Applications of Transportation Theory to Statistical Problems

BEVERLEY D. CAUSEY, LAWRENCE H. COX, and LAWRENCE R. ERNST*

The two-dimensional controlled selection problem and the problem of maximizing the overlap of old and new primary sampling units after restratification and change of selection probabilities have been studied for several decades but have never been completely solved until now. Using transportation theory, complete solutions are obtained here for these and other problems. The solution to the controlled selection problem is based on a specific transportation model that was originally developed, in a previous paper by Cox and Ernst (1982), to solve completely the controlled rounding problem, namely the problem of optimally rounding real-valued entries in a two-way tabular array to adjacent integer values in a manner that preserves the tabular (additive) structure of the array. This model is also applied to other statistical problems, such as raking and statistical disclosure for frequency count tabulations and microdata.

KEY WORDS: Controlled rounding; Controlled selection; Overlap of primary sampling units; Statistical disclosure; Raking.

1. INTRODUCTION

A transportation model is a system of linear constraints over a set of variables $\{y_{ij}\colon 1 \le i \le p,\ 1 \le j \le q\}$ of the form:

$$\sum_{i=1}^{p} y_{ij} = c_j, \qquad 1 \le j \le q$$

$$\sum_{j=1}^{q} y_{ij} = r_i, \qquad 1 \le i \le p$$

$$\sum_{j=1}^{q} \sum_{i=1}^{p} y_{ij} = t,$$

$$y_{ij} \ge 0;\ r_i,\ c_j,\ t \text{ constant.} \tag{1.1}$$

It is equivalent, and for our purposes convenient, to represent the transportation model as a tabular array

$$\begin{matrix} (y_{ij})_{p\times q} & (r_i)_{p\times 1} \\ (c_j)_{1\times q} & (t)_{1\times 1} \end{matrix}, \tag{1.2}$$

that is, a two-way table satisfying the additivity constraints (1.1).

The classical transportation problem is to minimize a linear combination of the $\mathbf{y}$'s subject to the constraints imposed by (1.2) (Dantzig 1963). Transportation problems are linear programming problems whose specialized tabular structure permits solution strategies that are extremely efficient computationally (Glover et al. 1974). Of particular importance is the fact that for transportation problems, integer-valued $\mathbf{r}_i$ and $\mathbf{c}_j$ guarantee that optimal solutions are integer valued.

In this article we present several applications of transportation theory in statistics. Several of these problems are solved by structuring the statistical problem as a "controlled rounding problem" that, by virtue of Cox and Ernst (1982), is solvable as a transportation problem. The problems of this type solved here fall into two categories: general statistical problems that involve replacing nonintegers by integers in tabular arrays and the controlled selection problem in which survey sampling units to be selected according to a specified probability model must also satisfy additional constraints (controls).

Section 2, summarizing Cox and Ernst (1982), presents the transportation solution of the controlled rounding problem. Section 3 presents several statistical applications in which controlled rounding can be used to replace nonintegers by integers in tabular arrays and gives a new application of transportation theory to raking two-way tables. In Section 4 controlled rounding is applied to solve the controlled selection problem. In Section 5 the important problem of maximizing overlap between sets of primary sampling units in redesigning statistical surveys is modeled as a general transportation problem. Complete solutions to the controlled selection and overlap problems have eluded discovery for decades.

2. CONTROLLED ROUNDING AS A TRANSPORTATION PROBLEM

In this section we define and summarize briefly the solution of the controlled rounding problem discovered by Cox and Ernst (1982). The term *rounding* here denotes the replacement of a real number $\mathbf{a}$ by an adjacent integer value $\mathbf{R}(\mathbf{a})$; that is, $R(a)$ equals either $[a]$ or $[a] + 1$, where $[a]$ is the integer part of $\mathbf{a}$. This is base 1 rounding; by dividing all entries by $\mathbf{B}$, the general problem of rounding to a positive integer base $\mathbf{B}$ is always reducible to this case. Let $\mathbf{A}$ denote the tabular array

$$\begin{matrix} (a_{ij})_{m\times n} & (a_{i.})_{m\times 1} \\ (a_{.j})_{1\times n} & (a_{..})_{1\times 1} \end{matrix}. \tag{2.1}$$

A controlled rounding of $\mathbf{A}$ is an array $\mathbf{R}(\mathbf{A})$ that satisfies the following conditions:

For each entry $\mathbf{a}$ of $\mathbf{A}$, including totals, $\mathbf{R}(\mathbf{a})$ is a rounding of $\mathbf{a}$. (2.2)

The array $\mathbf{R}(\mathbf{A})$ is tabular. (2.3)

* Beverley D. Causey and Lawrence R. Ernst are Mathematical Statisticians and Lawrence H. Cox is Senior Mathematical Statistician, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233. The authors thank the associate editor and the referees for their insightful review and one referee for strengthening the proof of (4.1).

