Indian Statistical Institute

Cost Robustness of an Algorithm for Optimal Integration of Surveys
Author(s): K. Krishnamoorthy and Sujit Kumar Mitra
Source: *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, Vol. 48, No. 2 (Aug., 1986), pp. 233-245

Published by: Springer on behalf of the Indian Statistical Institute
Stable URL: http://www.jstor.org/stable/25052444
Accessed: 29-08-2016 02:34 UTC

Sankhyā : The Indian Journal of Statistics
1986, Volume 48, Series B, Pt. 2, pp. 233-245.

COST ROBUSTNESS OF AN ALGORITHM FOR OPTIMAL INTEGRATION OF SURVEYS

By K. KRISHNAMOORTHY and SUJIT KUMAR MITRA
Indian Statistical Institute

SUMMARY. Let the cost of an integrated survey depend only on the number ν of distinct units in the survey and the cost function $C(\nu)$ be monotonic increasing in ν. Further let the increment $C(\nu+1)-C(\nu)$ diminish with ν. It is shown that the optimal integrated survey of Mitra and Pathak (1984) is cost optimal under such a cost function. We also present integration plans for three surveys which are cost optimal when $C(3)-C(2) > C(2)-C(1)$.

1. Introduction

Consider a finite population with N units serially numbered 1, 2, ..., N. Let S denote the set $\{1, 2, ..., N\}$. It is proposed to carry out k separate surveys on this population. The i-th survey assigns a probability of selection P_{ij} to the j-th unit and thus corresponds to a random variable X_i which assigns a probability P_{ij} to the integer j ($j = 1, 2, ..., N$). An integrated survey corresponds to a joint probability distribution of random variables $X_1, X_2, ..., X_k$ which realises for X_i the same marginal distribution as the one determined by the i-th survey. For $\boldsymbol{x}=(x_1, x_2, ..., x_k)$ in S^k the k-th cartesian power of S, $\nu(\boldsymbol{x})$ denotes the number of distinct integers appearing in the k coordinates of $\boldsymbol{x}$. An integrated survey is called optimal if $E[\nu(\boldsymbol{X})]$ is a minimum. It was pointed out in Maczynski and Pathak (1980) that an optimal integrated survey always exists but is not unique. Mitra and Pathak (1984) present algorithms for deriving optimal integrated surveys for $k = 2$ and 3. The present paper is a follow up of the Mitra-Pathak paper and aims firstly to clarify some doubts that may crop up through a cursory reading of this paper. Example 1 illustrates a stochastic matrix, specifying the selection probabilities for the k surveys, for which the optimal integrated survey is unique. Though the optimal integrated survey is in general not unique, one may speculate that each such survey plan would lead to a unique probability distribution for $\nu(\boldsymbol{X})$ the number of distinct units in the integrated survey. Examples 2 and 3 point out to the contrary. One may here be pleasantly surprised to find that the integration plan is still optimal even though he may have initially missed the bus by putting less than the maximum possible

AMS (1980) *subject classification :* 62D05.

Key words and phrases : Algorithm, Configuration, Cost optimality, Cost robustness, Integrated survey, Maximum matching, Safe available surplus.

mass to the event set $S_1 = \{\boldsymbol{x} : \nu(\boldsymbol{x}\} = 1\}$. In one of these examples an optimal integrated survey plan in fact assigns a probability 0 to S_1 even though $\theta_1 > 0$. In the other example an optimal integrated survey assigns a probability 0 to S_3 even though $1-\theta_2 > 0$ where S_2 and S_3 are defined the same way as S_1 noting that for $k = 3$, 2 and 3 are the only other values that $\nu(\boldsymbol{x})$ could assume and $\theta_i = \sum\limits_{1}^{N} P_{(i)j}$, where $P_{(i)j}$ is the i-th smallest value among P_{1j} , $P_{22}, \ldots, P_{kj}$.

