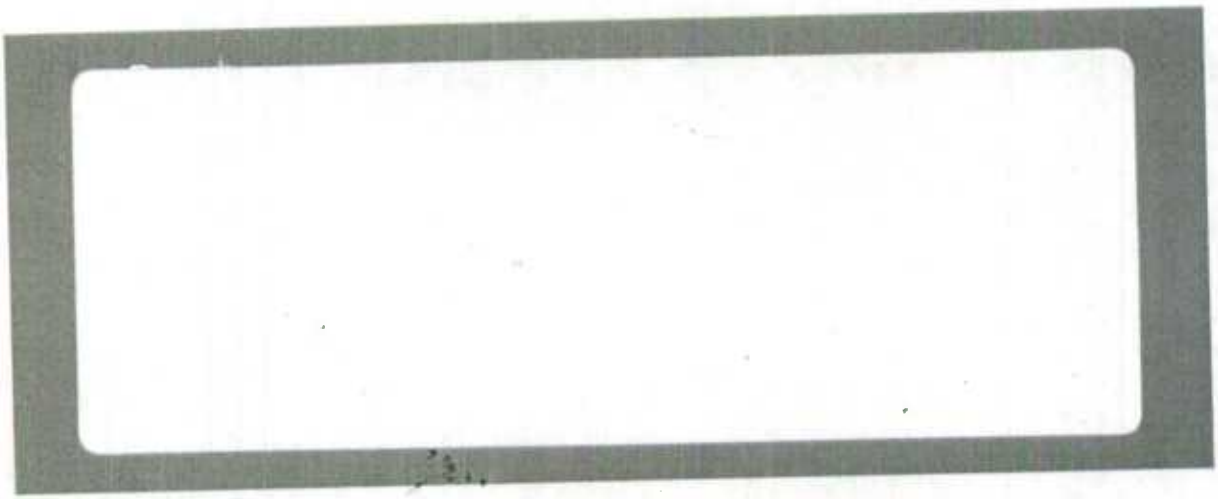

Methodology Branch

Household Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

THE USE OF THE TRANSPORTATION PROBLEM IN CO-ORDINATING THE SELECTION OF SAMPLES FOR BUSINESS SURVEYS

HSMD-2005-006E

Lenka, Mach, Philip T. Reiss, I. Şchiopu-Kratina

Household Survey Methods Division
Statistics Canada

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada.

THE USE OF THE TRANSPORTATION PROBLEM IN COORDINATING THE SELECTION OF SAMPLES FOR BUSINESS SURVEYS

Lenka Mach, Philip T. Reiss, Ioana Şchiopu-Kratina*

ABSTRACT

We view the problem of maximizing or minimizing the expected overlap of two surveys as a transportation problem (TP) and give simple selection algorithms for solving it. Although we use linear programming (LP) to formulate our problem and invoke results and techniques from optimization theory, the main purpose is to justify our selection algorithms and demonstrate the optimal properties of the solutions they generate. Our methods are relevant to surveys, such as typical government business surveys, with stratified frames and simple random sampling without replacement (SRSWOR) applied independently in all strata. We show that our coordinated selection preserves both sample designs and optimizes the expected overlap of samples. Our selection algorithms proceed in two steps, the details of which are determined by the specific problem. For example, when updating a population only for births, the first step consists of deciding on the number of births in the new sample through a random selection from a hypergeometric distribution. In the second step, the chosen number of births are selected by SRSWOR from the pool of new units. We then randomly select or deselect "old" units in order to obtain the required sample size. The procedure guarantees that the design is SRSWOR for the updated survey and that the expected number of selected births is minimal. Consequently, the expected cost of first contacts, which could otherwise be quite high, is minimized. Through a simulation study, we compare our method with methods based on the assignment of permanent random numbers (PRN) to all units in the population. Our mathematical approach has the advantage that we can theoretically prove the properties of our solutions. We show that the variance of our solutions cannot be improved within a class of comparable solutions.

---

* Lenka Mach, SSMD, Statistics Canada . Philip T. Reiss, Department of Biostatistics, Mailman School of Public Health, Columbia University. Ioana Schiopu-Kratina, HSMD, Statistics Canada.

L'UTILISATION DU PROBLÈME DE TRANSPORT DANS LA COORDINATION DE LA SÉLECTION DES ÉCHANTILLONS POUR DES ENQUÊTES AUPRÈS DES ENTREPRISES

Lenka Mach, Philip T. Reiss, Ioana Şchiopu-Kratina†

RÉSUMÉ

Nous envisageons le problème de la maximisation ou de la minimisation du chevauchement espéré de deux enquêtes comme un problème de transport (PT) et nous donnons des algorithmes de sélection simples pour le résoudre. Bien que nous utilisions la programmation linéaire (PL) pour formuler notre problème et que nous invoquions des résultats et les techniques de la théorie de l'optimisation, l'objectif principal est de justifier nos algorithmes de sélection et de démontrer les propriétés optimales des solutions qu'elles génèrent. Nos méthodes sont applicables aux enquêtes, telles que des enquêtes gouvernementales typiques auprès des entreprises, réalisées au moyen de bases de sondage stratifiées, selon une méthode d'échantillonnage aléatoire simple sans remise (EASSR) appliquée indépendamment dans chaque strate. Nous montrons que notre sélection coordonnée permet de respecter les deux plans de sondage en optimisant le chevauchement espéré des échantillons. Nos algorithmes de sélection comportent deux étapes, dont les détails sont déterminés par le problème à résoudre. Ainsi, si l'on met à jour une population uniquement pour tenir compte des nouvelles entreprises, la première étape consiste à décider du nombre de nouvelles entreprises dans le nouvel échantillon au moyen d'une sélection aléatoire à partir d'une distribution hypergéométrique. À la seconde étape, le nombre choisi de nouvelles entreprises est sélectionné par EASSR parmi les nouvelles unités. Ensuite nous sélectionnons ou désélectionnons aléatoirement parmi les « anciennes » unités afin d'obtenir la taille d'échantillon requise. La procédure garantit que le plan de sondage est EASSR pour l'enquête mise à jour et que le nombre espéré de nouvelles entreprises sélectionnées est minimal.

Par conséquent, le coût prévu des premiers contacts, qui pourrait autrement être assez élevé, est réduit au minimum. Au moyen d'une étude de simulation, nous comparons notre méthode aux méthodes fondées sur l'attribution d'un numéro aléatoire permanent (NAP) à chaque unitée de la population. Notre approche mathématique permet de prouver théoriquement les propriétés de nos solutions. Nous montrons que la variance de nos solutions ne peut pas être améliorée à l'intérieur d'une classe de solutions comparables.

---

† † Lenka Mach, SSMD, Statistics Canada . Philip T. Reiss, Department of Biostatistics, Mailman School of Public Health, Columbia University. Ioana Schiopu-Kratina, HSMD, Statistics Canada.