In the Public Domain
Journal of the American Statistical Association
December 1985, Vol. 80, No. 392, Theory and Methods

By choice of an integer $\mathbf{p}$, $1 \leq p < \infty$, $\mathbf{R}(\mathbf{A})$ is an optimal controlled rounding of $\mathbf{A}$ if, in addition to (2.2) and (2.3), $\mathbf{R}(\mathbf{A})$ minimizes

$$l_p[R(A), A] = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |R(a_{ij}) - a_{ij}|^p\right)^{1/p} \quad (2.4)$$

or, for $p = \infty$,

$$l_\infty[R(A), A] = \max\{|R(a_{ij}) - a_{ij}| : 1 \leq i \leq m,\ 1 \leq j \leq n\}. \quad (2.5)$$

Conventional rounding, that is, $R(a) = [a + .5]$, always minimizes (2.4) and (2.5); however, conventional rounding often fails (2.3). Causey (1979) presented a heuristic algorithm for controlled rounding; however, this algorithm is not optimal and fails (2.2) or (2.3) in specific examples.

Cox and Ernst (1982) modeled the problem of obtaining an optimal controlled rounding of A as a classical transportation problem in $\{0, 1\}$ variables. Using the general theory of transportation problems, they demonstrated that solutions always exist to the controlled rounding problem (2.2) and (2.3). (A referee pointed out that Baranyai 1975 had discovered an analogous result.) By mimicking the behavior of (2.4) and (2.5) by linear functions of the transportation variables, Cox and Ernst (1982) solved the optimal controlled rounding problem. Furthermore, extending their argument slightly, they showed that the same techniques guarantee and yield optimal solutions to the zero-restricted controlled rounding problem, that is, (2.2) and (2.3), subject to the restriction that integer values must round to themselves. Consequently, controlled rounding problems always may be solved by means of standard, highly efficient algorithms and software (Glover et al. 1974). By means of straightforward modification of (2.4) or (2.5), other familiar measures of overall distortion to $\mathbf{A}$ can also be minimized using these techniques [e.g., for $a_{ij} \neq 0$, divide each term of (2.4) by $|a_{ij}|$].

3. SOME STATISTICAL APPLICATIONS OF CONTROLLED ROUNDING AND TRANSPORTATION THEORY

Controlled rounding can improve the readability of data values and enhance their utility for analysis. It may also be applied to statistical problems for which a complete solution requires that real-number values be replaced by integers (or more generally, integer multiples of $\mathbf{B}$) with minimum overall distortion to $\mathbf{A}$ [e.g., in the sense of (2.4) and (2.5)].

An important application of controlled rounding is controlling statistical disclosure in tables of frequency counts. If small counts can be inferred from the released tables, then possibly individual respondents will be identified and confidential data divulged. The ability of users to infer small counts from the published data therefore must be thwarted. A solution to this problem is to round all entries in the published tables to an appropriately chosen base $\mathbf{B}$, for example, $B = 5$ or 10, and to do so with minimum overall distortion of the data.

Conventional rounding is an inferior method for controlling statistical disclosure: it fails to maintain additivity of detail items to totals, and because the conventional rounding of a value is unique, it is sometimes possible to reconstruct the original array from the rounded array. Nargundkar and Saveland (1972) rounded each frequency count randomly while maintaining the adjacency condition (2.2). For example, with $B = 5$ the value 2 rounds to 0 with probability .6 and to 5 with probability .4. Thus random rounding preserves expected values of individual entries and totals.

Unfortunately, random rounding often fails the additivity condition (2.3). Fellegi (1975) preserved adjacency, expected values, and additivity but only for one-way tables. A single controlled rounding of a two-way table preserves adjacency and additivity but not expected values. As described in Section 4, however, by virtue of (4.1)–(4.3) it is always possible to choose a set $\mathbf{N}$ of controlled roundings together with associated probabilities of selection such that under random selection of one controlled rounding from this set, the expectation of the rounded value for each entry equals the original unrounded value. Moreover, by suitable choice of the members of $\mathbf{N}$, data distortion can be reduced subject to preserving adjacency, additivity, and expected values.

Controlled rounding may also be used to prevent statistical disclosure in microdata release. Suppose for the two-way table $\mathbf{A}$ that $\mathbf{a}_{ij}$ is the sum of $\mathbf{k}_{ij}$ quantities $\mathbf{a}_{ijk}$, $k = 1, \ldots, k_{ij}$, each of which is to be rounded to a multiple of $\mathbf{B}$. Assume as before that $B = 1$. We perform controlled rounding on the array $\mathbf{A}$ so that each $\mathbf{a}_{ij}$ is replaced by an adjacent integer $\mathbf{R}(\mathbf{a}_{ij})$. Next, for cell (i, j) we want to round some of the values $\mathbf{a}_{ijk}$ up and some down to obtain a total $\mathbf{R}(\mathbf{a}_{ij})$ for the cell while minimizing overall distortion. Let $\mathbf{L}$ denote $R(a_{ij}) - \sum_k [a_{ijk}]$; we set the rounded value of $\mathbf{a}_{ijk}$ to $[a_{ijk}] + 1$ for the $\mathbf{L}$ largest values among the $\mathbf{k}_{ij}$ quantities $a_{ijk} - [a_{ijk}]$ and to $[a_{ijk}]$ otherwise. As a possible example of application of this method, the $\mathbf{a}_{ijk}$ might be personal incomes to be rounded to multiples of \$5,000 before presentation to preclude identification of particular individuals and their exact incomes, with $\mathbf{i}$ and $\mathbf{j}$ corresponding to sex and race categories. For each sex–race (i, j) cell we would reveal the rounded sum, rather than the exact sum, of the $\mathbf{k}_{ij}$ incomes for that category. In addition to controlling statistical disclosure, this technique maintains consistency between estimates derived either from $\mathbf{A}$ or from the microdata directly.