The optimal integrated survey is clearly cost optimal if the cost of observing and analysing an integrated sample with ν distinct units depends only on ν and the cost function $C(\nu)$ is linear in ν with a positive slope. It may be more realistic to assume that $C(\nu)$ increases monotonically with ν but the increments themselves diminish with $\nu[\Delta C(\nu) > 0,\ \Delta^2 C(\nu) < 0]$. The second object of this paper is to show that the optimal integrated survey derived in Mitra and Pathak (1984) is fairly cost robust in the sense that it retains its cost optimality even for this wide class of cost functions. The case of a cost function which exhibits a faster than linear growth presumably only of academic interest is treated in Section 3. We are able to present integrated surveys which are cost optimal under such a cost function.

We recall that a configuration in general, in the context of k surveys, a $k \times N$ array of nonnegative numbers with each row adding up to the same number not necessarily equal to 1 (Mitra and Pathak, 1984). Each step of the Mitra-Pathak algorithm transforms one configuration into another with the common row totals of the successive configurations shrinking to zero.

2. Three examples

Example 1 : Consider the stochastic matrix for three surveys given in Table 1.

TABLE 1. VALUES OF P_{ij}

$i \backslash j$	1	2	3	4
1	a	$1-a$	0	0
2	b	0	$1-b$	0
3	b	0	0	$1-b$

$1 > b \geqslant a > 0$

The algorithm given in Mitra and Pathak (1984) leads to the following optimal integrated survey

$$p_{111} = \Pr\{X_1 = 1,\ X_2 = 1,\ X_3 = 1\} = a$$

$$p_{234} = 1-b,\ p_{211} = b-a$$

where the P_{jkl}'s in general are analogously defined. Here $E[\nu(\boldsymbol{X})] = 1(a)+2(b-a)+3(1-b) = 3-a-b$. We now show that here the optimal integrated survey is uniquely determined. Clearly an integrated survey here is supported entirely on 8 points

$$(x_1, x_2, x_3), x_1 = 1, 2 ; x_2 = 1, 3 ; x_3 = 1, 4.$$

One considers the linear equations which the p_{jkl}'s must satisfy to realise the marginal distributions given in Table 1 along with the linear equation which will ensure optimality (i.e. $E[\nu(\boldsymbol{X})] = 3-a-b$). It is seen that these equations uniquely determine p_{234} as $p_{234} = 1-b$. This along with Pr $\{X_2 \neq 1\} =$ Pr $\{X_3 \neq 1\} = 1-b$ on account of nonnegativity of the p_{jkl}'s imply $p_{114} = p_{131} = p_{134} = p_{214} = p_{231} = 0$ which in turn implies $p_{111} = a$, $p_{211} = b-a$. The uniqueness of the optimal integrated survey is thus established.

The next two examples show that in general not only are the optimal integrated survey plans not unique, they do not even lead to a unique probability distribution for the number of distinct units. It is easily seen that for an optimal integrated survey the vector (Pr (S_1), Pr (S_2), Pr (S_3)) is unique upto a constant multiple of the vector $(-1, 2, -1)$.

Example 2 : Consider the stochastic matrix for 3 surveys given in Table 7 of Mitra and Pathak (1984) after correcting the obvious mistake therein (interchanging of rows and columns)

TABLE 2. VALUES OF P_{ij}

i \ j	1	2	3
1	.2	.5	.3
2	.3	.2	.5
3	.5	.3	.2

The algorithm leads to the following plan for an optimal integrated survey

$$p_{111} = p_{222} = p_{333} = .2 ; p_{211} = p_{232} = p_{331} = .1 ; p_{231} = .1.$$

An alternative optimal integrated survey is given by

$$p_{111} = .1, p_{222} = p_{333} = .2 ; p_{211} = .2, p_{232} = p_{331} = p_{131} = .1.$$

Note that we have transferred equal probability masses of .1 from (1, 1, 1)ϵ S_1 and (2, 3, 1) ϵ S_3 to (2, 1, 1) and (1, 3, 1) both in S_2. We have here Pr (S_1) = .5, Pr (S_2) = .5.

Example 3 : Consider the following stochastic matrix for 3 surveys.