1. INTRODUCTION

We discuss in this article some problems related to the coordination of samples selected for business surveys. For such surveys, SRSWOR samples of fixed size are selected independently from each stratum. We are interested in the problem of maximizing or minimizing the expected overlap of samples while attaining the inclusion probabilities of all orders (preserving the entire design) for all surveys considered. This last requirement is sometimes referred to in the literature as "integration of surveys," a term coined by Mitra and Pathak (1984), which is formally expressed by (TP) in Section 2. Thus we formulate our problem as a transportation problem (also abbreviated TP) and present simple algorithms for solving it. Our method is adequate for large samples, according to the classification given in the comprehensive article by Ernst (1999). We compare our method with three PRN methods: sequential SRSWOR for positive and negative coordination, the method of collocated samples (see Ohlsson (1995)), and the PRN method with full-stratum rotation.

There is a vast body of literature on coordination of samples for several surveys, starting with the pioneering work of Keyfitz (1951). For detailed descriptions of methods, we refer to Ernst (1999) and Ohlsson (1995). Even though we formulate each sample coordination problem as a transportation problem and show that a solution given by the Northwest Corner Rule (NWCR) is adequate, our approach is not technically an LP approach. We obtain an analytical form of the solution (see Proposition 1) and proceed with a two-stage selection algorithm (Theorem 1). Therefore, our approach is not directly comparable to other LP applications to integration of sample surveys (e.g. Causey, Cox and Ernst (1985), Ernst and Ikeda (1994)).

One of the first major applications of the transportation problem to survey sampling was developed by Raj (1956) and can be formulated as follows. Assume that two surveys are carried out on the same population concentrated in centers that are geographically far apart. The surveys must be integrated and the cost of transportation has to be minimized. This translates into the requirement that a maximum number of settlements be visited once to gather information for both surveys.

Another frequently encountered situation is the updating of the frame of a business survey. New units (births) must be added to the frame, while others, which are found inactive (deaths), must be discarded. The updates could also be the result of changes in the definition of strata, which requires a reclassification of units. In each new stratum, which we assume is an estimation

domain, one would like to retain as many units as possible from the previous sample so as to obtain comparable estimates over time. One is also interested in reducing the cost of first contacts, which means maximizing the expected number of "non-births" in the sample.

The minimization of the expected overlap is motivated by the following situation, often encountered in practice. Several surveys are carried out on the same population. Some units may be selected in more than one survey and the burden of providing a lot of information may lead to nonresponse and a general deterioration in the quality of the collected data. We would therefore like to minimize the "cost" of responding to several surveys. Mathematically, these problems can be treated in a similar fashion. It is interesting to note that in our formulation the minimization problem is not the mathematical dual of the maximization problem, since the objective functions (the functions we want to optimize) are different.

Once we define the meaning of "overlap" or "response burden", we can write the objective function and formalize each problem as a transportation problem (see Section 2), which can be solved using LP techniques. This LP problem may have a large number of variables. Even listing all possible samples in one survey is a difficult task. Attempts at reducing the problem for general designs have been given in, e.g., Aragon and Pathak (1990) and Ernst and Ikeda (1994). In our approach, because of the symmetry (exchangeability) of the SRSWOR design, we consider configurations (groups of samples) instead of the individual samples. To illustrate this, assume that before the updates to a stratum, we selected n units out of N available. Instead of considering the list of $\binom{N}{n}$ possible samples, we look at configurations defined by the number of dead units observed in the sample of n units after the updates. There are at most $n+1$ such configurations and each has an associated probability given by a hypergeometric distribution. We show that the problem of coordinating the selection of samples for two different surveys reduces to optimizing an overlap function defined at the level of configurations of samples.

The NWCR is a very efficient algorithm for finding a solution to a transportation problem. This algorithm gives an optimum value for the objective function if this function has the appropriate Monge property (in the literature inconsistently referred to as either our Definition 1 or 2). In Proposition 1, we calculate the joint distribution given by the NWCR. The NWCR cumulative distribution is the minimum of the cumulative distributions of the marginals, which are hypergeometric distributions (see Proposition 1). The actual selection proceeds in two stages. For a simultaneous selection for two surveys, we first perform a probability-proportional-to-size

selection from the NWCR solution to determine the actual number of common units in both samples. We then randomly select these units from the common pool and select additional units to complete each of the two samples. The same mechanism is adapted to updating a sample selected on a previous occasion.

One advantage of our method (henceforth called the NWCR method) over others is that we solve well-defined mathematical problems and so our solutions have verifiable properties. We prove that the surveys considered are integrated and that the expected overlap has the optimum desired properties, including minimum variance when minimizing the overlap of two surveys. We also show that all solutions to the maximization problem have the same variance. This means that one cannot find another solution to the maximization problem that would have a smaller variance than the NWCR solution. Furthermore, the NWCR method can be applied on an ad hoc basis, i.e. to surveys that are not necessarily run from a business register with PRNs assigned to all units. The algorithm that we propose is easy to apply and lists of samples or even configurations of samples are not needed. On the other hand, it is not clear how the NWCR method for maximizing the overlap can be adapted if the classification of units (for example, their industry or size code) is updated on the frame. The NWCR method for minimizing the overlap does not have a straightforward generalization to more than two surveys. In this latter situation, it might be worthwhile to investigate a bona fide LP application, since the number of variables (configurations) does not appear to be large. This would also give us some flexibility in the choice of the objective function when dealing with several surveys which may incur different response burden on respondents, as is provided by the software Salomon (see Rivière (2001)).

This article is organized as follows. Section 2 presents the mathematical formulation of the problem as a transportation problem and gives an analytical solution based on the NWCR (Proposition 1). Section 3 presents the NWCR solution to the problem of maximizing the overlap of surveys (Theorem 1). Section 4 is dedicated to the problem of minimizing the overlap of two surveys. In Section 5, we present the results of a simulation study which compares the NWCR method to the sequential SRSWOR for positive and negative coordination, the method of collocated samples and the PRN method with full-stratum rotation. We illustrate our definitions and methods on Example 1, which appears first in Section 3.1.

2. THE TRANSPORTATION PROBLEM FOR TWO SURVEYS

In this article, we consider two SRSWOR designs used for two surveys of a finite population. This covers the situation of two distinct surveys as well as two different selections for the same survey when the population of a stratum has been updated for deaths and births. Initially, we have N units in the stratum population. After the updates, there are N' units in the stratum. If D denotes the number of *deaths*, i.e. units that belong only to the initial (first) population, B the number of *births*, i.e. units only in the updated (second) population, and C the number of units that belong to both populations, then $N' = N - D + B = C + B$. For the first survey, an SRSWOR of n units is selected from the population of size N, and, for the second survey, an SRSWOR of n' from N'. We denote