Controlled rounding is also applicable to the problem of iterative proportional fitting (or raking) in two-way tables of counts as considered by Ireland and Kullback (1968). Given a two-way table $\mathbf{A}'$ of integer counts $\mathbf{a}'_{ij}$, we seek to construct a revised table of integer counts $\mathbf{A}$, whose row and column sums $\mathbf{a}_{i.}$ and $\mathbf{a}_{.j}$ have been predetermined, to minimize distortion to the original table. As an example, i might correspond to "race," j to "county," $\mathbf{a}'_{ij}$ to the count of persons by race and county according to the 1980 census, $\mathbf{a}_{i.}$ and $\mathbf{a}_{.j}$ to known numbers of persons in 1985 (based on demography, administrative records, or otherwise), and $\mathbf{a}_{ij}$ to estimated numbers of persons in 1985. The raking procedure is based on repeated uniform multiplication of entries in each row and then each column to satisfy the desired marginals. In general, the resulting $\mathbf{a}_{ij}$'s are noninteger, in which case controlled rounding with $B = 1$ can be applied to $\mathbf{A}$ to obtain an integer array as desired.

Raking requires zeros to appear in $\mathbf{A}$ whenever they appear in $\mathbf{A}'$. Although the iterative proportional fitting algorithm for raking will always converge to a solution table $\mathbf{A}$ with the desired marginals if such a solution exists (Darroch and Ratcliff 1972), sometimes the presence of too many zeros in $\mathbf{A}'$ makes raking impossible. These zeros may be either sampling or structural zeros. Using transportation theory, one may identify nonconvergence in advance (using the method of Fagan and Greenberg 1984). If raking will not converge, we may use transportation theory a second time, as follows, to obtain a reduced set of zero constraints for which raking will converge. (Typically, but not necessarily, we require that this set contain the structural zeros.)

Use general transportation theory to construct a table $\mathbf{A}''$ with the desired marginals and structural zeros (if such exists), such that the sum of the internal entries corresponding to the sampling zeros of $\mathbf{A}'$ is minimal. Thus there is minimal overall distortion from the original problem. Subtract from $\mathbf{a}_{i.}$, $\mathbf{a}_{.j}$, $\mathbf{A}'$, and $\mathbf{A}''$ the entries of $\mathbf{A}''$ corresponding to the sampling zeros and apply iterative proportional fitting to the new $\mathbf{A}'$. Because the new $\mathbf{A}''$ satisfies the zero and marginal constraints for raking the new $\mathbf{A}'$ with respect to the new marginals, iterative proportional fitting must converge (Darroch and Ratcliff 1972). By adding back all subtracted entries to this solution, a raking solution $\mathbf{A}$ with respect to the original $\mathbf{A}'$ and marginals is obtained. Our method minimizes the overall change to the original (sampling) zeros of $\mathbf{A}'$ (actually, we can do so for any linear representation of distortion); another approach is to "smooth" $\mathbf{A}'$ by a Bayesian technique (Bishop et al. 1975, chap. 12).

4. CONTROLLED SELECTION

Consider a population partitioned by two criteria of classification, $\mathbf{m}$ rows and $\mathbf{n}$ columns, resulting in a two-way table of $\mathbf{mn}$ classification cells. A sample of size $\mathbf{r}$ is to be selected, with $\mathbf{s}_{ij}$ denoting the expected number of sample units in the $\mathbf{ij}$th cell. Let $\mathbf{S}$ denote a tabular array with internal elements $(s_{ij})_{m\times n}$. We wish to limit the deviation from $\mathbf{s}_{ij}$ of the number of sample units in the $\mathbf{ij}$th cell to less than one and similarly limit the deviation for the row and column totals, while strictly maintaining the requirements of probability sampling. This will be done by construction of a finite sequence of arrays $N_1 = (n_{ij1}), N_2 = (n_{ij2}), \ldots, N_l = (n_{ijl})$ and associated probabilities $\mathbf{p}_1, \ldots, \mathbf{p}_l$, satisfying the following:

$\mathbf{N}_k$ is a zero-restricted controlled rounding of $\mathbf{S}$ for all $\mathbf{k}$, (4.1)

$$\sum_{k=1}^{l} p_k = 1, \tag{4.2}$$

$$E(n_{ijk} \mid i, j) = \sum_{k=1}^{l} n_{ijk}\, p_k = s_{ij},$$

$$1 \le i \le m + 1,\ 1 \le j \le n + 1. \tag{4.3}$$

For notational simplicity, throughout this section [including (4.3) above] the $\mathbf{i}$th row total, the $\mathbf{j}$th column total, and the grand total of a tabular array $\mathbf{A}$ will be denoted respectively by $\mathbf{a}_{i(n+1)}$, $\mathbf{a}_{(m+1)j}$, and $\mathbf{a}_{(m+1)(n+1)}$ as alternatives to $\mathbf{a}_{i.}$, $\mathbf{a}_{.j}$, and $\mathbf{a}_{..}$.

