APPLICATIONS OF TRANSPORTATION THEORY TO STATISTICAL PROBLEMS

Beverley D. Causey, Lawrence H. Cox, and Lawrence R. Ernst, Bureau of the Census

1. INTRODUCTION

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

$$\begin{aligned} &\sum_{i=1}^{p} y_{ij} = c_j, && 1\le j\le q \\ &\sum_{j=1}^{q} y_{ij} = r_i, && 1\le i\le p \qquad (1.1)\\ &\sum_{j=1}^{q}\sum_{i=1}^{p} y_{ij} = t, && \\ &y_{ij}\ge 0, && \\ &r_i,\ c_j,\ t \ \text{constant.} && \end{aligned}$$

It is equivalent and often convenient to represent (1.1) by the standard transportation array

$$\begin{array}{c|c} (y_{ij})_{p\times q} & (r_i)_{p\times 1} \\ \hline (c_j)_{1\times q} & (t)_{1\times 1} \end{array} \qquad (1.2)$$

This is a tabular array, i.e., the horizontal and vertical lines denote the property that detail entries along each row or column must add to the corresponding row or column total entry and, similarly, the r_i (respectively, the c_j) must add to the grand total entry t.

The transportation problem is that of minimizing a linear combination of the y's (called the objective function) subject to the constraints imposed by (1.2) (Dantzig 1963). Thus, transportation problems are linear programming problems. The specialized tabular structure of the transportation problem permits solution strategies which are extremely efficient computationally, even for problems which are large by conventional linear programming standards (Glover, et al. 1974). In addition, transportation problems enjoy the property that integer-valued r_i and c_j guarantee that optimal solutions are also integer-valued (Dantzig 1963, pg. 305).

The book by Raj (1968) presents some early work on applications of mathematical programming theory to statistical problems. In this paper, we present several applications of transportation theory in statistics. Several of these problems are solved by structuring the statistical problem as a so-called "controlled rounding problem" which, 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 which involve replacing non-integers by integers in tabular arrays, and the controlled selection problem in which survey sampling units which are to be selected according to a specified probability model also must satisfy additional constraints (or controls).

Portions of the original paper, principally the list of references and the proofs justifying the procedure of Section 4, have been omitted here due to lack of space. The complete paper is available from the authors.

2. CONTROLLED ROUNDING AS A TRANSPORTATION PROBLEM

In typical usage, the term "rounding" connotes conventional rounding to the base 1, in which a real number a is replaced by ("rounded to") the next closest integer value (with 0.5 rounded to 1 by convention). Although conventional rounding minimizes standard measures of the overall discrepancy between corresponding rounded and unrounded values, it has one important shortcoming: given a collection of unrounded values which add to a total value, the sum of the corresponding conventionally rounded values often fails to add to the conventional rounding of the total. For example, $0.9+0.5=1.4$ but $1+1\neq 1$. Therefore, to maintain additivity of rounded detail entries to rounded totals in base 1 rounding, the requirement that values round to the next closest integer value must be relaxed. The natural relaxation of this constraint in base 1 rounding is to allow each entry to be rounded to either of the two integer values adjacent to it. Under these conditions, the problem of maintaining additivity in the array of rounded values is trivial for the case of one-way tables. Also for one-way tables, it is somewhat more complex computationally but conceptually no more demanding to require, in addition to additivity, that the rounding minimize one among several standard measures of overall discrepancy between the rounded and unrounded arrays. Other requirements can be met with relative ease in one-way tables (Fellegi 1975). However, the problem of constructing roundings which are additive and achieve minimum discrepancy in two-way tables is much more profound. The first author to study this problem was Causey (1979). Causey postulated the notion of controlled rounding in two-way tables and provided a heuristic solution which maintained additivity in some but not all examples. The method of Cox and Ernst (1982) solved both the additivity and minimization of discrepancy problems completely for both one and two-way tables by modelling the so-called controlled rounding problem (defined below) as a transportation problem. The remainder of this section is devoted to a summary of their work and related questions. The reader unfamiliar with transportation theory should consult Dantzig (1963).