by S and S' the set of all possible samples in the first and second survey, respectively. The overlap of the samples $s \in S$ and $s' \in S'$, denoted by $o(s, s')$, is the number of units that the two samples have in common. Our ultimate objective is to find a joint distribution for all pairs of samples (s, s') that will maximize or minimize their expected overlap, given the marginal distributions in each of the two surveys. Consider the $\binom{N}{n} \times \binom{N'}{n'}$ matrix whose (i, j) entry is $o(s_i, s_j')$, where $(s_1, ..., s_{\binom{N}{n}})$ is an ordering of the samples from S which groups them into super-rows in a manner to be described below; and likewise the ordering $(s_1', ..., s_{\binom{N'}{n'}}')$ groups samples from S' into super-columns. We calculate from the two designs the probability distributions $P = (p_i)_{i=1,...,I}$ and $Q = (q_j)_{j=1,...,m}$ of super-rows and super-columns, respectively. These super-rows and super-columns form a matrix of blocks. Depending on our goal, certain pairs of samples within a given block will give an optimal overlap value: maximum overlap within the block in case of positive coordination, and minimal overlap for negative coordination. To encompass either case, we refer to this optimal value as the *block optimum*. The block optima constitute the coefficients of the linear function (the objective function) which we optimize at the first stage of our algorithm. The joint distribution (the probabilities assigned to blocks) is represented by a table whose rows and columns correspond to groups of samples from the old and new designs, respectively. We therefore seek a joint distribution $X = (x_{ij})_{i,j \geq 1}$ with P and Q as marginals, i.e. a solution to the transportation problem (TP), which optimizes an objective function to be defined subsequently:

(TP) $$x_{ij} \geq 0, \quad \sum_{j=1}^{m} x_{ij} = p_i, 1 \leq i \leq n, \quad \sum_{i=1}^{n} x_{ij} = q_j, 1 \leq j \leq m, \quad \sum_{i=1}^{n} p_i = \sum_{j=1}^{m} q_j = 1.$$

A solution to (TP) always exists (e.g., we can define $x_{ij} = p_i \times q_j, i = 1,...,n, j = 1,...,m$).

The NWCR gives a solution $X = (x_{ij})_{i,j \geq 1}$ to the TP and is a greedy algorithm, i.e., each x_{ij} is the maximum possible probability (mass) given the marginals p_i, q_j and the mass assigned to "previous" blocks, $x_{i'j'}$, with $i' \leq i$, $j' \leq j$ and $(i', j') \neq (i, j)$. Hoffman (1985) explains the NWCR algorithm as follows.

(NWCR): We set $x_{i0} = 0$, $x_{0j} = 0$ and, if x_{rs} has been defined for all pairs (r, s), $r \leq i$, $s \leq j$, $(r,s) \neq (i,j)$, then: $x_{ij} = \min\{p_i - \sum_{s=0}^{j-1} x_{is}, \; q_j - \sum_{r=0}^{i-1} x_{rj}\}$, $i, j \geq 1$.

The above expression represents the greatest assignable mass given the marginals, since it represents the as-yet-unassigned mass of that row or column, whichever is less. For a more explicit construction, see the Appendix. See also pp. 248-250 of Arthanari and Dodge (1981). ■

Let $P(i) = \sum_{k=1}^{i} p_k$, $Q(j) = \sum_{k=1}^{j} q_k$, $X(i, j) = \sum_{k=1}^{i} \sum_{l=1}^{j} x_{kl}$ be the cumulative distributions.

For any joint distribution X and each (i, j), $X(i, j) \leq \min\{P(i), Q(j)\}$ (see also (2.5) of Hoffman (1985)). The following proposition asserts that equality holds if X is given by the NWCR.

Proposition 1: If X is given by the NWCR, then $X(i, j) = \min\{P(i), Q(j)\}$, $i, j \geq 1$.

Proof:

Case 1: $x_{ij} > 0$. Without loss of generality, suppose that $x_{ij} = p_i - \sum_{s=0}^{j-1} x_{is}$. If $X(i, j) < P(i)$ then $\sum_{s=1}^{j} x_{ks} < p_k$ for some $k < i$, since $\sum_{s=1}^{j} x_{is} = p_i$. Given such a k, there exists $m > j$ with $x_{km} > 0$; but this inequality, taken together with $x_{ij} > 0$, contradicts $x_{kj} = \min\{p_k - \sum_{s=0}^{j-1} x_{ks}, q_j - \sum_{r=0}^{i-1} x_{rj}\}$.

Thus $X(i, j) \geq P(i) \geq \min\{P(i), Q(j)\}$ and so $X(i, j) = \min\{P(i), Q(j)\}$.

Case 2: $x_{ij} = 0$. By (NWCR), p_i (or q_j) has been distributed before column j (row i). Assume the distribution of q_j was completed in row $k < i$. Then, as in Case 1 above, we have $X(k, j) = Q(j) \leq X(i, j)$ (since $k < i$), so $X(i, j) \geq \min\{P(i), Q(j)\}$, which proves the equality. ■

Of particular interest to us are matrices of overlaps $R = (\rho_{ij})_{i,j\geq 1}$ satisfying one or the other of the following definitions, which are taken from Ross (1983).

Definition 1: The matrix $(\rho_{ij})_{i,j\geq 1}$ is supermodular (SM) if it satisfies:

(SM) $\rho_{ij} + \rho_{rs} \geq \rho_{is} + \rho_{rj}$, for all $1 \leq r \leq i, 1 \leq s \leq j$. ■

Definition 2: The matrix $(\rho_{ij})_{i,j\geq 1}$ is submodular (sM) if it satisfies:

(sM) $\rho_{ij} + \rho_{rs} \leq \rho_{is} + \rho_{rj}$, for all $1 \leq r \leq i, 1 \leq s \leq j$. ■

For a matrix of overlaps and a solution X to (TP), we define the expected overlap by $E_X(\rho) = \sum_{i,j} \rho_{ij} x_{ij}$ and the variance of the overlap by $V_X(\rho) = \sum_{i,j} \left[\rho_{ij} - E_X(\rho)\right]^2 x_{ij} = \left(\sum_{i,j} \rho_{ij}^2 x_{ij}\right) - \left[E_X(\rho)\right]^2 = E_X(\rho^2) - \left[E_X(\rho)\right]^2$.

It is well known that, if the matrix of overlaps is SM, than the solution X_0 obtained by the NWCR maximizes the expected overlap, i.e., $E_{X_0}(\rho) = \max \sum_{i,j} \rho_{ij} x_{ij}$, where the maximum is taken over all solutions X to (TP) (see Result 5.6.2 of Arthanari and Dodge (1981)). A similar result holds for sM matrices. In Proposition 2, we present a more general result for SM matrices, which also holds for sM matrices.

First, we introduce some notation. Define $\Delta_{kl}(i, j)$ to be the $n \times m$ matrix with only four nonzero entries, placed as follows at the vertices of a rectangle: 1 is placed in each of the cells (i, j) and $(i+k, j+\ell)$, while -1 is placed in each of the cells $(i, j+\ell)$, $(i+k, j)$. We call a constant multiple of such a matrix an *elementary difference matrix*. We write $E_{\Delta_{kl}(i,j)}(\rho) = \rho_{ij} + \rho_{i+k,j+\ell} - \rho_{i+k,j} - \rho_{i,j+\ell}$ and note that $E_{\Delta_{kl}(i,j)}(\rho) \geq 0$ if the matrix of overlaps is SM, and is negative if the matrix of overlaps is sM. The next proposition states that the difference between the NWCR solution and any other solution to (TP) can be written as a finite sum of elementary difference matrices. The proof is similar to that of Result 5.6.2 of Arthanari and Dodge (1981) and is given in the Appendix.