The preceding conditions, which can be generalized to dimensions greater than two, are a special case of conditions to be satisfied for the sampling technique known as controlled selection, first described by Goodman and Kish (1950). An example was given in that paper to illustrate the application of controlled selection, but no general method to solve such problems was presented. Bryant et al. (1960) developed a simple method for approaching the two-way stratification problem just described, but it does not in general satisfy all these requirements exactly. Jessen (1970) considered the identical requirements that we have imposed but presented no general procedure for obtaining a set of arrays that satisfies these conditions. Groves and Hess (1975) presented a formal algorithm for obtaining solutions to the two-dimensional and the much more complex three-dimensional problems. They made no claim, however, that their algorithm will always yield a solution; and there are indeed simple examples where it fails, even in the two-way case. Ernst (1981) presented the first complete procedure for the two-dimensional case, but that procedure is laborious. In this section, we present an efficient algorithm for controlled selection that employs the results described in Section 2 to completely solve the two-dimensional problem, together with an illustrative example. We also show by example that the three-dimensional problem does not always have a solution.

4.1 The Algorithm

We proceed to recursively define the sequences $\mathbf{N}_1, \ldots, \mathbf{N}_l$ and $\mathbf{p}_1, \ldots, \mathbf{p}_l$, satisfying (4.1)–(4.3). For fixed $\mathbf{k}$, to define $\mathbf{N}_k$, $\mathbf{p}_k$ we begin with the tabular array $A_k = (a_{ijk})$. $A_1 = S$; and for $k \ge 1$, $\mathbf{A}_{k+1}$ will be defined in terms of $\mathbf{N}_k$, $\mathbf{p}_k$.

$\mathbf{N}_k$ is simply a zero-restricted controlled rounding of $\mathbf{A}_k$. To define $\mathbf{p}_k$, first let

$$d_k = \max\{|n_{ijk} - a_{ijk}| : 1 \le i \le m + 1,\ 1 \le j \le n + 1\}, \tag{4.4}$$

and then let

$$\begin{aligned} p_k &= 1 - d_k && \text{if } k = 1, \\ &= \left(1 - \sum_{i=1}^{k-1} p_i\right)(1 - d_k) && \text{if } k > 1. \end{aligned} \tag{4.5}$$

If $d_k > 0$, define $\mathbf{A}_{k+1}$ by letting

$$a_{ij(k+1)} = n_{ijk} + (a_{ijk} - n_{ijk})/d_k \tag{4.6}$$

for all $\mathbf{i}$, $\mathbf{j}$, and then proceed to define $\mathbf{N}_{k+1}$, $\mathbf{p}_{k+1}$.

We now show that eventually there will be an integer $\mathbf{l}$ for which $d_l = 0$ and that this will terminate the algorithm; that is, $\mathbf{N}_1, \ldots, \mathbf{N}_l$ and $\mathbf{p}_1, \ldots, \mathbf{p}_l$ satisfy (4.1)–(4.3).

Proof of (4.1). We note that it suffices to show that for all $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$,

$$[a_{ijk}] \le a_{ij(k+1)} \le [a_{ijk}] + 1 \tag{4.7}$$

and

$$a_{ij(k+1)} = a_{ijk} \quad \text{if } \mathbf{a}_{ijk} \text{ is an integer}, \tag{4.8}$$

since (4.7) and (4.8) imply that a zero-restricted controlled rounding of any of the $\mathbf{A}_k$ is a zero-restricted controlled rounding of every preceding member of the sequence, including $A_1 = S$.

To establish (4.7) and (4.8), first let $Q = a_{ijk} - n_{ijk}$ and $R = n_{ijk} - [a_{ijk}]$. Then $0 \le R + Q/d_k \le 1$, because by the definitions of $\mathbf{N}_k$ and $\mathbf{d}_k$, either $R = 0$ and $0 \le Q/d_k \le 1$ or $R = 1$ and $-1 \le Q/d_k \le 0$. Since $a_{ij(k+1)} = [a_{ijk}] + R + Q/d_k$, (4.7) follows. Furthermore, if $\mathbf{a}_{ijk}$ is an integer, then $Q = R = 0$, hence (4.8).

Proof That $d_l = 0$ for Some l. For any (i, j), if $\mathbf{a}_{ijk}$ becomes an integer for some $\mathbf{k}$, then by (4.8) it will remain so for all subsequent values of $\mathbf{k}$. Furthermore, by (4.4) and (4.6), for each $k \ge 2$, this happens for at least one (i, j). Consequently, there exists a smallest integer $\mathbf{l}$ such that $\mathbf{a}_{ijl}$ is an integer for all (i, j) and $d_l = 0$ by (4.4).

