

AN ALGORITHM THAT MAXIMIZES THE CONDITIONAL OVERLAP OF SAMPLES FROM TWO CONSECUTIVE SURVEYS: APPLICATION TO THE CANADIAN LABOUR FORCE SURVEY

Ioana Şchiopu - Kratina¹

ABSTRACT

The problem of maximizing the expected overlap for two surveys is a transportation problem which can be solved using linear programming techniques. Since all samples in both surveys have to be listed and paired to form the variables, the size of this transportation problem is quite large and solving it even for small sample sizes is impractical. This article introduces the notion of conditional overlap and presents a simple algorithm that maximizes it while preserving the selection probabilities of all samples. This method does not require solving a large system of linear equations. In fact, a simple algorithm can be given. All sample selection probabilities must be obtainable. An application to the Canadian Labour Force Survey (LFS) is given.

RÉSUMÉ

Vu sous l'angle du transport, le problème de la maximisation du recouvrement moyen entre deux enquêtes peut être résolu à l'aide de techniques de programmation linéaire. Les échantillons des deux enquêtes devant être énumérés et appariés pour former les variables, la taille de ce problème de transport est toutefois considérable, ce qui le rend difficile à résoudre, même pour de petits échantillons. Cette constatation a conduit l'auteure à examiner le concept de recouvrement conditionnel, pour lequel elle présentera un algorithme simple de maximisation respectant les probabilités de sélection des différents échantillons. Loin de nécessiter la résolution d'un grand système d'équations linéaires, cet algorithme est en fait très simple. On doit pouvoir obtenir les probabilités de sélection des échantillons. L'enquête sur la population active servira à illustrer la méthodologie proposée.

1. INTRODUCTION

1.1 Description of the problem

Oftentimes several surveys are carried out on the same population. We distinguish two situations: different surveys are carried out at the same time or essentially the same survey is carried out on two consecutive occasions. The second situation is encountered when minor modifications are made to the survey design which affect the sample selection probabilities. For example, in case of a business survey, a change in classification or a change of survey frame may take place. As a different example, assume that the selection is made proportional to the size of units, which changes over time, and an update in the size of units results in different selection probabilities for all samples. We are dealing with a different design, even though the selection mechanism remains the same. In all situations, a large overlap may considerably reduce the travelling costs or the cost of listing primary selection units. On the other hand, if several surveys are carried out at the same time on a small population, it might be desirable to minimize the size of the overlap in order to reduce response burden. The problem of minimizing the overlap is no different mathematically from the problem of maximizing it. Finally, when the same survey is carried

out on two different occasions, it is desirable to have a large overlap as this will result in smaller variability in the estimates from one occasion to the next. Classical sampling literature treats this problem for samples of sizes 1 or 2. The only method that ensures a maximum expected overlap is presented by Keyfitz (1951) for samples of size 1. Using linear programming techniques, it is possible to formulate and solve such problems, in principle, for samples of any size. In practice, the sample size must be relatively small for the problem to be computationally feasible. In the application to the Labour Force Survey (LFS), which motivated this study, the sample size in the pertinent strata is 6, so it seems appropriate to investigate the use of linear programming methods.

1.2 Organization of the article

The next section contains the basic definitions and describes the concept of expected overlap. Section 3 introduces the concept of conditional overlap and illustrates the difference between the conditional and the expected overlap. Finally, we present an application in section 4.

¹ Ioana Şchiopu-Kratina, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

2. THE EXPECTED OVERLAP

2.1 Basic definitions

Let $(S, p), (S', p')$ be two sample designs, where S, S' represent the sets of samples and p, p' the associated probabilities. For simplicity, we assume that the same population is being sampled twice and that the sample size is fixed and equal to n on both occasions. In this situation, $S = S'$.