Proposition 2: If X_0 is given by the NWCR and X is any solution to (TP), we have

$X_0 = X + \sum_{\gamma=1}^{G} \alpha_\gamma \Delta_\gamma$ for some positive numbers $\alpha_1, \ldots, \alpha_G$, where $\Delta_\gamma = \Delta_{kl}(i, j)$, for some i, j, k, ℓ as above. ■

Corollary 1: If the matrix of overlaps is SM (sM) then $E_{X_0}(\rho)$ is maximal (minimal).

Proof: Since the matrix of overlaps is SM (sM), we have $E_{\Delta_\gamma}(\rho) \geq 0$ ($E_{\Delta_\gamma}(\rho) \leq 0$), $\gamma = 1,\ldots,G$. The maximality (minimality) of $E_{X_0}(\rho)$ follows directly by applying Proposition 2 and observing that $E_{X_0} = E_X + \sum_{\gamma=1}^{G} \alpha_\gamma E_{\Delta_\gamma}$. ■

For an illustration of the application of the NWCR, see Table 2 in Example 1 below.

Proposition 3: A matrix R is SM if and only if (SM') holds:

(SM') $$\rho_{i,j} + \rho_{i+1,j+1} \geq \rho_{i,j+1} + \rho_{i+1,j}, \; i = 1,\ldots,n-1, \; j = 1,\ldots,m-1.$$

The matrix R is sM if and only (sM') holds:

(sM') $$\rho_{i,j} + \rho_{i+1,j+1} \leq \rho_{i,j+1} + \rho_{i+1,j}, \; i = 1,\ldots,n-1, \; j = 1,\ldots,m-1.$$

Proof: We will prove the first statement only, since the second can be proved in a similar fashion. It is obvious that (SM) implies (SM'). We show the reverse implication by induction on the "perimeter of the rectangle." Without loss of generality, consider a rectangle with corners (i, j), (i, j+ℓ+1), (i+ k, j+ℓ+1), (i + k, j), i.e., of perimeter 2 k(ℓ+1). By the induction hypothesis applied to rectangles with perimeters 2kℓ and 2k, we have $\rho_{i,j} + \rho_{i+k,j+\ell} \geq \rho_{i,j+\ell} + \rho_{i+k,j}$ and $\rho_{i,j+\ell} + \rho_{i+k,j+\ell+1} \geq \rho_{i,j+\ell+1} + \rho_{i+k,j+\ell}$. We now add these inequalities and obtain $\rho_{ij} + \rho_{i+k,j+\ell+1} \geq \rho_{i+k,j} + \rho_{i,j+\ell+1}$, which proves (SM) for the rectangle with perimeter 2k(ℓ+1). ■

Lemma 1: If c_i, i = 1,...,n, and c_j', j = 1,...,m, are two decreasing sequences of positive numbers, then the $n \times m$ matrix with entries $\rho_{ij} = \min\{c_i, c_j'\}$ is SM.

Proof: By Proposition 3, it suffices to prove (SM'). Because the two sequences of positive numbers are decreasing, we can express the overlaps on the right hand side of (SM') as: $\rho_{i,j+1}$= $\min\{\rho_{ij}, c_{j+1}'\}$ and $\rho_{i+1,j} = \min\{\rho_{ij}, c_{i+1}\}$. The right hand side of (SM') can be now written as: $\min\{\rho_{ij}, \min\{c_{i+1}, c_{j+1}'\}\} + \min\{\rho_{ij}, \max\{c_{i+1}, c_{j+1}'\}\}$, since one number of the pair c_{i+1}, c_{j+1}' is the minimum and the other is the maximum of the two. (SM') then becomes ρ_{ij} + $\min\{\rho_{ij}, c_{i+1}, c_{j+1}'\} \geq \min\{\rho_{ij}, \min\{c_{i+1}, c_{j+1}'\}\} + \min\{\rho_{ij}, \max\{c_{i+1}, c_{j+1}'\}\}$, which

reduces to $\rho_{ij} \geq \min\{\rho_{ij}, \max\{c_{i+1}, c_{j+1}'\}\}$. We have equality if and only if $\rho_{ij} = \max\{c_{i+1}, c_{j+1}'\}$. ■

3. MAXIMIZING THE EXPECTED OVERLAP

3.1 Notation and Definitions

Consider SRSWOR selections from a stratum common to two surveys and recall the notation introduced in Section 2. We denote by C the set of units common to the two populations. For $s \in S$, $c(s)$ (or c) denotes the number of units from the set C (non-deaths). Similarly, $c'(s')$ (or c') represents the number of non-births in s', $s' \in S'$. We group the "old" samples s with the same value of c into super-rows, and order the super-rows in decreasing order of c. We likewise group the "new" samples s' into super-columns and order the super-columns in decreasing order of c'. Due to the SRSWOR design, the marginal probability of a super-row with a specified value of c is given by the hypergeometric distribution $p(c;N,n,C) = \binom{C}{c}\binom{D}{n-c}/\binom{N}{n}$. Similarly, the marginal probability of a super-column with a specified value of c' is given by the hypergeometric distribution $q(c';N',n',C) = \binom{C'}{c'}\binom{B}{n'-c'}/\binom{N'}{n'}$.

Remark 1: The number of units from C found in the first sample may be anywhere from $c_m = \max(n\text{-}D,0)$ to $c_M = \min(C, n)$; and the number in the second sample, from $c_m' = \max(n'\text{-}B, 0)$ to $c_M' = \min(C, n')$. With the decreasing-order arrangement, the i^{th} of the $(c_M - c_m + 1)$ super-rows consists of samples selected from the first population that contain exactly $(c_M - i + 1)$ units from C, while the j^{th} of the $(c_M' - c_m' + 1)$ super-columns consists of new samples with $(c_M' - j + 1)$ units from C.

The interior of the matrix defined by super-rows and super-columns consists of blocks. The block identified by (c, c') consists of all pairs of samples (s, s') with $c(s) = c$ and $c'(s') = c'$, and we define its *block optimum* by $\rho(c, c') = \min\{c, c'\}$. The block optimum represents the largest possible overlap of a pair of samples within the block. We note that the matrix of block optima is SM by Lemma 1.

Example 1: Let $N = 6$, $n = 3$, $D = 3$ and $B = 2$. Then $N' = 5$, from which we select $n' = 4$ units by SRSWOR. In Tables 1 and 2, the left margin refers to the original survey while the top margin

refers to the new survey. The margins of Table 1 are the possible values of c and c', and the entries give the resulting block optimum, $\rho(c, c') = \min\{c, c'\}$. The margins of Table 2 contain the hypergeometric probabilities corresponding to the super-rows and super-columns, $p(c)$ and $q(c')$, and the entries are the joint probabilities assigned by the NWCR.