TABLE 3. VALUES OF P_{ij}

$i \backslash j$	1	2	3	4
1	.1	.1	.1	.7
2	0	.5	.2	.3
3	.3	.1	.6	0

The algorithm leads to the following plan for an optimal integrated survey

$$p_{222} = p_{333} = .1 \; ; \; p_{121} = p_{433} = .1 \; ; \; p_{443} = .3 \; ; \; p_{421} = .2, \; p_{423} = .1.$$

An alternative optimal integrated survey is given by

$$p_{121} = p_{221} = p_{422} = p_{323} = .1 \; ; \; p_{433} = .2 \; ; \; p_{443} = .3 \; ; \; p_{421} = .1.$$

Note that we have transferred equal probability masses from S_1 and S_3 to S_2. We have here Pr $(S_2) = .9$, Pr $(S_3) = .1$.

3. The main results

We shall prove the following theorem.

Theorem 1 : *Let the cost of an integrated survey depend only on the number ν of distinct units in this survey and the cost function $C(\nu)$ be monotonic increasing in ν. Further let the increments $C(\nu+1)-C(\nu)$ diminish with ν. The optimal integrated survey derived through the algorithm in Mitra and Pathak* (1984) *is cost optimal under such a cost function.*

Proof : The case of two surveys is trivial. We confine our attention to $k = 3$. First the case $\theta_2 < 1$. Here the optimal integrated survey assigns a probability of θ_1 to S_1, $(\theta_2-\theta_1)$ to S_2 and $(1-\theta_2)$ to S_3. If this is not cost optimal let there exist an integrated survey assigning probability of p_i to S_i with a lower expected cost, that is

$$C(1)p_1+C(2)p_2+C(3)p_3 < C(1)\theta_1+C(2)\,(\theta_2-\theta_1)+C(3)\,(1-\theta_2) \qquad \dots \quad (1)$$

This implies on subtracting $C(2)$ from both sides of (1)

$$[C[1]-C(2)]p_1+[C(3)-C(2)]p_3 < [C(1)-C(2)]\theta_1+[C(3)-C(2)]\,(1-\theta_2)$$

$$\Rightarrow [C(3)-C(2)]p_3 < [C(1)-C(2)]\,(\theta_1-p_1)+[C(3)-C(2)]\,(1-\theta_2). \qquad \dots \quad (2)$$

The R.H.S. of (2) is further seen to be less than

$$[C(3)-C(2)]\,(1-\theta_2) \text{ since } [C(1)-C(2)]\,(\theta_1-p_1) \leqslant 0.$$

This implies $p_3 < (1-\theta_2)$.

Inequality (2) implies

$$[C(1)-C(2)]p_1+[C(2)-C(1)]p_3 < [C(1)-C(2)]\theta_1+[C(2)-C(1)]\,(1-\theta_2) \quad \ldots \quad (3)$$

$$\text{since } [C(2)-C(1)] > [C(3)-C(2)]$$

$$\Rightarrow [C(2)-C(1)]\,(p_1+2p_2+3p_3) < [C(2)-C(1)]\,[\theta_1+2(0_2-\theta_1)+3(1-\theta_2)]$$

$$\text{adding } 2[C(2)-C(1)] \text{ to both sides of (3)}$$

$$\Rightarrow p_1+2p_2+3p_3 < \theta_1+2(\theta_2-\theta_1)+3(1-\theta_2) \text{ since } [C(2)-C(1)] > 0$$

which is impossible since the algorithm for optimal integration minimises the expected value of the number of distinct units.

The case $\theta_2 \geqslant 1$ can be similarly established making use of the facts that $\theta_1 -$ Pr $(S_1) \geqslant 0$ and $C(\nu)$ is monotonic increasing. Q.E.D.

Theorem 2 : *Let the cost function $C(\nu)$ be monotonic increasing in ν and*

$$C(3)-C(2) > C(2)-C(1).$$

Let there exist an optimal integrated survey which assigns a probability 0 *to S_3. This survey is also cost optimal under such a cost function.*

Proof : Let the optimal integrated survey assign a probability p_i^0 to S_i and $p_3^0 = 0$. If this is not cost optimal let there exist an integrated survey assigning probability p_i to S_i with a lower expected cost

$$C(1)p_1+C(2)p_2+C(3)p_3 < C(1)p_1^0+C(2)p_2^0$$

$$\Rightarrow C(1)p_1+C(2)p_2+[2C(2)-C(1)]p_3$$

$$\leqslant C(1)p_1+C(2)p_2+C(3)p_3 < C(1)p_1^0+C(2)p_2^0$$

$$= C(1)p_1^0+C(2)p_2^0+[2C(2)-C(1)]p_3^0$$

$$\Rightarrow [C(2)-C(1)](p_1+2p_2+3p_3) < [C(2)-C(1)](p_1^0+2p_2^0+3p_3^0)$$

adding $C(2)-2C(1)$ to both sides.