Let $[a]$ denote the integer part of the real number a. Given a tabular array A

$$\begin{array}{c|c} (a_{ij})_{m\times n} & (a_{i.})_{m\times 1} \\ \hline (a_{.j})_{1\times n} & (a_{..})_{1\times 1} \end{array} \qquad (2.1)$$

and a positive integer B, called the rounding base, a controlled rounding of A is an array R(A) which satisfies the conditions:

for each entry a of A (including
the totals entries), the corresponding (2.2)
entry R(a) of R(A) equals either B[a/B]
or B([a/B]+1), viz., each entry of A is
rounded either down or up to an
adjacent integer multiple of B,

and,

the array R(A) is tabular. (2.3)

Usually optimal controlled roundings are sought. These are controlled roundings which, seeking to minimize the distortion to the data resulting from replacing A by R(A), minimize a predetermined measure of discrepancy between A and R(A) such as the sum of squares of differences between rounded and unrounded entries.

By dividing all entries of A by B and subtracting the integer parts of the resulting internal entries a_{ij} from corresponding totals entries, an equivalent problem is achieved in which $0\leq a_{ij}<1$ for $1\leq i\leq m$, $1\leq j\leq n$, and for which B=1. Throughout the remainder of this section, we assume that these conditions hold. Cox and Ernst (1982) formulate the controlled rounding problem in terms of the {0,1} - variables

$$\begin{aligned} x_{ij} &= R(a_{ij}), \\ x_{i.} &= R(a_{i.}) - [a_{i.}], \\ x_{.j} &= R(a_{.j}) - [a_{.j}], \\ x_{..} &= R(a_{..}) - [a_{..}]. \end{aligned} \qquad (2.4)$$

and demonstrate that the existence of an R(A) satisfying (2.2)-(2.3) is equivalent to the existence of a set of {0,1} values in the x-variables satisfying the constraints imposed within the tabular array

$$\begin{array}{cc|c} (x_{ij})_{m\times n} & (1-x_{i.})_{m\times 1} & ([a_{i.}+1])_{m\times 1} \\ (1-x_{.j})_{1\times n} & (x_{..})_{1\times 1} & \left(\sum_{j=1}^{n}[a_{.j}+1]-[a_{..}]\right)_{1\times 1} \\ \hline ([a_{.j}+1])_{1\times n} & \left(\sum_{i=1}^{m}[a_{i.}+1]-[a_{..}]\right)_{1\times 1} & \left(\sum_{i=1}^{m}[a_{i.}+1]+\sum_{j=1}^{n}[a_{.j}+1]-[a_{..}]\right)_{1\times 1}. \end{array} \qquad (2.5)$$

The authors choose the conventional ℓ_p norms, $1\leq p<\infty$, as measures of discrepancy between R(A) and A, viz.,

$$\ell_p(R(A),A)=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|R(a_{ij})-a_{ij}|^p\right)^{1/p}, \qquad (2.6)$$

and proceed to demonstrate that, for fixed p, the set of all { 0,1 } - solutions to (2.5) which minimize (2.6) equals the set of all { 0,1 } - solutions to (2.5) which minimize the (linear) function

$$z_p=\sum_{i=1}^{m}\sum_{j=1}^{n}((1-a_{ij})^p-(a_{ij})^p)x_{ij}. \qquad (2.7)$$

In effect, z_p is a reformulation of the ℓ_p norm which is linear. In addition, the authors obtain an analogous but more complicated reformulation of the ℓ_∞ norm not reported here (Cox and Ernst 1982, pp. 428-429).

The controlled rounding problem thus is formulated as a linear programming problem which, a priori, may not have { 0,1 } - solutions. Transportation theory is then brought to bear to demonstrate that { 0,1 } - solutions do exist. and to produce such solutions, as follows.

The tabular array (2.5) represents a system of linear constraints in the x-variables of transportation type, and (2.7) represents a linear objective function in these variables whose minimizing solutions over the set of { 0,1 } values for the x's are sought. Because the totals entries of (2.5) are integer-valued, the triangular basis property of transportation arrays guarantees that solutions to (2.5) which optimize (2.7), if they exist, will be integer-valued (Dantzig 1963, pp. 303-305). This property is not lost when the authors further restrict (2.5) to sets of [0,1] - values for the x's by introducing the capacity constraints

$$0\leq x\leq 1 \qquad (2.8)$$

on the x's. Therefore, optimal controlled roundings of A, if they exist, correspond precisely to solutions of the transportation problem (2.5), (2.7)-(2.8). It remains to show that (optimal) controlled roundings of A always exist. Cox and Ernst accomplish this by explicit construction of a feasible solution to (2.5), (2.7)-(2.8). The desired optimal controlled roundings of A then are simply the integer-valued (i.e., { 0,1 } -valued) optimal solutions of the capacity constrained transportation problem (2.5), (2.7)-(2.8).