Table 1. Matrix of block optima

$c \backslash c'$	3	2
3	3	2
2	2	2
1	1	1
0	0	0

Table 2. Matrix of block probabilities

$p(c) \backslash q(c')$	8/20	12/20
1/20	1/20	0
9/20	7/20	1/10
9/20	0	9/20
1/20	0	1/20

There are four super-rows and two super-columns in each table. Thus, there are only eight blocks for which we need joint probabilities. By contrast, even for the small population and sample sizes in Example 1 there are 100 possible pairs of samples (s, s'). In Table 2, the joint probabilities represent a solution to (TP) which can be obtained by directly applying (NWCR) or using Proposition 1. ■

3. 2 Simultaneous Selection for Maximization

Construction of a joint density for maximization. Given two SRSWOR surveys, a solution to (TP) which maximizes the expected overlap can be obtained by a two-stage procedure:

1. Form a table whose marginals are the super-row and super-column probabilities arranged in descending order of c and c' as above, and derive the joint distribution of the blocks by the NWCR.
2. Within each block, divide the mass equally among those individual pairs of samples (s, s') which attain the largest possible overlap for that block.

To clarify the second step: The super-row consisting of all old samples with c units from C and the super-column consisting of all new samples with c' units from C form a block (c, c') containing $\binom{C}{c}\binom{D}{n-c}\binom{C}{c'}\binom{B}{n'-c'}$ pairs of samples (s, s'). Of these, there are $\binom{D}{n-c}\binom{C}{\min(c,c')}\binom{C-\min(c,c')}{\max(c,c')-\min(c,c')}\binom{B}{n'-c'} = \binom{D}{n-c}\binom{C}{\max(c,c')}\binom{\max(c,c')}{\min(c,c')}\binom{B}{n'-c'}$ pairs (s, s') sharing $\min(c, c')$ units and thus attaining the largest possible overlap within the block, i.e.,

$o(s,s') = \rho(c,c') = \min(c,c')$. The proposed conditional-on-block distribution divides the mass $x_{c,c'}$ equally among these latter pairs.

Theorem 1: (a) The joint density X_0 defined on the set of pairs of samples (s, s') by the two-stage procedure above has SRSWOR marginals.

(b) X_0 has the maximum expected overlap within the set of joint densities with the given marginals.

(c) All joint densities which satisfy (a) and (b) have the same variance.

Proof: (a) We show that the probability of selecting any sample $s \in S$ is $1/\binom{N}{n}$. Consider a sample s in super-row i. In super-columns with $c_j' \le c_i$, there are exactly $\binom{c_i}{c_j'}\binom{B}{n'-c_j'}$ samples s' that share c_j' units with s. In super-columns with $c_j' > c_i$, there are exactly $\binom{C-c_i}{c_j'-c_i}\binom{B}{n'-c_j'}$ samples s' that share c_i units with s. The probability $p(s)$ of selecting a sample s in super-row i is the sum of all probabilities assigned in step 2 above on row s and across all super-columns:

$$p(s) = \sum_{c_j' \le c_i} \left(\sum_{\{s' \in j:\, o(s,s')=c_j'\}} \frac{x_{ij}}{\binom{D}{n-c_i}\binom{C}{c_i}\binom{c_i}{c_j'}\binom{B}{n'-c_j'}} \right) + \sum_{c_j' > c_i} \left(\sum_{\{s' \in j:\, o(s,s')=c_i\}} \frac{x_{ij}}{\binom{D}{n-c_i}\binom{C}{c_i}\binom{C-c_i}{c_j'-c_i}\binom{B}{n'-c_j'}} \right)$$

$$= \sum_{j=1}^{m} \frac{x_{ij}}{\binom{D}{n-c_i}\binom{C}{c_i}} = \frac{p_i}{\binom{D}{n-c_i}\binom{C}{c_i}} = \frac{1}{\binom{D}{n-c_i}\binom{C}{c_i}} \times \frac{\binom{C}{c_i}\binom{D}{n-c_i}}{\binom{N}{n}} = \frac{1}{\binom{N}{n}},$$

where the second equality follows from (i) and (ii) above. Similarly, it can be shown that the probability of selecting any sample $s' \in S'$ is $1/\binom{N'}{n'}$.

(b) To show optimality, it suffices to apply Corollary 1 at the level of blocks because the entire mass of a block is divided only among those pairs (s, s') that have the maximum possible overlap. The matrix of block optima is SM (Lemma 1), so the NWCR gives the maximum expected overlap by Corollary 1.

(c) To prove the result about the variance, we again use Proposition 2. If X represents a solution with the maximum expected overlap and X_0 is the NWCR solution, it suffices to show that $E_{X_0}(\rho^2) = E_X(\rho^2)$. Since the matrix of block optima is SM, we must have $E_{\Delta_\gamma}(\rho) = 0$, $\gamma = 1,\ldots,G$, in the representation given in Proposition 2. By Lemma A.1 in the

Appendix, $E_{\Delta_\gamma}(\rho^2)=0, \gamma=1,...,G$. We now apply Proposition 2 to the matrix $(\rho_{ij}^2)_{i,j\geq 1}$ and obtain $E_{X_0}(\rho^2)=E_X(\rho^2)$. ■

Example 2: We refer to Example 1 and expand below the block corresponding to cell (2, 2). The "marginals" are now the samples. We use *u*, *d* and *b* to denote the units from the common pool, deaths and births respectively. Here $c = c' = 2$. The top margins list all new samples with two births, while the left margins list all old samples containing one death. Table 3 contains the actual overlap of pairs of samples, and Table 4 the probabilities of selection in the NWCR solution.

Table 3. Overlaps within block (2, 2)

	$u_1u_2b_1b_2$	$u_3u_2b_1b_2$	$u_1u_3b_1b_2$
$u_1u_2d_1$	2	1	1
$u_3u_2d_1$	1	2	1
$u_1u_3d_1$	1	1	2
$u_1u_2d_2$	2	1	1
$u_3u_2d_2$	1	2	1
$u_1u_3d_2$	1	1	2
$u_1u_2d_3$	2	1	1
$u_3u_2d_3$	1	2	1
$u_1u_3d_3$	1	1	2

Table 4. Probabilities within block (2, 2)

	$u_1u_2b_1b_2$	$u_3u_2b_1b_2$	$u_1u_3b_1b_2$
$u_1u_2d_1$	1/90	0	0
$u_3u_2d_1$	0	1/90	0
$u_1u_3d_1$	0	0	1/90
$u_1u_2d_2$	1/90	0	0
$u_3u_2d_2$	0	1/90	0
$u_1u_3d_2$	0	0	1/90
$u_1u_2d_3$	1/90	0	0
$u_3u_2d_3$	0	1/90	0
$u_1u_3d_3$	0	0	1/90

Recall that the probability of the block is 1/10. We distribute it equally among the 9 pairs of samples with overlap of 2, so each receives a mass of 1/90. All other pairs in the block have probability 0 of being selected. ■

Remark 2: Using the results of Bein, Brucker, Park and Pathak (1995), Theorem 1 can be extended to apply to more than two surveys. In this case, we have a choice of objective functions. For three surveys, we could, for example, associate a different weight to the overlap of pairs of samples for each of the 3 possible pairs of surveys and add it to the overlap of the three surveys. Note that a linear combination of SM functions with positive coefficients is an SM function. ■

3.3 Applications

The following algorithm for selecting a pair of samples for two different surveys is justified by the proof of Theorem 1.