Since $C(2)-C(1) > 0$, this contradicts our assumption that the optimal integrated survey minimises the expected number of distinct units. Q.E.D.

Example 2 illustrates such a situation.

The following theorem can be proved on similar lines. We omit the proof.

Theorem 3 : *Let the cost function $C(\nu)$ be monotonic increasing in ν and*

$$C(3)-C(2) > C(2)-C(1).$$

Let there exist an optimal integrated survey which assigns probability 0 *to* S_1. *This survey is also cost optimal under such a cost function.*

Example 3 illustrates such a situation.

Plans for optimally integrating three surveys derived through the algorithm in Mitra and Pathak (1984) are of two types. One that uses stages 1, 2 and 3 leads to the probability distribution of the number of distinct units given by

$$\Pr\{S_1\} = \theta_1,\ \Pr\{S_2\} = \theta_2-\theta_1,\ \Pr\{S_3\} = 1-\theta_2. \qquad \ldots \quad (4)$$

The other uses stages 1, 2* and 3* and leads to

$$\Pr\{S_1\} = \theta_1,\ \Pr\{S_2\} = 1-\theta_1. \qquad \ldots \quad (5)$$

Clearly as noted in Mitra and Pathak (1984) the first alternative works only if

$$\theta_2 \leqslant 1. \qquad \ldots \quad (6)$$

Theorem 4 shows that condition (6) is also sufficient.

Theorem 4 : *For an optimal integration of three surveys with the probability distribution as in* (4) *to exist it is necessary and sufficient that* (6) *holds.*

Proof: The necessity part is trivial since (4) $\Rightarrow$

$$\Pr\{S_3\} = 1-\theta_2 \geqslant 0.$$

For the sufficiency part we show that if (6) holds all the steps in stage 2 will go through smoothly without any hindrance. Let us consider the configuration at the commencement of stage 2. We classify the N columns of the configuration as follows. Let I_{i1} denote the set of indices of those columns for which the i-th row contains the smallest column entry (0), I_{i2} the set of indices of those columns for which the i-th row contains the second smallest column entry and I_{i3} the set of indices of those columns for which the i-th row contains the maximum column entry, ties being arbitrarily broken-up.

Clearly,

$$I_{i1} \bigcup I_{i2} \bigcup I_{i3} = S, \quad i = 1, 2, 3.$$

A typical column in I_{11}, written as a row vector, is thus

$$(0, a, b) \text{ or } (0, b, a), a \leqslant b.$$

If this is column j, the second smallest entry in this column is zeroed out by assigning a mass δ_u to a point (u, j, j) in S^3 such that $u \neq j$ and $\sum_{u \neq j} \delta_u = a$. The target is to execute this operation without affecting the second smallest entry in any column. This implies that $\delta_u > 0$ only if $u \epsilon I_{13}$ and in any case δ_u should not exceed $P_{(3)u} - P_{(2)u}$. The available surplus in the columns of I_{13} should be adequate to meet the demands made by the columns in I_{11} that is

$$\text{Supply} - \text{demand} = \sum_{u \epsilon I_{13}} \{P_{(3)u} - P_{(2)u}\} - \sum_{u \epsilon I_{11}} \{P_{(2)u} - P_{(1)u}\}$$

$$= \sum_{u \epsilon I_{13}} \{P_{1u} - P_{(2)u}\} + \sum_{u \epsilon I_{11}} \{P_{1u} - P_{(2)u}\} + \sum_{u \epsilon I_{12}} \{P_{1u} - P_{(2)u}\}$$

(noting that for u in I_{12}, $P_{1u} - P_{(2)u} = 0$)

$$= \sum_{u \epsilon S} P_{1u} - \sum_{u \epsilon S} P_{(2)u} = 1 - \theta_2 \geqslant 0.$$

Hence if (6) holds, in row 1, supply is adequate to meet the demand. The same argument also holds for the other rows. Q.E.D.

