities, at least half of the restrictions of the problem are satisfied. What is left then is the optimal determination of the remaining unknown nondiagonal probabilities, i.e. $\mathbb{P}(x_1 = s_m, x_2 = s_n)$, for only some of $m <> n$. This reduced problem is at most one-fourth the size of the original problem. For example, application of this technique to Example 1 reduces the original problem with 36 variables to a smaller problem with only 9 variables. The reduced problem is stated below and an optimal solution summarized in Table 3.

$$\text{minimize} \qquad \sum_{m=1}^{3}\sum_{n=1}^{3} P(x_1 = s_m, x_2 = s_n) \cdot d(s_m, s_n)$$

$$\text{subject to} \qquad \sum_{n=1}^{3} P(x_1 = s_m, x_2 = s_n) = P_{1m},$$

$$\sum_{m=1}^{3} P(x_1 = s_m, x_2 = s_n) = P_{2n},$$

$$P(x_1 = s_m, x_2 = s_n) \geq 0,$$

$$\forall\ m, n = 1,\ldots,3.$$

Table 3: A Solution to the Reduced Version of Example 1
$P(x_1 = s_m, x_2 = s_n)$, m, $n = 1,\ldots,3$

Survey II

		1, 2	1, 4	3, 4	
	1, 3	0.01			0.01
Survey I	2, 3	0.06			0.06
	2, 4	0.11	0.08	0.03	0.22
		0.18	0.08	0.03	0.29

Note that the marginal probabilities P_{1m} and P_{2n} are given by the marginal entries in Table 3. And that Table 3 was obtained from Table 1 after the assignment of the main diagonal probabilities $\mathbb{P}(s, s)$. Thus at the first stage of this size reduction technique, we set $\mathbb{P}(s_1, s_1) = .04$, $\mathbb{P}(s_1, s_2) = \ldots = \mathbb{P}(s_1, s_6) = 0$, $\mathbb{P}(s_2, s_1) = 0$, $\mathbb{P}(s_2, s_2) = .15,\ldots,\mathbb{P}(s_6, s_6) = .14$. Then at the second stage, the reduced problem is solved by using a standard transportation problem solver. The two solutions when combined together furnish a complete solution to the original problem of Example 1.

In order to present this size reduction technique in greater generality and scope, a slight digression from the main theme of the paper is necessary. We turn now to the following so-called Hitchock transportation problem (Chvatal, 1983, p. 345):

Problem 2

$$
\begin{array}{lll}
\text{minimize} & \sum_i \sum_j c_{ij} x_{ij} & \\
\text{subject to} & \sum_j x_{ij} = p_i, & (i = 1,\ldots,m), \\
 & \sum_i x_{ij} = q_j, & (j = 1,\ldots,n), \\
 & x_{ij} \geq 0, & \forall\ i, j.
\end{array}
$$

Moreover, suppose that in the preceding transportation problem, there are cells in the cost matrix $C = \{c_{ij}\}$ with the following property of negative variation:

Definition 1

A cell (i, j) of the cost matrix $C = \{c_{ij}\}$ is said to have *negative variation* if for all $k < > i$ and $l < > j$, the following inequality holds:

$$(c_{ij} + c_{lk}) - (c_{il} + c_{kl}) \leq 0.$$

Similarly, the cell (i, j) is said to have *positive variation* if the above difference is always non-negative.

If the cost matrix of a given transportation problem has cells with negative variation, then the given problem can be reduced to a new problem of a smaller size. Specifically, if the original problem is of size $m \times n$ and has c cells with negative variation, then the reduced problem is of size $(m-a) \times (n-b)$ with $a + b \geq c$. This size reduction is a consequence of the following theorem:

Theorem 2

Suppose that a given cell, say (1, 1), of the cost matrix of Problem 2 has negative variation. Then there exists an optimal feasible solution $X = \{x_{ij}\}$ with $x_{11} = min(p_1, q_1)$.