Definition 2.1.1: (Mitra & Pathak, 1984): The sample designs (S, p) and (S', p') are integrated by $q, q: S \times S' \rightarrow [0, 1]$, if:

$$(1) \quad \sum_{s'} q(s', s) = p(s), \forall s \in S \text{ and} \\ \sum_s q(s', s) = p'(s'), \forall s' \in S',$$

i.e., p and p' are the marginals of q .

Remark 2.1.1: It follows from (1) that:

$$(2) \quad \sum \sum_{s, s'} q(s', s) = 1.$$

It follows from (1) and the positivity constraint that $q(s', s) \leq \min\{p'(s'), p(s)\}$.

Remark 2.1.2: The problem as formulated in (1) with (2) is a classical transportation problem which admits infinitely many solutions. The integration of two surveys viewed as a transportation problem was pursued by Causey *et al.* (1985).

Result 2.1.1: The following method provides a solution for any $I \times J$ table, as long as (1) and (2) hold. A cell $c(i, j)$ is selected first and assigned $q(i, j) = \min\{p'(i), p(j)\}$. The available total in the marginals is revised. This step assigns 0 mass to the remaining cells on the i^{th} row, if $q(i, j) = p'(i)$ and 0 mass to the remaining cells on column j , if $q(i, j) = p(j)$. Thus, at the end of this step, we will have assigned values to all variables on a row or/and a column. We then repeat the step by selecting one of the remaining cells. The process terminates successfully because of (2) after at most $\max\{I, J\}$ iterations.

Example 2.1.1: We present below an illustration of the procedure for a 2×2 table with given marginals:

	1	2	p'
1	1/3	1/6	1/2
2	0	1/2	1/2
p	1/3	2/3	1

We assign the cell $c(1, 1)$ the value $q(1, 1) = \min\{1/3, 1/2\} = 1/3$ which reduces the available mass on the first row to $1/2 - 1/3 = 1/6$ and to 0 on the first column. This uniquely determines $q(1, 2) = 1/6$, $q(2, 1) = 0$ and $q(2, 2) = 1/2 = \min\{1/2, 2/3\}$.

2.2 The expected overlap

Definition 2.2.1: The expected overlap associated with q in (1) is $E(q) = E_q(\rho) = \sum \sum_{s, s'} q(s', s) \rho(s', s)$, where $\rho(s', s) = \#(s \cap s')$, the number of units in the common part of the samples.

Problem I (Causey *et al.*, 1985): Find $q: S \times S' \rightarrow [0, 1]$ which integrates two surveys $(S, p), (S', p')$ and is such that the expected overlap $E(q)$ is a maximum.

Example 2.1.1 revisited: A relatively recent article by Aragon and Pathak (1990) gives a necessary condition for q to be a solution to Problem I. Their condition states that all entries on the main diagonal of the solution must equal the minimum of the corresponding marginals. Note that the solution presented in Example 2.1.1 satisfies this condition. Because Aragon and Pathak's condition leads to a unique solution in this case, we may conclude that the solution presented in Example 2.1.1 is a solution to problem I, if $\{1\}$ and $\{2\}$ represent the only possible samples in each of the two surveys.

The formulation and solution to Problem I leads to a large number of variables, even if the sample size n is small. For instance, in a stratum of size $N=30$ there are $I=593,775$ possible samples of size $n=6$ and I^2 pairs of samples which may enter in the formulation of Problem I. This would be now the maximum size of the problem for LFS strata which motivated this study. However, an increase of N or n leads to a substantial increase in I^2 . It is therefore preferable to find methods that are less computationally intensive. Another disadvantage of the formulation above is the fact that the probability of selection of all samples must be calculable. In the case of the LFS, it is not difficult to do so, as the units are selected systematically with probability proportional to the size of the units, once the units have been randomly ordered.