Consider a point in S_1 and a point in S_3. We say that these two points are matched if they agree in one coordinate. Since the coordinates of a point in S_1 are all identical and those of a point in S_3 are necessarily distinct these two points could agree at most in one coordinate e.g. the pair of points (1, 1, 1) and (2, 3, 1) in Example 2. Since the algorithm assigns a mass of $\delta = .2$ to (1, 1, 1) and $\delta' = .1$ to (2, 3, 1), a mass of $.1 = \min(\delta, \delta')$ could be removed from each of the points $(1, 1, 1) \epsilon S_1$ and $(2, 3, 1) \epsilon S_3$ and redistributed to each of the points (2, 1, 1) and (1, 3, 1) both in S_2. The resulting plan is still optimal since these manoeuvres do not affect $E\{\nu(\boldsymbol{X})\}$. However alternative plans derived through the algorithm of Mitra and Pathak (1984) may allow for differing degrees of matching. The arguments given in the proof of Theorem 4 suggest a strategy for 'maximum matching'. This is described below, keeping in view possible uses in connection with Theorems 2, 3 and also 5 which we present later in this section. We have said earlier that for $u \epsilon I_{13}$ the available surplus is $P_{(3)u} - P_{(2)u}$. Since here the plan assigns a mass of $P_{(1)u}$ to $(u, u, u) \epsilon S_1$ if the entire surplus is used in stage 2 which only generates points in S_2, the possibility of a matching of a point in S_3 with the point

(u, u, u) in S_1 will be annihilated in the bud. Ideally out of the surplus $P_{(3)u}-P_{(2)u}$, stage 2 should leave an amount of $P_{(1)u}$ for stage 3. We therefore define the notion of safe available surplus α_u as follows.

$$\alpha_u = \begin{cases} 0 & \text{if } P_{(3)u}-P_{(2)u} \leqslant P_{(1)u} \\ P_{(3)u}-P_{(2)u}-P_{(1)u} & \text{otherwise.} \end{cases} \quad \ldots \quad (7)$$

We have maximum matching in the first coordinate if

$$\sum_{u \epsilon I_{13}} \alpha_u \geqslant \sum_{u \epsilon I_{11}} \{P_{(2)u}-P_{(1)u}\} \quad \ldots \quad (8)$$

and the extent of matching is measured by

$$\mu_1 = \sum_{u \epsilon I_{13}} \beta_u \quad \ldots \quad (9)$$

where $$\beta_u = \begin{cases} P_{(1)} & \text{if } \alpha_u > 0 \\ P_{(3)u}-P_{(2)u} & \text{otherwise.} \end{cases} \quad \ldots \quad (10)$$

Since the demand in any case has to be met, if (8) is not satisfied the maximum matching as given in (9) is reduced by the corresponding amount. One has here the following revised expression for μ_1

$$\mu_1 = \sum_{u \epsilon I_{13}} (\beta_u+\alpha_u) - \sum_{u \epsilon I_{11}} \{P_{(2)u}-P_{(1)u}\} = 1-\theta_2. \quad \ldots \quad (11)$$

Compute similarly μ_2 and μ_3. The maximum matching in a plan for integrating 3 surveys is given by

$$\min (\mu_1+\mu_2+\mu_3, 1-\theta_2) \quad \ldots \quad (12)$$

noting that once the configuration prior to stage 3 is so arrived at, with each column containing atmost one nonnull entry, operations in stage 3 could be so directed as to attain the bound set in (12). Clearly no change in the steps under stage 3 of Mitra and Pathak (1984) are called for if $\mu_i = 1-\theta_2$ for $i = 1, 2$, or 3 since the maximum matching $1-\theta_2$ can be secured through the i-th row itself. We shall therefore consider the case where $\mu_i < 1-\theta_2$ for $i = 1, 2, 3$. Here it will be convenient to split the nonnull entry γ_u in column u of the configuration (prior to stage 3) as $\gamma_u = \beta_u+\eta_u$ $(\eta_u \geqslant 0)$, β_u and η_u, representing respectively the critical and the noncritical mass. The distinction is however only superficial and will help us in describing the required modifications in the steps of stage 3. We first zero out the critical masses β_u in the first row with the help of noncritical masses in other rows. This will be possible if