Selection Algorithm 1 (Simultaneous selection for maximization): Consider two surveys with sample sizes n and n'. From the joint distribution given by Proposition 1, we randomly select a block labelled (i, j), say. If $\rho(c_i, c_j') = \min\{c_i, c_j'\} = c_j'$, we first randomly select c_j' units from C. To complete the selection of the sample s', we randomly select $n' - c_j'$ units from the B births. To complete the selection of the sample s, we randomly select $c_i - c_j'$ units from the remaining $C - c_j'$ common units, and then $n - c_i$ more units from the D deaths. The case $\min\{c_i, c_j'\} = c_i$ is similar and will be omitted. ■

Assume now that a sample s has already been drawn and that it belongs to super-row i. As is often the case in ongoing surveys, we need to select s' from the updated population so that the expected overlap is maximized and the estimates are unbiased. This last requirement is satisfied if the new conditional selection corresponds to a joint distribution with prescribed marginals as described in Section 3.2 and justified in Theorem 1. Since all rows within a super-row have equal probabilities, it suffices to consider the conditional distribution of super-columns given that a super-row i has been selected.

Application 1 (Conditional distribution given a super-row): Recall the notation and the proof of Proposition 1 and set $P(0) = 0$. For each super-row i, $i = 1,\ldots,n$, the conditional cumulative distribution of super-columns, given this super-row i, is $[X(i, j) - P(i-1)]/p_i$ for all j such that $Q(j) > P(i-1)$, and 0 otherwise. ■

Selection Algorithm 2 (Sequential selection for maximization): Let us assume that the old sample s belongs to super-row i. To obtain a new sample, we first randomly select a super-column from the conditional distribution above - say, the super-column j of new samples s' with c_j' units from C. If $c_i < c_j'$ then the new sample s' is formed by retaining the c_i common units from s and randomly selecting $c_j' - c_i$ more units from the remaining $C - c_i$ units in C and $n' - c_j'$ units from the B births. If $c_j' < c_i$, we deselect randomly $c_i - c_j'$ units from the c_i units in $s \cap C$ and then complete s' by randomly selecting $n' - c_j'$ births. ■

Example 3: We consider Examples 1 and 2 and assume that $s = u_1 u_2 d_1$ was selected first, in super-row 2. From Table 2 we have $x_{21} = 7/20$ and $x_{22} = 2/20$. We pick a random number $r \leq 9/20$, say $r = 3/20$. Since $3/20 < 7/20$, the new sample s' is in super-column 1 with $c' = 3$. To

select s', we retain u_1u_2, add u_3 from C and randomly select one of the two births, say b_2, so the new sample is $s'=u_1u_2u_3b_2$. ■

The implementation of the selection algorithm is particularly simple when we are updating for births (or deaths) only or simply changing the sample size.

Application 2 (Births only): Assume that $D = 0$, so we have only one super-row with $c(s) = n$ for all $s \in S$, and we determine the number of births from the appropriate hypergeometric distribution. We then select the units as above. We applied this method in selecting a sample for Statistics Canada's Local Government Surveys. ■

Application 3 (Deaths only): Now $B = 0$, so there is only one super-column. We simply select or deselect sequentially to attain the new sample size n'. An even simpler case is when both B and D are 0, i.e. only the sample size has changed. ■

4. MINIMIZING THE EXPECTED OVERLAP

4.1 Simultaneous Selection for Minimization

Consider surveys with a stratified SRSWOR design and a common stratification as in Section 3.1. For the purpose of minimizing the expected overlap, we define the block optimum of the block labeled (c, c') as $\rho(c, c') = \max\{0, c + c' - C\}$. In this section the block optimum is the smallest possible overlap that can be attained by a pair of samples within that block. In this section, in order to obtain sM matrix, we must label the rows and the columns of the matrix of overlaps in opposite order.

Construction of a joint density for minimization. Given two SRSWOR surveys, a solution to (TP) which minimizes the expected overlap can be obtained by a two-stage procedure.

1. Form super-rows and super-columns as in Section 3.1. Label the super-rows in increasing order of c and the super-columns in decreasing order of c'. Form a table whose marginals are super-row and super-column probabilities ordered as above and derive the joint distribution of the blocks by the NWCR.
2. Within each block, divide the mass equally among those pairs of samples (s, s') that have the smallest possible overlap for that block.

Theorem 2: (a) The joint density X_0 defined on the set of pairs of samples (s, s') by the two-stage procedure above has SRSWOR marginals.

(b) X_0 has the minimum expected overlap within the set of joint densities with the given marginals.

(c) X_0 has the minimum variance within the set of joint densities which satisfy (a) and have the minimum expected overlap.

Proof: (a) This follows as in the proof of Theorem 1.

(b) By Corollary 1, it suffices to show that the matrix of overlaps is sM, for which we use Proposition 3. The super-rows were labeled in increasing order of c, while the super-columns were labeled in decreasing order of c'. For a block with the optimum overlap $\rho_{ij} = \max\{0, c_i + c_j' - C\}$, we distinguish three cases:

(i) $c_i + c_j' > C$. Then $\rho_{i,j+1} = \rho_{ij} - 1$, $\rho_{i+1,j} = \rho_{ij} + 1$ and $\rho_{i+1,j+1} = \rho_{ij} + 1 - 1$, so (sM) holds with equality.

(ii) $c_i + c_j' = C$. Then $\rho_{i+1,j+1} = \rho_{ij} = 0$, $\rho_{i+1,j} = \rho_{ij} + 1 = 1$ and $\rho_{i,j+1} = 0$, and so (sM) holds with strict inequality.

(iii) $c_i + c_j' < C$. Then all four entries are 0, and so (sM) holds again with equality.

(c) We note that the matrix of the squares of the block overlap is also sM. Thus, $E_{X_0}(\rho^2)$ is minimal, and consequently, since $E_{X_0}(\rho)$ is equal to the minimum expected overlap and hence fixed, it follows that $V_{X_0}(\rho) = E_{X_0}(\rho^2) - [E_{X_0}(\rho)]^2$ is also minimal. ■

Example 4: Consider Example 1 with the overlap function defined in this section. The corresponding matrix R is sM:

Table 5. Matrix of block optima for minimization

$c \backslash c'$	3	2
0	0	0
1	1	0
2	2	1
3	3	2

Remark 3: Even if $n + n' \leq C$ and the minimum overlap in each block is 0, we must still decide how many units from the common pool must belong to each sample. Here we need not use the NWCR to obtain a feasible solution. We could, for instance, give each block a probability equal

to the product of the marginals. This does not mean selecting independently in each survey, since within blocks we do coordinate the selection of the samples as above. ■

4.2 Applications

Selection Algorithm 3 (Simultaneous selection for minimization): The matrix with super-rows and super-columns is constructed as in Section 4.1. The selection of the super-row and super-column is as in Selection Algorithm 1. We select the actual units as follows: Using an SRSWOR selection scheme, we first randomly select ρ units from C. Next, we randomly select from the remaining units in C, $c' - \rho$ units for s' and a different set of $c - \rho$ units for s. We complete s' by randomly picking $n' - c'$ births. Similarly, we select $n - c$ deaths for s.