In a different guise, this theorem can be found hidden in the seminal work of the noted French mathematician Monsieur Monge (1781). In a totally different context in operations research, it has been used by A.J. Hoffman (1963). We independently discovered it in the context of integration of surveys and controlled selection. We take the liberty of referring to this theorem as the Monge-Hoffman size reduction theorem.

Corollary 1

If the objective of Problem 2 is maximization instead of minimization, then the above theorem goes through provided the cell (1, 1) has positive variation.

Now consider the Transportation Problem 2 and assume that the cell (1,1) of its cost matrix has negative variation. Also, without loss of generality assume that $p_1 < q_1$. Then the preceding theorem implies that the original problem can be replaced by the following smaller problem:

minimize $$\sum_i \sum_j c_{ij}x_{ij},\ 2 \leq i \leq m,\ 1 \leq j \leq n$$

subject to $$\sum_j x_{ij} = p_i,\ (i = 2,\ldots,m),$$

$$\sum_i x_{ij} = q_j,\ (j = 1,\ldots,n),$$

$$x_{ij} \geq 0,\ \forall\ i, j.$$

Clearly any optimal solution of this problem, along with $x_{11} = p_1$, $x_{12} = \ldots = x_{1n} = 0$, will provide an optimal solution to the original problem. This effectively reduces the number of variables from $m \times n$ to $(m-1) \times n$. If instead of $p_1 < q_1$, we have $q_1 < p_1$, a similar consideration will show that an optimal solution now is given by $x_{11} = q_{11}$, $x_{21} = \ldots = x_{m1} = 0$ and the values of the remaining variables are obtained by solving an analogous reduced problem involving $m \times (n-1)$ variables. Finally if $p_1 = q_1$, then an optimal solution is given by $x_{11} = p_1 = q_1$, $x_{12} = \ldots = x_{1n} = x_{21} = \ldots = x_{m1} = 0$, and the values of the remaining variables are obtained by solving another analogous reduced problem of $(m-1) \times (n-1)$ variables.

Note that if the cost matrix in the transportation problem has multiple cells with negative variation then the preceding size reduction algorithm can be carried out sequentially until all the cells with negative variation have been removed. In the special case of $c_{ij} = |j - i|$, this size reduction procedure can be carried out to the very end and an optimal solution can be obtained without ever having to invoke any solvers of the transportation problem. This last result is a consequence of the following theorem.

Theorem 3

Consider the $m \times n$ matrix $D = \{d_{ij}\}$ in which $d_{ij} = |j - i|$, and suppose that the first r rows and the first c columns of D have been removed. Let $s = max(r, c)$ and $t = min(m, n)$. Then the following holds:

a) All the cells on the shortest path joining the cells $(1, 1)$, $(s-r+1, s-c+1)$, $(t-r, t-c)$ and $(m-r, n-c)$ have negative variation.

b) All the cells (i, j) with $i \geq t - r$, $j \leq s - c + 1$, and all cells (k, l) with $k \leq s - r + 1$, $l \geq t - c$ have positive variation.

Corollary 1

The above theorem also holds if some of the very last rows and columns of the matrix are removed as well. (This should be self-evident since the removal of rows and columns from the end leaves original structure of the matrix D intact.)

Corollary 2

Suppose that $d(x, y)$ has the following property of a distribution function in two dimensions:

$$d(x+h, y+k) - d(x, y+k) - d(x+h, y) + d(x, y) \geq 0.$$

for all $h, k \Rightarrow 0$. Then the northeast and the southwest cells of the matrix D have positive variation, while the northwest and the southeast cells have negative variation.

An immediate consequence of the above corollary is that for distance functions such as $d(x, y) = xy$, the conventional northwest (greedy) algorithm provides an optimum solution for the Hitchock Transportation Problem 2.

Example 2

The purpose of this example is to graphically illustrate the statement of Theorem 3. Consider a 9 × 7 matrix D for Theorem 3. Tables 4 through 6 summarize the variations of D when certain initial rows and columns are removed from it.