$$\sum_{u\epsilon I_{23}} \eta_u \geqslant \sum_{u\epsilon I_{13}} \beta_u, \sum_{u\epsilon I_{33}} \eta_u \geqslant \sum_{u\epsilon I_{13}} \beta_u \quad \dots \quad (13)$$

i.e. $$1-\theta_2 \geqslant \mu_1+\mu_2 \text{ and } 1-\theta_2 \geqslant \mu_1+\mu_3. \quad \dots \quad (14)$$

If one of these inequalities is not true for example if $1-\theta_2 < \mu_1+\mu_3$ the noncritical masses in third row are insufficient for this purpose. That is, while zeroing out the critical masses in the first row not only are the entire noncritical masses in third row used up but one has to draw from part of the critical masses in third row as well. Hence when one zeroes out next the balance of the critical masses in third row which add upto $1-\theta_2-\mu_1$, the noncritical masses in the first row are just sufficient for the purpose and irrespective of what happens in the second row the bound $1-\theta_2$ in (12) is clearly attained.

If (13) is satisfied the critical masses in row 2 are next zeroed out with the help of noncritical masses in other rows. Since in the first row only noncritical masses remain, no special efforts are needed here. However in the third row the available noncritical masses should suffice for this purpose, that is

$$\sum_{u\epsilon I_{33}} \eta_u - \sum_{u\epsilon I_{13}} \beta_u \geqslant \sum_{u\epsilon I_{23}} \beta_u \quad \dots \quad (15)$$

or equivalently

$$1-\theta_2 \geqslant \mu_1+\mu_2+\mu_3. \quad \dots \quad (16)$$

If (15) is satisfied, the critical masses in third row are next zeroed out with the help of available masses in the other two rows which are incidentally all noncritical. Thus the bound $\mu_1+\mu_2+\mu_3$ in (12) is attained. If (15) is not true the bound $1-\theta_2$ in (12) is attained in a manner similar to what we have described above. It may be wise to distinguish the points in S_1 that are to be used in maximal matching by putting a * against the coordinate that supports a pivotal critical mass. Thus $p_{12*3} = .1$ indicates that the point (1, 2, 3) is matched with (2, 2, 2) and a mass of .1 can be removed from each of these points for transfer to (1, 2, 2) and (2, 2, 3).

Consider a plan for maximally matched optimal integrated survey and a plan for optimal integrated survey derived from the same by transferring equal masses, to the maximum extent possible from S_1 and S_3 to points in S_2. It is shown in Theorem 5 that the resulting plan (call it $\mathcal{P}_1$) is cost optimal for

$$1 < [C(3)-C(2)]/[C(2)-C(1)] \leqslant 2 \quad \dots \quad (17)$$

even if Pr (S_1) and Pr (S_3) are both nonnull in this case.

Consider the plan $\mathscr{P}_1$ so derived in the preceding paragraph. There is zero matching now between the points in S_3 and those in S_1. However transfers to S_2 can still take place under slightly more unfavourable condition. Thus given a mass 2δ attached to the point $(u, u, u) \epsilon S_1$ and a mass δ at the point $(j, k, l) \epsilon S_3$ with u, j, k, l all distinct, a mass of δ could be transferred to each of the points (j, u, u), (u, k, u), (u, u, l) in S_2. These transfers though not profitable under (17) will turn out to be profitable if

$$2 < [C(3)-C(2)]/[C(2)-C(1)]. \qquad \ldots \quad (18)$$

We keep on making these transfers until zero mass is left either in S_1 or in S_3 or in both. It is shown in Theorem 5 that the resulting plan $\mathscr{P}_2$ is cost optimal under (18).