Application 4: Note that if the two populations are the same and only the sampling fractions differ, then $c = n$, $c' = n'$ and the algorithm simplifies greatly since we have only one block. In a minimization scheme with $n + n' \leq N$, we always randomly select n units for the first survey and n' different units for the second survey. ■

5. COMPARISON WITH OTHER METHODS

We compare the NWCR method to three PRN methods: sequential SRSWOR, the method of collocated samples and the PRN method with full-stratum rotation. In the PRN methods, all units in the frame are independently assigned random numbers from the uniform distribution on [0, 1], which are then retained permanently. Each of the three studied methods uses these PRNs to select samples.

1. Sequential SRSWOR: The frame is sorted in the order of the PRNs. To select a sample, we choose a starting point $a \in [0, 1]$ and a direction (right or left). To select n units according to an SRSWOR design, we take the units corresponding to the first n PRNs to the right (or left) of a. This method is due to Fan, Muller and Rezucha (1962). It is shown in Ohlsson (1992) that this technique produces an SRSWOR design. When the frame is updated, births are independently assigned newly generated PRNs and the dead units are discarded with their PRNs.

Sequential SRSWOR for positive coordination (maximizing the overlap) of two surveys uses the same starting point (origin) and direction to select pairs of samples to maximize the expected overlap. For negative coordination (minimizing the overlap), distinct and preferably far apart origins a_1 and a_2 are chosen and the directions are either the same or opposite (Ohlsson (1995)).

2. The method of collocated samples: In collocated sampling, the units are first arranged in a random order and then assigned a sample selection number (SSN) by transforming their rank so that the SSNs are equally spaced on the interval [0, 1]. All units whose SSNs lie within the sampling interval $[a, a + n/N]$ are included in the sample. The random ordering can be done by sorting the units in the order of their PRNs. For more details, see Ohlsson (1995) and Srinath and Carpenter (1995). We use repeated collocated sampling as described by Srinath and Carpenter (1995). After removing deaths and adding births, the SSNs within a stratum are no longer equally spaced, and thus this method gives variable sample size n'.

3. PRN method with full-stratum rotation: This method is for minimizing the overlap and is described in Ernst, Valliant and Casady (2000). It assumes that a sequential SRSWOR sample has been selected on the first occasion by moving to the right. Let a_0 be the largest PRN associated with a unit in this sample. On the second occasion, a_0 becomes the starting point for the selection of the second sample. Ernst et al. (2000) showed that this method has a slight *selection bias* toward births.

Some formal properties of the PRN methods, like the expected overlap, do not appear to have been studied. To compare methods, we conducted a simulation study and performed 100,000 repetitions using the data in Example 1. The three measures used for comparison are: E, V and the selection bias, where E is the expected overlap and V the corresponding variance. Let $Q = \left(q(s')\right)_{s'\in S'}$ be the design probability on the second occasion. Ernst et al. (2000) defined the selection bias by $\sum_{s'\in S'} b(s')q(s')/n' - B/N'$, where $b(s')$ is the number of births in s'. This is the difference between the expected sample proportion of births and the population proportion of births. Since the NWCR method requires (TP) to hold, we obtain an SRSWOR design on the second occasion with no selection bias.

For positive coordination, we compared the NWCR and sequential SRSWOR methods and the two methods performed equally well as shown in Table 6. They have no selection bias and attain the maximum expected overlap. As it appears that the sequential SRSWOR method possesses the same optimal properties as the NWCR method, we did not include the method of repeated collocated sampling, for which n' varies, in the study.

Table 6. Positive coordination of two surveys

Method	E	V	Selection bias

NWCR method	1.50	0.45	0
Sequential SRSWOR	1.49	0.43	0

Table 7. Negative coordination of two surveys

Method	E	V	Selection bias
NWCR method	0.90	0.19	0
Sequential SRSWOR	0.91	0.52	0
PRN, full-stratum rotation	0.69	0.39	0.054
Repeated collocated	0.88	0.39	0

The NWCR method performs very well for minimizing the overlap. It has the smallest variance and thus the overlap of each pair (s, s') selected by the NWCR method is likely to be close to E. The full-stratum rotation method has the smallest expected overlap but it does not produce an SRSWOR design on the second occasion. Repeated collocated sampling performs better than sequential SRSWOR, but has the disadvantage that the associated sample size n' is a random variable (here $2 \le n' \le 5$). On the whole, the simulations confirm that the NWCR method gives solutions with very good properties. For the negative sample coordination, it has the advantage of minimizing both the expected overlap and its variance, while attaining the required SRSWOR design on the second occasion.

APPENDIX

Explicit construction of the NWCR solution

The following is a step-by-step algorithm for NWCR.

1. Define $x_{i0} = 0, x_{0j} = 0$.
2. Set $i = j = 1$.
3. There are three cases.
   (a) If $p_i - \sum_{s=0}^{j-1} x_{is} < q_j - \sum_{r=0}^{i-1} x_{rj}$ then $x_{ij} = p_i - \sum_{s=0}^{j-1} x_{is}$; $x_{is} = 0$ for $s>j$; increment i by 1.
   (b) If $p_i - \sum_{s=0}^{j-1} x_{is} > q_j - \sum_{r=0}^{i-1} x_{rj}$ then $x_{ij} = q_j - \sum_{r=0}^{i-1} x_{rj}$; $x_{rj} = 0$ for $r>i$; increment j by 1.
   (c) If $p_i - \sum_{s=0}^{j-1} x_{is} = q_j - \sum_{r=0}^{i-1} x_{rj}$ then set x_{ij} to their common value; $x_{is} = 0$ for $s>j$; $x_{rj} = 0$ for $r>i$; increment i and j by 1.
4. If i is now $n+1$ or j is $m+1$, stop; otherwise return to step 3 with the updated value of (i, j).

Proof of Proposition 2

The nonzero entries of X_0 can be ordered as in the above NWCR construction. We suppose that the first k, $1 \le k \le n \times m$ nonzero entries of X_0 equal to the corresponding entries of X and describe a procedure to obtain $X^* = X + \sum_{\gamma=1}^{G^*} \alpha_\gamma \Delta_\gamma$, another solution of (TP), such that the first k+1 nonzero entries of X_0 equal the corresponding entries of X^*. This will suffice to prove the proposition, since by applying this procedure repeatedly, we can obtain a solution of (TP) of the desired form, which has all the same nonzero entries as X_0, and which therefore equals X_0.