Proof of (4.2). Simply substitute $k = l$ and $d_l = 0$ in (4.5).

Proof of (4.3). Solving (4.6) for $\mathbf{n}_{ijk}$ and multiplying by $\mathbf{p}_k$, we obtain

$$p_k n_{ijk} = p_k/(1 - d_k)(a_{ijk} - d_k a_{ij(k+1)}),$$

which together with (4.5) yields

$$p_1 n_{ij1} = a_{ij1} + (p_1 - 1)a_{ij2},$$

$$p_k n_{ijk} = \left(1 - \sum_{s=1}^{k-1} p_s\right)a_{ijk} + \left(\sum_{s=1}^{k} p_s - 1\right)a_{ij(k+1)}, \qquad 2 \le k \le l - 1.$$

Consequently, $\sum_{k=1}^{l-1} p_k n_{ijk} = a_{ij1} + (\sum_{s=1}^{l-1} p_s - 1)a_{ijl}$, which is then combined with the relations $a_{ij1} = s_{ij}$, $n_{ijl} = a_{ijl}$, and (4.2) to obtain (4.3).

4.2 An Example

Figure 1 presents an example. Note that the solution given is not unique.

$$\text{Let } S = \begin{array}{cccc} .4 & 2.0 & .0 & 2.4 \\ 1.2 & .0 & 1.0 & 2.2 \\ .2 & .0 & .0 & .2 \\ 1.2 & .4 & .2 & 1.8 \\ 1.0 & .6 & .2 & 1.8 \\ .0 & .4 & .4 & .8 \\ .0 & .2 & .4 & .6 \\ .0 & .0 & .2 & .2 \\ & & & \\ 4.0 & 3.6 & 2.4 & 10.0 \end{array}$$

$$\text{Then } A_1 = S,\quad N_1 = \begin{array}{cccc} 1 & 2 & 0 & 3 \\ 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 \\ 1 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ & & & \\ 4 & 4 & 2 & 10 \end{array}$$

$d_1 = .6,\ p_1 = .4,$

$$A_2 = \begin{array}{cccc} 0 & 2 & 0 & 2 \\ 1\frac{1}{3} & 0 & 1 & 2\frac{1}{3} \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 1\frac{1}{3} & 0 & \frac{1}{3} & 1\frac{2}{3} \\ 1 & \frac{1}{3} & \frac{1}{3} & 1\frac{2}{3} \\ 0 & \frac{2}{3} & 0 & \frac{2}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} & 1 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ & & & \\ 4 & 3\frac{1}{3} & 2\frac{2}{3} & 10 \end{array} \qquad N_2 = \begin{array}{cccc} 0 & 2 & 0 & 2 \\ 2 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ & & & \\ 4 & 4 & 2 & 10 \end{array}$$

$d_2 = \frac{2}{3},\ p_2 = .2,$

$$A_3 = \begin{array}{cccc} 0 & 2 & 0 & 2 \\ 1 & 0 & 1 & 2 \\ .5 & 0 & 0 & .5 \\ 1.5 & 0 & 0 & 1.5 \\ 1 & 0 & .5 & 1.5 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & .5 & .5 \\ & & & \\ 4 & 3 & 3 & 10 \end{array} \qquad N_3 = \begin{array}{cccc} 0 & 2 & 0 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ & & & \\ 4 & 3 & 3 & 10 \end{array}$$

$d_3 = .5,\ p_3 = .2,$

$$N_4 = A_4 = \begin{array}{cccc} 0 & 2 & 0 & 2 \\ 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 2 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ & & & \\ 4 & 3 & 3 & 10 \end{array}$$

$d_4 = 0,\ p_4 = .2.$

Figure 1. An Example.

Remark 4.1. From the proof given in Section 4.1 that $d_l = 0$ for some $\mathbf{l}$, it immediately follows that an upper bound on $\mathbf{l}$ is $(m + 1)(n + 1)$, and this can be lowered to $\mathbf{mn}$ because of the constraints imposed by the totals. It can also be shown that another upper bound on $\mathbf{l}$ is the greatest common divisor of the denominators of all of the entries in $\mathbf{S}$, when each entry is expressed in reduced form. For the example illustrated in Figure 1, this last upper bound is 5, only one more than the actual value of $\mathbf{l}$.

Remark 4.2. One potential use of this controlled selection algorithm occurs in the selection of sample primary sampling units (PSU's; contiguous geographic areas) for a redesign of the household surveys conducted by the U.S. Bureau of the Census. Each row in the controlled selection problem can be considered as representing a stratum; when the design calls for one PSU per stratum, the row totals would be 1. The columns would arise from a classification variable that is not part of the original stratification. For example, the state in which a PSU is located has been used as a classification variable in previous redesigns. Each PSU's position in $\mathbf{S}$ would then be determined by its stratum and state. Since the PSU's are selected with probability proportional to size, as determined by population data from the most recent census, the entry in each internal cell of $\mathbf{S}$ would be the sum of the probabilities of the PSU's in that cell. The column totals would be the expected number of PSU's to be selected from each state. Controlled selection would guarantee that the number of sample PSU's in each state is within one of the expected number. The selected $\mathbf{N}_k$ would be a $\{0, 1\}$ array with a single 1 in each row, which would indicate the state in which the sample PSU would be located in that stratum. In any stratum in which there is more than one PSU from the selected state, a single PSU would be selected from among them with probability proportional to size.