In summary, the authors have modelled an otherwise complicated combinatorial or integer programming problem completely as a feasible, capacity constrained transportation problem in the x-variables whose optimal solutions necessarily take on { 0,1 } - values. These solutions may be computed using standard transportation algorithms and computer software which is extremely efficient computationally (Glover, et al. 1974).

Extending these techniques, the authors also solve completely the stricter zero-restricted controlled rounding problem which, in addition to (2.2)-(2.3), requires that, for B=1,

$$|R(a)-a|<1. \qquad (2.9)$$

This is equivalent to the statement that, whenever an entry a of A is an integer, R(a)=a is required. (Note: In the general case, (2.9) becomes |R(a)-a|<B and multiples a of B are required to round to themselves.) The zero-restricted controlled rounding problem is modelled as a transportation problem by re-

placing the individual capacity constraints of (2.8) by the restriction

$$0 \leq x \leq 0 \qquad (2.10)$$

for precisely those x-values whose corresponding entries $\underline{a}$ of $\underline{A}$ are integers. Though posed artifically, the restriction (2.10) preserves the feasibility of the problem. Thus, the zero-restricted controlled rounding problem always can be solved. The practical value of a procedure which requires that integer values always round to themselves is obvious. Its importance is illustrated clearly in terms of controlled selection in Section 4.

3. APPLICATIONS OF CONTROLLED ROUNDING IN STATISTICS

In addition to improving the readability of data values and enhancing their utility for analysis, we may apply controlled rounding to several statistical problems. In each of these problems, a complete solution requires that non-integer values (respectively, non-integer multiples of $\underline{B}$) be replaced by integers (respectively, integer multiples of $\underline{B}$) with minimum overall distortion to the data.

An important application of controlled rounding is that of controlling statistical disclosure in tables of frequency counts. Data gathered from individual respondents by organizations such as the Census Bureau typically must be kept confidential. The publication of tables of frequency counts which directly or indirectly disclose small counts pose a threat to individual respondent confidentiality because small counts permit users to identify individual respondents and attribute characteristics to them. 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 a fixed integer base $\underline{B}$, e.g., B=5 or 10.

Conventional rounding fails to maintain additivity of detail items to totals and also may be undone to infer small counts. Nargundkar and Saveland (1972) address this problem for the case of one-way tables by rounding frequency count data <u>randomly</u> to a sufficiently large base $\underline{B}$. For example, with B=5, the value 2 rounds to 0 with probability 0.6 and to 5 with probability 0.4. Under their method, although the expected value of the sum of rounded values always equals the sum of the unrounded values, values rounded randomly may fail to add to rounded totals. Fellegi (1975) was able to achieve both randomness and additivity to totals for the case of one-way tables. In two-way tables, whereas a single controlled rounding would preserve the additivity condition (2.3), the individual entries could not be said to have been rounded randomly. However, as described in Section 4, it is always possible to choose a set of controlled roundings together with associated probabilities of selection such that the expectation of the rounded value for each entry equals the original unrounded value. Therefore, controlled rounding offers a viable alternative to random rounding as a technique for controlling statistical disclosure in frequency count tables; it preserves additivity and, subject to this condition, can be made to come as close as possible to maintaining conventional rounding throughout the table (Cox and Ernst 1982, pg. 429).