3. THE CONDITIONAL OVERLAP

3.1 The case of two consecutive surveys

In the context of two consecutive surveys, e.g., when the sampling scheme remains the same but the inclusion probabilities must change as a result of updating the size of the units, maximizing the conditional overlap (which we define next) seems to be the more appropriate thing to

do. In such a situation, the transition from the old to the new design takes place 'in one shot'. What matters then is to retain as much as possible of the sample selected according to the old design under the constraints of the new design. These constraints are not very stringent (see (1), (2) and Result 2.1.1). We do not aim at attaining the probabilities of the second design based on the selection of one sample on the first occasion. The only integration scheme that would reach this goal would be an independent selection of samples on the second occasion. This is hardly an acceptable solution, for reasons which were stated in Subsection 1.1.

Definition 3.1.1: (D. McDonald): The conditional overlap associated with $q: S \times S' \rightarrow [0,1]$ and the sample $s \in S$ selected on the first occasion is:

$$E(q/s) = \sum_{s'} q(s',s) \rho(s',s) / \rho(s), \quad \rho(s',s) = \#(s' \cap s), \\ \forall s' \in S'.$$

Problem II: Consider the sample designs $\mathcal{P} = (S, p)$ & $\mathcal{P}' = (S', p')$ and $s \in S$ selected on the first occasion. Find:

$$\max \{E(q/s): q \text{ integrates } \mathcal{P} \text{ \& } \mathcal{P}'\}$$

Remark 3.1.1: The solution to Problem II depends on the sample s selected on the first occasion.

3.2 Maximizing the conditional overlap

With s representing the sample selected on the first occasion, we suggest:

- (a) If $p(s) \leq p'(s)$, retain s , i.e., $q(s,s) = p(s)$. Place 0's for all entries on the corresponding column, i.e., $q(s',s) = 0, \forall s' \neq s$. Revise the marginal $p'(s)$.
- (b) If $p(s) > p'(s)$, retain s with probability $q(s/s) = p'(s)/p(s)$, i.e., set $q(s,s) = p'(s)$, then assign $q(s',s)$ out of $p(s) - p'(s)$ to $s' \in S'$ in an overlap-greedy fashion and such that $q(s',s) \leq p'(s')$. Revise the marginals $p'(s')$, $s' \neq s$. The 'greedy' distribution is achieved by placing all samples s' in a decreasing order of overlap with s , say $s'(i_j), j \in I$ and then distributing $p(s) - p'(s)$ in that order so that $s'(i_j)$ gets $q'(s'(i_j),s) \leq p'(s'(i_j)), j \in J$. (see Lemma 3.2.1).
- (c) At the end of steps (a) or (b), values for $q(s',s)$ will have been assigned, $\forall s' \in S'$ and the row marginal(s) will have been revised. We must now fill out the entries of a matrix $A(s)$ of dimension at most $I \times (J-1)$ with given marginals. We do so in any way

consistent with the marginals and obtain a solution to Problem II. To be on the safe side, we may require that the expected overlap be maximized for the matrix $A(s)$, i.e., solve Problem I for $A(s)$.

The lemma below formalizes steps (a)-(b) above, with $d(j) = p'(s'(j)), \{s'(j)\}_{j \in J} = S'$ and $b = p(s)$.

Lemma 3.2.1: Consider $x = \{q(j)\}$ and positive constants $\{\rho(j)\}, d = \{d(j)\}, b \in \mathbb{R}$, where $0 \leq q(j) \leq d(j), j \in J$. The solution to the problem: $\max \{\sum_i q(i) \rho(i) \mid q \ni \sum q(i) = b\}$ can be found using a greedy algorithm.

The proof of this lemma is immediate.

However, $p(s) - p'(s)$ is generally small compared to $p'(s')$, $s' \neq s$. In the LFS application with s selected first and $n=6$, (see also example 3.2.3) it may suffice to calculate $p'(s')$ for s' such that $\rho(s' \cap s) > 4$ only, i.e., for at most $6 \times 24 = 144$ samples.

A greedy algorithm is a natural way to go about solving a maximum overlap problem. However, even for a simple two dimensional problem it need not lead to an optimal solution. In the following example, each cell is assigned a 'weight' and we wish to maximize the expected weight over the entire table, for the given marginal probabilities.