Table 4: Variations of the Matrix D
(0 rows and 0 columns are removed)

	1	2	3	4	5	6	7
1	—						+
2		—					
3			—				
4				—			
5					—		
6						—	
7	+						—
8	+						—
9	+						—

Table 5: Variations of the Matrix D
(3 rows and 2 columns are removed)

	3	4	5	6	7
4	—	—			+
5			—		
6				—	
7	+	+			—
8	+	+			—
9	+	+			—

Table 6: Variations of the Matrix D
(2 rows and 3 columns are removed)

	4	5	6	7
3	—			+
4	—			+
5		—		
6			—	
7	+			—
8	+			—
9	+			—

When the underlying cost coefficients c_{ij}'s of Problem 2 are given by the matrix D, the northeast algorithm (Algorithm I) provides a complete solution to the Problem 2. To determine the variations of the cells of an arbitrary matrix, the algorithms similar to the positive variation algorithm (Algorithm II) can be used. Both of these algorithms are given at the end of this paper.

Example 3

This example is taken from the paper by Causey, Cox, and Ernst (1985) on the problem of maximizing the overlap between two surveys. The sampling scheme is summarized in Table 7. The cost matrix $C = \{c_{ij}\}$, where $c_{ij} = \#(s_i \cap s_j)$, along with its variations are summarized in Table 8. The object is to:

$$\text{maximize} \qquad \textstyle\sum_i \sum_j c_{ij} p_{ij},\ 1 \le i \le 12,\ 1 \le j \le 5$$

$$\text{subject to} \qquad \textstyle\sum_j p_{ij} = p_i,\ 1 \le i \le 12,$$

$$\textstyle\sum_i p_{ij} = q_j,\ 1 \le j \le 5,$$

$$p_{ij} \ge 0,\ \forall\ i, j.$$

It follows from Corollary 1 of Theorem 2 that there exists an optimal solution for this problem such that for $1 \le i, j \le 5$, $P(s_i, s_j) = min(p_{1i}, p_{2i})$ for $i = j$ and zero otherwise. This partial solution reduces the size of the original problem from $12 \times 5 = 60$ to $7 \times 5 = 35$ variables as summarized in Table 9. The reduced problem can now be solved using any standard solver of transportation problems.

Table 7: Sampling Scheme for Example 3

Survey I		Survey II	
s_m	P_{1m}	s_n	P_{1n}
(1)	.15	(1)	.40
(2)	.018	(2)	.15
(3)	.012	(3)	.05
(4)	.24	(4)	.30
(5)	.04	(5)	.10
(1,4)	.30		
(1,5)	.05		
(2,4)	.036		
(2,5)	.006		
(3,4)	.024		
(3,5)	.004		
(0)	.12		

Table 8: The Cost Matrix and Its Variation

	1	2	3	4	5
1	1+	0	0	0	0
2	0	1+	0	0	0
3	0	0	1+	0	0
4	0	0	0	1+	0
5	0	0	0	0	1+
6	1	0	0	0	1
7	1	0	0	1	0
8	0	1	0	0	1
9	0	1	0	0	1
10	0	0	1	1	0
11	0	0	1	0	1
12	0	0	0	0	0

Table 9: Sampling Scheme for the Reduced Problem

Survey I		Survey II	
s_m	P_{1m}	s_n	P_{1n}
(1, 4)	.30	(1)	.25
(1, 5)	.05	(2)	.132
(2, 4)	.036	(3)	.038
(2, 5)	.006	(4)	.06
(3, 4)	.024	(5)	.06
(3, 5)	.004		
(0)	.12		

Example 3

This example establishes the optimality of Lahiri's selection scheme (1954). This scheme requires a serpentine ordering of the psu's as illustrated in Figure 1 below so that geographically contiguous units occur next to each other in the sampling frame.