Controlled rounding may also be used to prevent statistical disclosure in microdata release. Suppose for the two-way table $\underline{A}$ that $\underline{a}_{ij}$ is the sum of $\underline{k}_{ij}$ quantities $\underline{a}_{ijk}$, $k=1, \ldots, k_{ij}$, each of which is to be rounded to a multiple of $\underline{B}$. Assume as before that B=1. We perform controlled rounding on the array $\underline{A}$, so that each $\underline{a}_{ij}$ is replaced by an adjacent integer $R(a_{ij})$. Next, for cell (i,j), we want to round some of the values $\underline{a}_{ijk}$ up and some down, so as to obtain a total $R(a_{ij})$ for the cell. To do this with a minimum of distortion to the cell entries $\underline{a}_{ijk}$, let $\underline{L}$ denote $R(a_{ij}) - \sum_k [a_{ijk}]$; we set the rounded value of $\underline{a}_{ijk}$ to $[a_{ijk}]+1$ for the $\underline{L}$ largest values among the $\underline{k}_{ij}$ quantities $a_{ijk} - [a_{ijk}]$, and to $[a_{ijk}]$ otherwise. As a possible example of application of this method, the quantities $\underline{a}_{ijk}$ might be personal incomes, to be rounded to the nearest multiple of $5000 before presentation so as to preclude identification of particular individuals and/or their exact incomes, with $\underline{i}$ and $\underline{j}$ 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 $\underline{k}_{ij}$ incomes for that category.

Controlled rounding is also applicable to "raked" two-way tables of counts as considered by Ireland and Kullback (1968). Given a two-way table $\underline{A}$ of integer counts $\underline{a}_{ij}$, we seek to construct a revised table of integer counts $\underline{A}$ whose row and column sums $\underline{a}_{i.}$ and $\underline{a}_{.j}$ have been predetermined, so as to minimize distortion to the original table. As an example, "i" might correspond to "race," "j" to "county," $\underline{a}_{ij}$ to the count of persons by race and county according to the 1970 Census, $\underline{a}_{i.}$ and $\underline{a}_{.j}$ to known numbers of persons in 1978 (based on demography, administrative records or otherwise), and $\underline{a}_{ij}$ to estimated numbers of persons in 1978. This procedure, <u>iterative proportional fitting</u>, is based upon repeated uniform multiplication of entries in each row and then each column to satisfy the desired marginals. In general, this procedure may result in non-integer values for the $\underline{a}_{ij}$'s, in which case controlled rounding with B=1 can be applied to $\underline{A}$ to obtain an integer array as desired.

For two-way stratified random sampling, one often uses Neyman allocation or a similar allocation scheme for choice of stratum sample sizes so as to minimize total sample size or a derived measure of cost subject to satisfying prespecified variance constraints. In such procedures, one obtains a two-way table of non-integer values which may likewise be optimally rounded to integers by means of controlled rounding.

In the area of raking, moreover, one will sometimes have a table $\underline{A}$ with so many zero entries that the iterative proportional fitting algorithm or an alternative fails to converge to a table $\underline{A}$ which has the desired new row and column sums $\underline{a}_{i.}$ and $\underline{a}_{.j}$. These zeros may be either sampling or structural zeros. This would happen because raking requires zeros to appear in

$\underline{A}$ whenever they appear in $\underline{A}$. Under these circumstances one may use a transportation theory approach to attempt to regain solvability, as follows. Use general transportation theory to construct a table $\underline{A}$, if it exists, with the desired marginals and structural zeros, for which the sum of the internal entries corresponding to the sampling zero entries of $\underline{A}$ is minimal. Then, if this sum is nonzero, subtract from $\underline{a}_{i.}$ and $\underline{a}_{.j}$ the entries of $\underline{A}$ corresponding to the sampling zeros and try raking $\underline{A}$ again with the modified marginals. If, perhaps after repeated applications of this procedure, the raking algorithm converges, all subtracted elements are added to the corresponding entries of the resulting raked table to produce the final raked array. Another approach to this problem is to "smooth" $\underline{A}$ by a Bayesian technique (Bishop, et al. 1975, Ch. 12).