Example 3.2.1: The following matrix displays 'weights' for the internal cells and probabilities for the marginals:

2	1.5	1/2
1.5	0	1/2
1/2	1/2	1

The maximum weight is 2, so we set $q(1,1) = 1/2 \Rightarrow q(2,2) = 1/2, q(1,2) = q(2,1) = 0 \Rightarrow \sum_{i,j} q(i,j) \rho(i,j) = 2 \times 1/2 = 1$. A larger expected weight would have been obtained if $q(1,2) = 1/2, q(1,1) = q(2,2) = 0$ and $q(2,1) = 1/2 \Rightarrow \sum_{i,j} q(i,j) \rho(i,j) = 1.5$.

Example 3.2.2: Assume that the population $P = \{1,2,3,4\}$ is sampled twice. The sample size on each occasion is $n = 2$. Here $S = S' = \{s(1), s(2), s(3), s(4), s(5), s(6)\}$ where $s(1) = \{1,2\}, s(2) = \{1,3\}, s(3) = \{1,4\}, s(4) = \{2,3\}, s(5) = \{2,4\}, s(6) = \{3,4\}$. The marginal probabilities are given in the matrix of sample overlaps presented below.

Example 3.2.3: For simplicity, let us denote $p'(i) = p'(s(i)), i \in I, p(j) = p(s(j)), j \in J$. The values of

$p(i), i \in I$, will stay the same as in Example 3.2.2 and the table below but $p'(1) = 2/24, p'(2) = 3/24, p'(3) = p'(4) = p'(5) = 4/24, p'(6) = 7/24$. In this example, the only sample with a strictly larger probability of selection on the second occasion is $s(6)$, which has 0 overlap with $s(1)$. Any assignment that maximizes the conditional overlap will violate the necessary condition for maximizing the expected overlap stated in Example 2.1.2 revisited.

We work out Example 3.2.2 next.

Assume that $s(1)$ was selected on the first occasion. Set $q(1,1) = \min\{1/12, 1/6\} = 1/12$. We have $1/6 - 1/12 = 1/12$ left over from $p(1)$. Set $q(3,1) = \min\{5/24, 1/12\} = 1/12$. The entries on the first column have been placed and we must now fill in the entries of a 5×5 matrix, since in this case all entries of the first row other than $q(s,s)$ must be 0. For (c), by the result of Aragon and Pathak (1990), we must place for

values on the diagonal of $A(s)$ the minimum of the corresponding revised marginals. These conditions determine the final solution which is presented in the table below. Notice that the entire solution $q(i,j)$ depends on the sample $s(1)$, even for samples $s \neq s(1)$ selected in the first survey. As a consequence of this asymmetry, if the two surveys were to be integrated continuously, the overlap could be quite small for some samples $s \neq s(1)$ selected in the first survey. As we will see in Example 3.3.1, there is also no guaranty that the expected overlap over all possible selections on the first occasion is small. Notice that the probabilities of selecting the samples in both designs (the marginals) are attained.

The matrix of sample overlaps is:

	s(1)	s(2)	s(3)	s(4)	s(5)	s(6)	p'
s(1)	2	1	1	1	1	0	1/12
s(2)	1	2	1	1	0	1	3/24
s(3)	1	1	2	0	1	1	5/24
s(4)	1	1	0	2	1	1	1/6
s(5)	1	0	1	1	2	1	5/24
s(6)	0	1	1	1	1	2	5/24
p	1/6	1/6	1/6	1/6	1/6	1/6	1

	s(1)	s(2)	s(3)	s(4)	s(5)	s(6)	p'
s(1)	1/12	0	0	0	0	0	1/12
s(2)	0	3/24	0	0	0	0	3/24
s(3)	1/12	0	3/24	0	0	0	5/24
s(4)	0	0	0	1/6	0	0	1/6
s(5)	0	0	1/24	0	1/6	0	5/24
s(6)	0	1/24	0	0	0	1/6	5/24
p	1/6	1/6	1/6	1/6	1/6	1/6	1

3.3 Comparison with Problem I.

Example 3.3.1 (Example 3.2.2 revisited): The solution to Problem II given in Example 3.2.2 is not a solution to Problem I. According to Aragon and Pathak, the solution to Problem I has on the diagonal of the (6×6) matrix: $q_1(1,1) = 1/12, q_1(2,2) = 3/24, q_1(3,3) = 4/24, q_1(4,4) = 1/6, q_1(5,5) = 1/6, q_1(6,6) = 1/6$.