Theorem 5 : *Let $C(\nu)$ be an increasing function of ν. The integration plans $\mathscr{P}_1$ and $\mathscr{P}_2$ are cost optimal under conditions* (17) *and* (18) *respectively.*

Proof : Let $C(\nu) \uparrow \nu$ and

$$[C(3)-C(2)]/[C(2)-C(1)] > 1. \qquad \ldots \quad (19)$$

Further let $\mathscr{P}$ be an integration plan which is cost optimal under such a cost function. We consider the points in S_1, S_2 and S_3 which received positive mass under $\mathscr{P}$. The respective subsets are denoted by S_1^*, S_2^* and S_3^* and let these correspond to configurations[1] $¢_1$, $¢_2$ and $¢_3$ respectively. We now deduce certain properties of $¢_1$, $¢_2$ and $¢_3$ as a consequence of the cost optimality of $\mathscr{P}$. Firstly we note that there is zero matching between the points in S_3^* with those in S_1^*, because otherwise certain positive mass δ could be removed from a point in S_3^* and a matching point in S_1^* and then transferred to points in S_2 as indicated earlier. When (19) holds these transfers would result in strict improvement in respect of the expected cost contradicting the cost optimality of $\mathscr{P}$. This implies that if a column in $¢_1$ has nonnull entries then the corresponding column in $¢_3$ has only null entries. Similarly we note that a column in $¢_3$ can have atmost one nonnull entry, because otherwise an application of the Mitra-Pathak algorithm to $¢_3$ would result in shifting positive masses from S_3^* to certain points in S_1 and/or S_2 and these shifts again would result in a strict improvement in respect of the expected cost. Let the j-th column of $¢_1$ be null and without any loss of generality let the single nonnull entry c in the j-th column of $¢_3$ occur in the first row. This implies that the j-th column of $¢_2$, written as a row vector would look like

$$(P_{1j}-c, P_{2j}, P_{3j}) = (q_{1j}, q_{2j}, q_{3j}) \text{ (say)}$$

[1] Each configuration here is a $3 \times N$ table with the three rows describing the partial marginal probability distributions for the three surveys.

where P_{tj}'s are the entries of the original configuration specifying the marginal probability distributions for the three separate surveys. We now show that q_{1j} cannot be the smallest or the second smallest entry in that column. The points in S_2 with j appearing in at least one of the three coordinates are of six types (1) (j, j, k_1), (2) (j, k_2, j), (3) (k_3, j, j), (4) (k_4, k_4, j), (5) (k_5, j, k_5) and (6) (j, k_6, k_6) while the points in S_3 with j appearing in the first coordinate are of one type—(j, i, l). Clearly the point (k_3, j, j) could not appear in S_2^*. If it did, certain positive mass δ could be transferred from each of the points (k_3, j, j) in S_2^* and (j, i, l) in S_3^* to the point (k_3, i, l) in S_3 or S_2 and (j, j, j) in S_1 and this would result in a strict improvement in respect of the expected cost. Also the point (k_4, k_4, j) (and for a similar reason the point (k_5, j, k_5)) could not appear in S_2^*. For example if the point (k_4, k_4, j) did, a positive mass δ could be transferred from (j, i, l) in S_3^* and (k_4, k_4, j) in S_2^* to points (k_4, k_4, l) and (j, i, j) both in S_2. These transfers would result in a strict improvement in respect of the expected cost. Thus there are only points of type (1), (2) and (6) in S_2^*. Let the total mass received by points of type (u) be denoted by a_u. We have therefore

$$a_1+a_2+a_6 = q_{1j}$$

$$a_1 = q_{2j}$$

$$a_2 = q_{3j}$$

$$\Rightarrow \quad q_{1j} \geqslant q_{2j},\, q_{1j} \geqslant q_{3j} \Rightarrow q_{1j} = q_{(3)j}.$$