4. CONTROLLED SELECTION

Consider a population stratified by two criteria of stratification, $\underline{m}$ rows and $\underline{n}$ columns, resulting in a two-way table of $\underline{mn}$ strata cells. A sample of size $\underline{r}$ is to be selected, with $\underline{s}_{ij}$ denoting the expected number of sample units in the $\underline{ij}$-th cell. Let $\underline{S}$ denote a tabular array with internal elements $(s_{ij})_{m \times n}$. We wish to limit the deviation from $\underline{s}_{ij}$ of the number of sample units in the $\underline{ij}$-th cell to less than one, and also 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 $\underline{N}_k$, $1 \le k \le \ell$, of zero-restricted controlled roundings of $\underline{S}$ (with B=1), together with associated probabilities of selection, where $\underline{n}_{ijk}$, the $\underline{ij}$-th entry for the $\underline{k}$-th array $1 \le i \le m$, $1 \le j \le n$, $1 \le k \le \ell$, is the number of sample units in the $\underline{ij}$-th cell if the $\underline{k}$-th array is selected, and satisfies $E(n_{ijk}|i,j)=s_{ij}$.

The above 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, Hartley, and Jessen (1960) developed a simple method for approaching the two-way stratification problem that we described above. However, their method does not in general satisfy all the requirements stated in the previous paragraph 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 also the much more complex three-dimensional problem. 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.

In this section we present an algorithm which employs the results described in Section 2 to completely solve the two-dimensional problem, together with an illustrative example, and outline how this construction may be modified to additionally satisfy restrictions on subarray totals. It is also shown in the complete paper, but omitted here due to lack of space, that the three-dimensional problem does not always have a solution.

4.1 The Algorithm

For purpose of notational simplicity, throughout this section and the Appendix, the $\underline{i}$-th row total, the $\underline{j}$-th column total and the grand total of a tabular (m+1)x(n+1) array $\underline{A}$ will be denoted respectively by $\underline{a}_{i(n+1)}$, $\underline{a}_{(m+1)j}$ and $\underline{a}_{(m+1)(n+1)}$, as alternatives to $\underline{a}_{i.}$, $\underline{a}_{.j}$ and $\underline{a}_{..}$.

We proceed to obtain a solution to the controlled selection problem $\underline{S}$ by recursively defining a sequence of arrays $\underline{N}_1=(n_{ij1})$, $N_2=(n_{ij2}),\ldots$, $N_\ell = (n_{ij\ell})$, and associated probabilities $p_1, \ldots, p_\ell$, satisfying

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

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

$$E(n_{ijk}|i,j) = \sum_{k=1}^{\ell} n_{ijk}\, p_k = s_{ij}, \quad 1 \le i \le m+1, \quad 1 \le j \le n+1. \tag{4.3}$$

To define $\underline{N}_k,\underline{p}_k$ we begin with the tabular array $A_k = (\overline{a}_{ijk})$. $A_1 = S$, while for $k \ge 1$, $\underline{A}_{k+1}$ will be defined in terms of $\underline{N}_k,\underline{p}_k$.

$\underline{N}_k$ is simply a zero-restricted controlled rounding of $\underline{A}_k$, while to define $\underline{p}_k$, 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

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

Now if $d_k = 0$, then $\sum_{i=1}^{k} p_i=1$ and we are done, that is, $N_1, \ldots, N_k$ together with the associated probabilities $p_1, \ldots, p_k$ provide a solution to the controlled selection problem. Otherwise, we define $\underline{A}_{k+1}$ by letting

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

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

4.2 An Example.

For this example

$$A_1 = S = \begin{array}{ccc|c} 0.4 & 2.0 & 0.0 & 2.4 \\ 1.2 & 0.0 & 1.0 & 2.2 \\ 0.2 & 0.0 & 0.0 & 0.2 \\ 1.2 & 0.4 & 0.2 & 1.8 \\ 1.0 & 0.6 & 0.2 & 1.8 \\ 0.0 & 0.4 & 0.4 & 0.8 \\ 0.0 & 0.2 & 0.4 & 0.6 \\ 0.0 & 0.0 & 0.2 & 0.2 \\ \hline 4.0 & 3.6 & 2.4 & 10.0, \end{array} \qquad N_1 = \begin{array}{ccc|c} 1 & 2 & 0 & 3 \\ 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 \\ 1 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 4 & 4 & 2 & 10, \end{array}$$