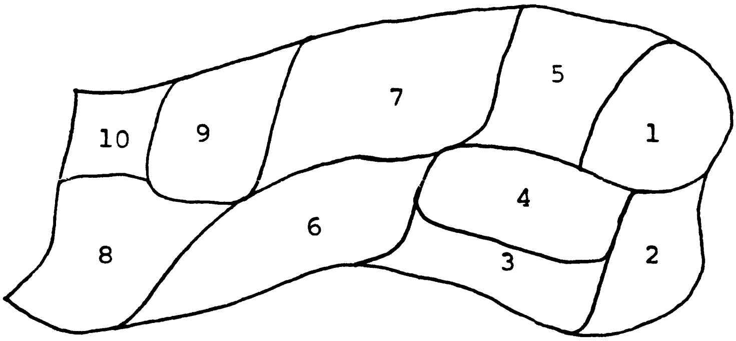

Fig. 1. Lahiri's Serpentine Ordering of PSU's

Algorithm I (Northeast)

```
begin
    $p_{ij} := 0; 1 \leq i \leq m, 1 \leq j \leq n;$
    $i := 1; j := n;$
    while ($j > 0$) or ($i < m$) do
        if $p[i] < q[j]$ then
        begin
            $p_{ij} := p[i];$
            $q[j] := q[j] - p[i];$
            $i := i + 1$
        end
        else if $p[i] \geq q[j]$ then
        begin
            $p_{ij} := q[j];$
            $p[i] := p[i] - q[j];$
            $j := j - 1$
        end
        else
        begin
            $p_{ij} := q[j];$
            $i := i + 1;$
            $j := j - 1$
        end
end.
```

Algorithm II (Positive Variation)

```
begin
    for $i := 1$ to $m$ do;
        for $j := 1$ to $n$ do;
        begin
            positive := true;
            $k := 1;$
            repeat
                $l := 1;$
                repeat
                    if variation($i,j,k,l$) < 0
                        then positive :=
                        false
                    $l := l + 1$
                    until (not positive) or
                          ($l > n$);
                  $k := k + 1$
             until (not positive) or ($k > m$)
end.
```

References

Aragon, J. and Pathak, P. K. (1990): A transportation algorithm for optimal integration of two surveys, to appear in *Sankhya.*

Arthanari, T. S. and Dodge, Y. (1981): *Mathematical Programming in Statistics*, Wiley, New York.

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985): Applications of transportation theory to statistical problems, *J. Amer. Statist. Assoc.*, 80, 903-909.

Chvatal, V. (1983): *Linear Programming*, Freeman, New York.

Felligi, I. P. (1966): Changing the probabilities of selection when two units are selected with pps without replacement, *Proceedings of the Social Statistics Section, Amer. Statist. Assoc.*, 434-442.

Goodman, R. and Kish, L. (1950): Controlled selection–A technique in probability sampling, *J. Amer. Statist. Assoc.* 45, 350-372.

Hanson, R. J. and Hiebert, K. L. (1981): A sparse linear programming subprogram, *Sandia National Laboratory Report* SAND81-0297.

Hoffman, A. J. (1963): On simple linear programming problems, *Proc. Sympos. Pure Math.* 7, 317-327.

Keyfitz, N. (1951): Sampling with probability proportional to size: adjustments for changes in probabilities, *J. Amer. Statist. Assoc.* 46, 105-109.

Kish, L. and Scott, A. (1971): Retaining units after changing strata and probabilities, *J. Amer. Statist. Assoc.* 66, 461-470.

Krishnamoorthy, K. and Mitra, S. K. (1986): Cost robustness of an algorithm for optimal integration of surveys, *Sankhya B* 48, 233-245.

Lahiri, D. B. (1954): Technical paper on some aspects of the development of the sample design, *Sankhya* 14, 264-316.

Maczynski, M. J. and Pathak, P.K. (1980): Integration of surveys, *Scan. J. Statist.* 7, 130-138.

Mitra, S. K. and Pathak, P. K. (1984): Algorithms for optimum integration of two or three surveys, *Scan. J. Statist.* 11, 257-263.

Monge, G. (1781): Mémoire sur la théorie des déblais et des remblais, *Mémoires de L'Académie des Sciences* 19, 666-704.

Raj, D. (1957): On the method of overlapping maps in sample surveys, *Sankhya* 17, 89-98.