Since $q(3,3) = 3/24 < q_1(3,3) = 4/24$ a solution that maximizes the conditional overlap need not maximize the expected overlap.

Example 3.3.2 It is possible sometimes to refine (b) of the algorithm to obtain a solution that is closer to a solution to Problem I. In Example 3.2.2, instead of assigning $1/12$ to $s(3)$, we may distribute it evenly between $s(3)$ and $s(5)$, i.e., assign to cells $c(3,1)$ and $c(5,3)$ $1/24$ mass each (see the first column of the matrix above). Now $c(3,3)$ must be assigned $4/24$ and so the necessary condition for maximizing the expected overlap for the entire matrix is satisfied. Note that both $s(3)$ and $s(5)$ have larger probabilities of selection on the second occasion and that the increase in probabilities from one occasion to the next is the same, i.e., $5/24 - 1/6 = 1/24$. On the other hand, we cannot always obtain solutions to both problem I and II, as a consideration of Example 3.2.3 would show.

As formulation for integration of two consecutive surveys, Problem II presents several advantages. Firstly, the maximum number of variables is I as opposed to I^2 for problem I. Secondly, it can be solved using a greedy algorithm, which is very efficient. Thirdly, the conditional overlap is a maximum, which is what really matters. A disadvantage of Problem II is the fact that its use is limited to a one-time integration of two consecutive surveys. The entire solution depends on the last sample selected according to the old design. If we were to integrate the surveys continuously, we could obtain a very poor overlap on other occasions.

4. APPLICATION

4.1 General remarks and application to LFS.

Since the selection mechanism is the same on both occasions, one might attempt to relate and then 'adjust' this selection mechanism to reflect the changes in the probabilities of selection. The following example shows that this cannot be done in an obvious way. We confine ourselves to a sample selection that is taken systematically and proportional to the size of the units, from a population that has been randomly ordered first. This type of selection is performed in some areas of the LFS.

Example 4.1.1: Consider a population $P = \{1,2,3,4\}$ of 4 units and sizes: $\{5, 30, 20, 45\}$ in that order, from which we wish to select $n=2$ units (note that $(5 + 30 + 20 + 45)/n = 100/2 = 50$). Assume that $r=17 \in [0, 50]$ has been drawn randomly, which means that the sample $s = \{2,4\}$ has been selected before the updates, since $17 + 50 = 67$. After the updates, assume that the sizes are (in the same order): $\{20, 30, 5, 45\}$. The random number r would require that the sample $s' = \{1,4\}$ be selected, when in fact s should be retained since its probability of selection is unchanged on the second occasion. The example also shows that we cannot condition on the random order of units. The $\{1, 2, 3, 4\}$ ordering gives a higher chance of selection to $\{2, 4\}$ on the first occasion. Only by taking into account the ordering $\{3, 2, 1, 4\}$ on both occasions we see that the probabilities of selecting s , i.e., $\{2\}$ and $\{4\}$ in any order are the same before and after the updates.

REFERENCES

- Aragon, J., and Pathak, P.K. (1990). "An algorithm for optimal integration of two surveys", *Sankhya*, 52, Series B, Pt. 2, 198-203.
- Causey, B.D., Cox, L.H., and Ernst, L. (1985). "Application of transportation theory to statistical problems", *Journal of the American Statistical Association*, 80, 903-909.
- 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, 257-163.