4.3 Three-Way Stratification

The following example illustrates that the three-way stratification problem does not always have a solution.

Consider a population subject to a $2 \times 2 \times 2$ stratification from which a sample of size 2 is to be drawn. The expected number of sample units in the **ijk**th stratum cell, $\mathbf{s}_{ijk}$, is as follows:

$$s_{111} = s_{221} = s_{122} = s_{212} = .5,$$

$$s_{121} = s_{211} = s_{112} = s_{222} = 0.$$

We demonstrate that there is no solution by proving that there exists no $2 \times 2 \times 2$ integer-valued matrix $N = (n_{ijk})$ such that

$$|n_{ijk} - s_{ijk}| < 1 \quad \text{for } i = 1, 2, j = 1, 2, k = 1, 2, \tag{4.9}$$

$$|n_{i\cdot\cdot} - s_{i\cdot\cdot}| < 1 \quad \text{for } i = 1, 2, \tag{4.10}$$

$$|n_{\cdot j\cdot} - s_{\cdot j\cdot}| < 1 \quad \text{for } j = 1, 2, \tag{4.11}$$

$$|n_{\cdot\cdot k} - s_{\cdot\cdot k}| < 1 \quad \text{for } k = 1, 2, \tag{4.12}$$

$$n_{\cdots} = 2. \tag{4.13}$$

To show this we note that there are six possible **N**'s that will satisfy (4.9) and (4.13). The nonzero elements of **N** for each of these possibilities are as follows: $n_{111} = n_{221} = 1$, $n_{111} = n_{122} = 1$, $n_{111} = n_{212} = 1$, $n_{221} = n_{122} = 1$, $n_{221} = n_{212} = 1$, and $n_{122} = n_{212} = 1$. The first possibility fails, since $n_{\cdot\cdot 1} = 2$ and $s_{\cdot\cdot 1} = 1$ and hence (4.12) is not satisfied. Similarly, the others fail to satisfy (4.10), (4.11), (4.11), (4.10), and (4.12), respectively.

5. OPTIMAL METHOD FOR MAXIMIZING THE OVERLAP BETWEEN SURVEYS

Consider a periodic survey with a multistage stratified design, where the PSU's are contiguous geographic areas. Typically, to increase the precision of the estimates, PSU's with similar characteristics are grouped together in the same stratum and the sample PSU's are selected proportional to some measure of size associated with each PSU, such as the population in the most recent census. At some point a redesign is undertaken in which the PSU's remain the same but the strata and selection probabilities change because more current data have become available, such as data from a new census, that would lead to improved stratification and measures of size. The new sample PSU's may of course be selected independently of the initial PSU's. However, additional costs are generally incurred with each change of sample PSU. Consequently, it may be considered desirable to maximize the expected number of old PSU's retained in the new design, while strictly maintaining the requirements of probability sampling.

Keyfitz (1951) presented an optimum procedure for one PSU per stratum designs in the special case in which the initial and new strata are identical, with only the selection probabilities changing. For the more general one PSU per stratum problem for which the strata definitions can change in the redesign, Perkins (1970) and Kish and Scott (1971) presented procedures that are not optimal. Fellegi (1966) considered a particular type of two PSU's per stratum problem, but his procedure is also not optimal.

In this section a relatively simple optimal procedure is obtained by formulating the problem as a transportation problem. This procedure is very general, with no restrictions on changes in strata definitions or number of PSU's per stratum. Raj (1968) previously employed the transportation problem approach, but only with the restrictive assumptions considered by Keyfitz.

5.1 The Procedure

Each stratum **S** in the new design represents a separate problem. Corresponding to each possible set of initial sample PSU's is a subset of **S** consisting of all PSU's in **S** that were in the initial sample. Let $\mathbf{I}_1, \ldots, \mathbf{I}_m$ denote all possible such subsets of **S**, and let $\mathbf{p}_i$ $(1 \le i \le m)$ be the probability that $\mathbf{I}_i$ is the set of initial sample PSU's in **S**. For example, if the initial design was **k** PSU's per stratum, then a subset **I** of **S** would be among the $\mathbf{I}_i$'s iff no more than **k** PSU's from each initial stratum are in **I** and at least **k** PSU's in each initial stratum are not in $S \sim I$. Similarly, let $S_1, \ldots S_n$ denote all possibilities for the subset of **S** consisting of all the new sample PSU's in **S**, and let $\boldsymbol{\pi}_j$ $(1 \le j \le n)$ be the probability that $\mathbf{S}_j$ is the set of new PSU's in **S**. For example, if the new design is **k** PSU's per stratum without replacement, then $\mathbf{S}_1, \ldots, \mathbf{S}_n$ would be all subsets of **S** with exactly **k** elements. For $1 \le i \le m$ and $1 \le j \le n$, let $\mathbf{x}_{ij}$ denote the joint probability that $\mathbf{I}_i$ is the set of initial sample PSU's in **S** and $\mathbf{S}_j$ is the set of new PSU's in **S**, and let $\mathbf{c}_{ij}$ be the number of PSU's in $I_i \cap S_j$. The $\mathbf{p}_i$'s, $\boldsymbol{\pi}_j$'s, and $\mathbf{c}_{ij}$'s are known values, whereas the $\mathbf{x}_{ij}$'s are variables for which optimal values are to be determined.