Assume that x_{ij}^0 corresponds to the $(k+1)^{st}$ nonzero entry of X_0 and that $x_{ij}^0 \neq x_{ij}$. Because of the marginal and positivity constraints, X also coincides with X_0 on all zero entries created by the application of the NWCR algorithm (see Step 3 in the NWCR algorithm above), prior to calculating x_{ij}^0. Consequently, on row i, $x_{is} = x_{is}^0$ for $s < j$; on column j, $x_{rj} = x_{rj}^0$ for $r < i$; and thus $x_{ij} < x_{ij}^0$. For X to have the correct marginals, there must then exist nonempty sets $\mathcal{R} = \{r > i : x_{rj} > x_{rj}^0\}$ and $\mathcal{S} = \{s > j : x_{is} > x_{is}^0\}$. Pick $r \in \mathcal{R}$ and $s \in \mathcal{S}$, let $\alpha_1 = \min\{x_{rj} - x_{rj}^0, x_{is} - x_{is}^0\}$, and define $X_1 = X + \alpha_1 \Delta_{r-i,s-j}(i,j)$. X_1 is another solution to (TP) such that (using obvious notation) either $\mathcal{R}_1 = \{r > i : x_{rj}^1 > x_{rj}^0\}$ or

$\mathcal{S}_1 = \{s > j : x_{is}^1 > x_{is}^0\}$ has one fewer element than did $\mathcal{R}$ or $\mathcal{S}$. Form X_2 from X_1, $\mathcal{R}_1$ and $\mathcal{S}_1$ in the same way, that we formed X_1 from X, $\mathcal{R}$ and $\mathcal{S}$. We continue in this way, forming X_3, X_4, …, X_{G^*} and stopping when either $\mathcal{R}_{G^*+1}$ or $\mathcal{S}_{G^*+1}$ is empty. At this point, we have $x_{ij}^{G^*} = x_{ij}^0$, so X_{G^*} is the desired X^* such that the first k+1 nonzero entries of X_0 equal the corresponding entries of X^*. ■

Lemma A.1

With the notation introduced before Proposition 2, set $\Delta = \Delta_{k\ell}(i,j)$. If the matrix (ρ_{ij}) is a matrix of block optima constructed as in Section 3.1, then, $E_\Delta(\rho^2) \ge 0$ and $E_\Delta(\rho^2) = 0$ if and only if $E_\Delta(\rho) = 0$.

Proof: By Lemma 1, the matrix (ρ_{ij}^2), with $\rho_{ij}^2 = \min^2\{c_i, c_j'\}$, $i = 1,...,n$, $j = 1,...,m$, is SM and so $E_\Delta(\rho^2) \ge 0$. We prove the statement about the zero expectations using mathematical

induction on k+ℓ. First consider $\Delta_{1,1}(i,j)$. In the proof of Lemma 1, we showed that $\rho_{i,j}+\rho_{i+1,j+1}=\rho_{i,j+1}+\rho_{i+1,j}$, and hence $E_{\Delta_{1,1}(i,j)}(\rho) = 0$, if and only if $\rho_{ij} = \max\{c_{i+1}, c_{j+1}'\}$. For $E_{\Delta_{1,1}(i,j)}(\rho^2) = 0$, we need $\rho_{i,j}^2 + \rho_{i+1,j+1}^2 = \rho_{i,j+1}^2 + \rho_{i+1,j}^2$. Using the proof of Lemma 1 again, we see that the equality for ρ^2 is attained if and only if $\rho_{ij}^2 = \max^2\{c_{i+1}, c_{j+1}'\}$, or equivalently, if and only if $E_{\Delta_{1,1}(i,j)}(\rho) = 0$. This proves the lemma for k =1, ℓ =1.

Without loss of generality, it suffices to prove the statement for $\Delta_{k+1,\ell}(i,j)$, assuming that it is true for all (k',ℓ'), $k'+\ell' < k+1+\ell$. We write $\Delta = \Delta_{k+1,\ell}(i,j) = \Delta_{k,\ell}(i,j) + \Delta_{1,\ell}(i+k,j) = \Delta_1+\Delta_2$ and note that $E_{\Delta_1}(\rho^2)+E_{\Delta_2}(\rho^2) = E_{\Delta}(\rho^2) = 0$ holds if and only if both terms are zero. Since $k + \ell < k + \ell + 1$, we use the inductive hypothesis and conclude that this is equivalent to $E_{\Delta_1}(\rho) = 0$ and $E_{\Delta_2}(\rho) = 0$, or $E_{\Delta}(\rho) = 0$. ■

ACKNOWLEDGEMENTS

We would like to thank J.N.K. Rao and Pierre Lavallée for their helpful suggestions.

REFERENCES

Aragon, J., and Pathak, P.K. (1990), "An algorithm for optimal integration of two surveys," *Sankhyā*, Ser. B, 52, 198-203.

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

Bein, W.W., Brucker, P., Park, J.K., and Pathak, P.K. (1995), "A Monge property for the *d*-dimensional transportation problem," *Discrete Applied Mathematics*, 58, 97-109.

Causey, B. D., Cox, L.H., and Ernst, L. (1985), "Applications of transport theory to statistical problems," *Journal of the American Statistical Association,* 80, 903-909.

Ernst, L.R., and Ikeda, M.M. (1994), "A reduced size transportation algorithm for maximizing the overlap between surveys," *Survey Methodology*, 21, 147-157.

Ernst, L.R. (1999), "The maximization and minimization of sample overlap problems: a half century results," *Bulletin of the International Statistical Institute, Proceedings*, Tome LVII, Book 2, pp 293-296.

Ernst, L.R, Valliant, R., and Casady, R.J. (2000), "Permanent and collocated random number sampling and the coverage of births and deaths," *Journal of Official Statistics*, 16, 211-228.

Fan, C.T., Muller, M.E., and Rezucha, I. (1962), "Development of sampling plans by using sequential (item by item) techniques and digital computers," *Journal of the American Statistical Association*, 57, 387- 402.

Hoffman, A.J. (1985), "On greedy algorithms that succeed," *Surveys in Combinatorics*, ed. I. Anderson: Cambridge U. Press, pp. 97-112.

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

Mitra, S.K., and Pathak, P.K. (1984), "Algorithm for optimal integration of three surveys," *Scandinavian Journal of Statistics*, 11, 257-316.

Ohlsson, E. (1992), "SAMU- the system for co-ordination of samples from the business register at Statistics Sweden-a methodological description," Research and Development Report 1992:18, Stockholm: Statistics Sweden.

Ohlsson, E. (1995), "Coordination of samples using Permanent Random Numbers," *Business Survey Methods*, eds. B.G. Cox, D. A. Binder, D. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, New York: Wiley, pp. 153-169.

Raj, D. (1956), "On the method of overlapping maps in sample surveys," *Sankhyā*, 17, 89-98.

Rivière, P. (2001), "Random permutations of random vectors as a way of co-ordinating samples," Working paper, University of Southampton.

Ross, S.M. (1983), *Introduction to Stochastic Dynamic Programming*, San Diego: Academic Press.

Srinath, K.P. and Carpenter, R.M. (1995), "Sampling methods for repeated business surveys," *Business Survey Methods*, eds. B.G. Cox, D. A. Binder, D. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott. New York: Wiley, pp.171-183.