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

$$A_2 = \begin{array}{ccc|c} 0 & 2 & 0 & 2 \\ 1\,1/3 & 0 & 1 & 2\,1/3 \\ 1/3 & 0 & 0 & 1/3 \\ 1\,1/3 & 0 & 1/3 & 1\,2/3 \\ 1 & 1/3 & 1/3 & 1\,2/3 \\ 0 & 2/3 & 0 & 2/3 \\ 0 & 1/3 & 2/3 & 1 \\ 0 & 0 & 1/3 & 1/3 \\ \hline 4 & 3\,1/3 & 2\,2/3 & 10 \end{array}, \quad N_2 = \begin{array}{ccc|c} 0 & 2 & 0 & 2 \\ 2 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \hline 4 & 4 & 2 & 10 \end{array},$$

$d_2 = 2/3$, $p_2 = .2$,

$$A_3 = \begin{array}{ccc|c} 0 & 2 & 0 & 2 \\ 1 & 0 & 1 & 2 \\ 0.5 & 0 & 0 & 0.5 \\ 1.5 & 0 & 0 & 1.5 \\ 1 & 0 & 0.5 & 1.5 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0.5 & 0.5 \\ \hline 4 & 3 & 3 & 10 \end{array}, \quad N_3 = \begin{array}{ccc|c} 0 & 2 & 0 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ \hline 4 & 3 & 3 & 10 \end{array},$$

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

$$N_4 = A_4 = \begin{array}{ccc|c} 0 & 2 & 0 & 2 \\ 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 2 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ \hline 4 & 3 & 3 & 10 \end{array},$$

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

Note that the solution given is not unique.

4.3 Subarray Restrictions

Suppose that $\underline{S}$ is partitioned into a set $S_1, \ldots, S_s$ of subarrays. Let $\underline{A}_{kj}$, $\underline{N}_{kj}$ denote the subarrays of $\underline{A}_k$ and $\underline{N}_k$ respectively corresponding to $\underline{S}_j$. It is desired to obtain $\underline{N}_k$, $\underline{p}_k$, $k=1, \ldots, \ell$, satisfying not only (4.1)-(4.3), but also the additional condition that $\underline{N}_{kj}$ is a zero-restricted controlled rounding of $\underline{S}_j$ for all $\underline{k},\underline{j}$. Using preliminary results of Cox, subject to one limitation, this may be accomplished by proceeding exactly as in Section 4.1 with the following two exceptions. First, for each $\underline{k}$, $\underline{N}_k$ is now chosen so that not only is $\underline{N}_k$ a restricted controlled rounding of $\underline{A}_k$, but also $\underline{N}_{kj}$ is a restricted controlled rounding of $\underline{A}_{kj}$ for all $\underline{j}$. Secondly, $\underline{d}_k$ is now defined to maximize the absolute value for $N_k - A_k$ not only among all internal entries, and row and column totals, but also for the row and column subtotals for each subarray of the partition. In particular, this approach yields a formal procedure for solving controlled selection problems, such as the example of Goodman and Kish (1950, pg. 356), in which there is control on subadditivity in one dimension.

5. OPTIMAL METHOD FOR MAXIMIZING THE OVERLAP BETWEEN SURVEYS

5.1 Introduction

Consider a periodic survey with a multistage stratified design. At some point a redesign is undertaken in which the primary sampling units (PSU's) remain the same but the strata and selection probabilities change because of new data or changes in the design. The new sample PSU's may of course be selected independently of the initial PSU's. However, generally additional costs are incurred with each change of sample PSU. Consequently, it may be considered desirable that as many of the initial PSU's as possible be retained in the new sample, while strictly maintaining the requirements of probability sampling.

The first result for the problem of maximizing the expected number of retained PSU's was obtained by Keyfitz (1951). He presented an optimum procedure for one PSU per stratum designs in the special case when 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, 1971), 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 also is 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) had previously employed the transportation problem approach, but only with the restrictive assumptions considered by Keyfitz.

5.2 Notation

For any set $\underline{T}$ let card (T) denote the number of elements of $\underline{T}$. Let $\underline{S}$ denote a stratum in the new design, that is, $\underline{S}$ is a set of PSU's $S_1, \ldots, S_r$. Let $\underline{F}$, $\underline{G}$ be random sets denoting the set of PSU's in the initial and new samples respectively that are in $\underline{S}$, and let $C_1, \ldots, C_m$, $D_1, \ldots, D_n$ be the eligible choices for $\underline{F}$ and $\underline{G}$ respectively.