The goal of the procedure is to obtain probabilities for the new outcomes, conditional on the initial outcome, that maximize the expected number of initial sample PSU's retained in the new sample, while preserving the unconditional probabilities $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n$. We are further constrained by the given $\mathbf{p}_i$'s. Of course, for this to be done legitimately, the full set of conditional probabilities corresponding to all pairs of initial and new outcomes must first be determined without regard to the actual initial outcome, and then from this full set, the subset corresponding to the actual initial outcome will be used. To obtain our goal we first solve the following transportation problem: Determine $x_{ij} \ge 0$ that maximize

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}, \tag{5.1}$$

subject to

$$\sum_{j=1}^{n} x_{ij} = p_i, \qquad 1 \le i \le m, \tag{5.2}$$

$$\sum_{i=1}^{m} x_{ij} = \pi_j, \qquad 1 \le j \le n. \tag{5.3}$$

Table 1. Sets of Initial Sample PSU's and Corresponding p_i's

	m											
	1	2	3	4	5	6	7	8	9	10	11	12
Sample PSU's	{1}	{2}	{3}	{4}	{5}	{1, 4}	{1, 5}	{2, 4}	{2, 5}	{3, 4}	{3, 5}	∅
p_i	.15	.018	.012	.24	.04	.30	.05	.036	.006	.024	.004	.12

Note that in this transportation problem, the (objective) function (5.1) is the expected number of PSU's in **S** that are in both the initial and new samples, and note that constraints (5.2) and (5.3) are required by the definitions of the $\mathbf{p}_i$'s, $\boldsymbol{\pi}_j$'s, and $\mathbf{x}_{ij}$'s.

Then once the optimal $\mathbf{x}_{ij}$'s have been obtained, the conditional new selection probabilities given that $\mathbf{I}_i$ is the set of initial sample PSU's in **S** are x_{ij}/p_i, $1 \le j \le n$.

5.2 An Example

In this example the initial and new designs are both one PSU per stratum. **S** consists of five PSU's with new selection probabilities .40, .15, .05, .30, and .10. Then $n = 5$ and the given probabilities are the values of $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_5$, respectively.

We are further given that the initial selection probabilities for these five PSU's were .50, .06, .04, .60, and .10; that the first three PSU's were in one initial stratum and the other two in a second initial stratum; and that the sampling was done independently in each initial stratum. Then $m = 12$ with the possible sets of initial sample PSU's and corresponding $\mathbf{p}_i$'s given in Table 1. Furthermore, the $\mathbf{c}_{ij}$'s are then as given in Table 2.

On maximizing (5.1) subject to (5.2) and (5.3) with the given $\mathbf{p}_i$'s, $\boldsymbol{\pi}_j$'s, and $\mathbf{c}_{ij}$'s, an optimal set of $\mathbf{x}_{ij}$'s, given in Table 3, is found. The conditional probabilities of selection given that $\mathbf{I}_i$ is the set of initial sample PSU's in **S** can be obtained by dividing the ith row of Table 3 by $\mathbf{p}_i$. The corresponding maximum value of the objective function is .880, compared to an overlap probability of .402 if the new sample is selected independently of the initial sample.

Remark 5.1. If it is desired to minimize the expected number of retained sample PSU's instead of maximizing this quantity, simply minimize (5.1) subject to (5.2) and (5.3).

Remark 5.2. Although modern computers are capable of solving rather large transportation problems rapidly, the values of both **m** and **n** can be very large if **S** is of moderate size and more than, say, two PSU's are selected per stratum. Furthermore, **m** can also be very large in two PSU's per stratum design if **S** contains PSU's from several initial strata. Therefore, in any potential application of the procedure, it is first necessary to compute the size of the resulting transportation problems to determine whether this procedure is practical. When the amount of computation required for the optimal procedure is too large for certain strata, a combined procedure that employs a computationally simpler overlap procedure for those strata can be used instead, and the optimal procedure can be employed for the other strata.

Remark 5.3. The procedure that has been described in this section requires that the joint selection probability in the initial sample be known for any set of PSU's that are in the same stratum in the new design; that is, the $\mathbf{p}_i$'s must be known. If the initial sample, however, was not chosen independently from stratum to stratum, this information may not be available. In Ernst (1985) an alternative overlap procedure was presented that only requires knowledge of the joint selection probabilities in the initial sample for sets of PSU's that are in the same initial and new strata, which, in certain circumstances, is optimal among all procedures that require only this amount of information. This alternative procedure formulates the overlap problem as a linear programming problem, but not a transportation problem.