Finally, for $1 \le i \le m$, $1 \le j \le n$, $1 \le k \le r$, abbreviate

$$\begin{aligned} p_i &= P(F=C_i), \quad \pi_j = P(G=D_j). \\ p'_k &= P(S_k \in F), \quad \pi'_k = P(S_k \in G), \\ x_{ij} &= P(F=C_i \text{ and } G=D_j), \quad c_{ij} = \text{card } (D_j \cap C_i). \end{aligned}$$

5.3 The Procedure

We are now able to state the overlap problem more precisely. First note that each new stratum $\underline{S}$ represents a separate problem. For each such $\underline{S}$ we seek to maximize

$$E[\text{card } (G \cap F)],$$

subject to

$$P(F=C_i) = p_i, \quad 1 \le i \le m,$$

and

$$P(G=D_j) = \pi_j, \quad 1 \le j \le n.$$

However, this is equivalent to solving the following transportation problem. Find $x_{ij} \geq 0$ which maximize

$$E\ [\text{card}\ (G \cap F)] = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}, \qquad (5.1)$$

subject to

$$\sum_{j=1}^{n} x_{ij} = p_i,\ 1 \leq i \leq m, \qquad (5.2)$$

$$\sum_{i=1}^{m} x_{ij} = \pi_j,\ 1 \leq j \leq n. \qquad (5.3)$$

Once the optimal $\underline{x}_{ij}$'s have been obtained, the conditional selection probabilities given that C_i was selected originally are simply

$$P(G = D_j | F = C_i) = \frac{x_{ij}}{p_i},\ 1 \leq j \leq n.$$

5.4 An Example

In this example the intitial and new designs are both one PSU per stratum, and the sampling was done independently in each initial stratum. We have r=5 and the following probabilities:

1. p'_i's and π'_i's

i	1	2	3	4	5
p'_i	.50	.06	.04	.60	.10
π'_i	.40	.15	.05	.30	.10

We are also given that $\underline{S}_1$, $\underline{S}_2$, $\underline{S}_3$ were all in a single initial stratum and $\underline{S}_4$, $\underline{S}_5$ were in a second initial stratum. The $\underline{C}_i$'s and $\underline{D}_i$'s are then labeled:

$C_i = D_i = \{S_i\}$, $1 \leq i \leq 5$,
$C_6 = \{S_1, S_4\}$, $C_7 = \{S_1, S_5\}$, $C_8 = \{S_2, S_4\}$,
$C_9 = \{S_2, S_5\}$, $C_{10} = \{S_3, S_4\}$, $C_{11} = \{S_3, S_5\}$, $C_{12} = \emptyset$.

The p_i's and π_i's are then as in Table 2.

2. p_i's and π_i's

i	1	2	3	4	5	6	7	8	9	10	11	12
p_i	.15	.018	.012	.24	.04	.30	.05	.036	.006	.024	.004	.12
π_i	.40	.15	.05	.30	.10							

To solve the problem we find x_{ij}'s ≥ 0 which maximize (5.1) subject to (5.2) and (5.3), with m=12, n=5, and the following $\underline{c}_{ij}$'s:

3. c_{ij}'s

i \ j	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

An optimal set of $\underline{x}_{ij}$'s is given by Table 4, and the corresponding maximum value of the objective function is .880.

4. x_{ij}'s which maximize (5.1)

i \ j	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

Remark 5.1: Although the previous literature has only been concerned with the problem of maximizing the expected number of retained PSU's, there are certain situations where it is desired to minimize this quantity instead. This could occur, for example, if an entirely different set of ultimate sample units is wanted in the new sample, but retention of some small PSU's would also require the retention of some ultimate sample units. The procedure for solving the minimization problem is essentially identical to the maximization problem, except now (5.1) is minimized subject to (5.2) and (5.3).

Remark 5.2: 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. However, if the initial sample was not chosen independently from stratum to stratum, this information may not be available. In Ernst (1982) an alternative overlap procedure is presented which only requires knowledge of the joint selection probabilities in the inital sample for sets of PSU's that are in the same initial and new strata, and which, in certain circumstances, is optimal among all procedures which require only this amount of information. This alternative procedure formulates the overlap problem as a linear programming problem, but not a transportation problem.