The household surveys conducted by the U.S. Bureau of the Census were recently redesigned. An overlap procedure was used in the selection of sample PSU's for two of these surveys. In both cases the initial sample PSU's had not been chosen independently from stratum to stratum and the $\mathbf{p}_i$'s were not known. Consequently, the procedure of Ernst (1985) was used instead of the procedure described in this section. The reasons for this lack of independence were that both controlled

Table 2. Values of c_{ij}

	j				
i	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	1	0
7	1	0	0	0	1
8	0	1	0	1	0
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 3. Values of $\mathbf{x}_{ij}$ That Maximize (5.1)

	j				
i	1	2	3	4	5
1	.150	.000	.000	.000	.000
2	.000	.018	.000	.000	.000
3	.000	.000	.012	.000	.000
4	.000	.000	.000	.240	.000
5	.000	.000	.000	.000	.040
6	.240	.000	.000	.060	.000
7	.000	.000	.000	.000	.050
8	.000	.036	.000	.000	.000
9	.000	.000	.000	.000	.006
10	.000	.000	.024	.000	.000
11	.000	.000	.000	.000	.004
12	.010	.096	.014	.000	.000

selection and overlap procedures had been used in previous redesigns of these surveys, and these procedures can be shown to destroy such independence. In redesigning a survey, however, in which neither controlled selection nor an overlap procedure had been used previously, the overlap procedure described here may be usable.

[Received May 1983. Revised April 1985.]

REFERENCES

Baranyai, Z. S. (1975), "On the Factorization of the Complete Uniform Hypergraph," *Infinite and Finite Sets,* 10, 91–108.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice,* Cambridge, MA: MIT Press.

Bryant, E. C., Hartley, H. O., and Jessen, R. J. (1960), "Design and Estimation in Two-Way Stratification," *Journal of the American Statistical Association,* 55, 105–124.

Causey, Beverley (1979), "Approaches to Statistical Disclosure," in *Proceedings of the Social Statistics Section, American Statistical Association,* pp. 306–309.

Cox, Lawrence, and Ernst, Lawrence (1982), "Controlled Rounding," *INFOR,* 20, 423–432.

Dantzig, George (1963), *Linear Programming and Extensions,* Princeton, NJ: Princeton University Press.

Darroch, J., and Ratcliff, D. (1972), "Generalized Iterative Scaling for Log-Linear Models," *Annals of Mathematical Statistics,* 43, 1470–1480.

Ernst, Lawrence R. (1981), "A Constructive Solution for Two-Dimensional Controlled Selection Problems," in *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 61–64.

——— (1985), "Maximizing the Overlap Between Surveys When Information Is Incomplete," Report CENSUS/SRD/RR-84/31, U.S. Bureau of the Census, Statistical Research Division.

Fagan, James, and Greenberg, Brian (1984), "Making Tables Additive in the Presence of Zeroes," in *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 195–200.

Fellegi, Ivan P. (1966), "Changing the Probabilities of Selection When Two Units Are Selected With PPS Without Replacement," in *Proceedings of the Social Statistics Section, American Statistical Association,* pp. 434–442.

——— (1975), "Controlled Random Rounding," *Survey Methodology,* 1, 123–135.

Glover, Fred, Karney, Dennis, Klingman, Darwin, and Napier, Albert (1974), "A Computation Study on Start Procedures, Basic Change Criteria and Solution Algorithms for Transportation Problems," *Management Sciences,* 20, 793–813.

Goodman, Roe, and Kish, Leslie (1950), "Controlled Selection—A Technique in Probability Sampling," *Journal of the American Statistical Association,* 45, 350–372.

Groves, R. M., and Hess, I. (1975), "An Algorithm for Controlled Selection," in *Probability Sampling of Hospitals and Patients* (2nd ed.), eds. I. Hess, D. C. Riedel, and T. B. Fitzpatrick, Ann Arbor, MI: Health Administration Press, pp. 82–102.

Ireland, C. T., and Kullback, S. (1968), "Contingency Tables With Given Marginals," *Biometrika,* 55, 179–188.

Jessen, R. J. (1970), "Probability Sampling With Marginal Constraints," *Journal of the American Statistical Association,* 65, 776–795.

Keyfitz, Nathan (1951), "Sampling With Probabilities Proportionate to Size: Adjustment for Changes in Probabilities," *Journal of the American Statistical Association,* 46, 105–109.

Kish, Leslie, and Scott, A. (1971), "Retaining Units After Changing Strata and Probabilities," *Journal of the American Statistical Association,* 66, 461–470.

Nargundkar, Mukund, and Saveland, Walter (1972), "Random Rounding: A Means of Preventing Disclosure of Information About Individual Respondents in Aggregate Data," in *Proceedings of the Social Statistics Section, American Statistical Association,* pp. 382–385.

Perkins, Walter M. (1970), "1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata," memorandum to Joseph Waksberg, U.S. Bureau of the Census.

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