Let ξ_i $(i = 1, 2, 3)$ denote the sum of the i-th minimum column entries of all the N columns of ϕ_2. ξ_1, ξ_2 and ξ_3 thus correspond to θ_1, θ_2 and θ_3 respectively as defined for the original configuration. Let $\theta_1-\gamma_1$ be the total mass assigned to the points in S_1^*. The structure of ϕ_2 we have just established implies $\xi_1 = \gamma_1$, $\xi_2-\xi_1 = \theta_2-\theta_1$ and for each of the three rows of ϕ_2, the row total is equal to $\bar{\xi} = (\xi_1+\xi_2+\xi_3)/3$. If $\bar{\xi} \geqslant \xi_2$, an application of the Mitra-Pathak algorithm to C_2 would result in shifting from S_2^* a mass of ξ_1 to S_1 and $\bar{\xi}-\xi_2$ to S_3, the balance mass $\xi_2-\xi_1$ would remain in S_2. With the allocation already made in S_1^* and S_3^* remaining undisturbed, these shifts would result in a plan for which

$$P(S_1) = \gamma_1+\theta_1-\gamma_1 = \theta_1$$

$$P(S_2) = \xi_2-\xi_1 = \theta_2-\theta_1$$

$$P(S_3) = 1-\theta_2$$

which is clearly optimal in the sense of Mitra and Pathak (1984). In other words the integration plan $\mathscr{P}$ can be derived from a plan which is optimal in this sense by shifting masses from S_1 and S_3 to S_2. This can be more clearly seen as follows : Assume without any loss of generality that the first column of $¢_2$ has nonnull entries in each row. As enumerated earlier six types of points in $¢_2$ could have contributions in the first column of $¢_2$. (1) $(1, 1, k_1)$, (2) $(1, k_2, 1)$, (3) $(k_3, 1, 1)$, (4) $(k_4, k_4, 1)$, (5) $(k_5, 1, k_5)$ and (6) $(1, k_6, k_6)$. For reasons we have stressed earlier S_2^* cannot exclusively have points of type (4), (5) and (6). It should in fact have points of atleast two of types (1), (2) and (3). For example if it has only points of type (1) and none of type (2) or (3) then for the first column to have the stipulated property it is necessary that S_2^* should have some points of type (4). Again a point of type (1) cannot coexist with a point of type (4) in S_2^* for precisely the same reason. Thus either S_2^* has exclusively points of all the three types (1), (2) and (3) (with k_1, k_2 and k_3 all distinct) or has only two of them e.g. of types (1) and (2) (with $k_1 \neq k_2$) with or without points of type (6). If plan $\mathscr{P}$ assigns masses p, q, r respectively to $(1, 1, k_1)$, $(1, k_2, 1)$ and $(k_3, 1, 1)$ and $p \geqslant q \geqslant r$, an application of the Mitra-Pathak algorithm to the resulting configuration will lead to a redistribution of masses as follows :

mass	point
$q+r$	$(1, 1, 1) \in S_1$
$p-q$	$(1, 1, k_1) \in S_2$
$q-r$	$(1, k_2, k_1) \in S_3$
r	$(k_3, k_2, k_1) \in S_3$.

It is seen that if one transfers masses to S_2 symmetrically from the point $(1, 1, 1)$ in S_1 and matching point $(1, k_2, k_1)$ in S_3 and then asymmetrically from the point $(1, 1, 1)$ in S_1 and the nonmatching point (k_3, k_2, k_1) in S_3 the original distribution of masses to the three points $(1, 1, k_1)$, $(1, k_2, 1)$ and $(k_3, 1, 1)$ is restored. The argument is similar for the case where the plan $\mathscr{P}$ assigns masses p and q to the points $(1, 1, k_1)$ and $(1, k_2, 1)$. It is interesting to observe that even if the plan $\mathscr{P}$ had assigned a mass s to the point $(1, k_6, k_6)$, $(k_1 \neq k_2 \neq k_6)$ and Mitra-Pathak algorithm applied to the resulting configuration then the mass assigned to the point $(1, k_6, k_6)$ will not be affected by the redistribution. This means that when the smallest nonzero entry in different columns similar to column 1 (are zeroed out) disjoint sets of points are affected by the redistribution. The claim made earlier in this paragraph is thus substantiated. The plan $\mathscr{P}_1$ which ensures maximal matching betwen points in S_1 and S_3 before these shifts take place in a symmetric manner is thus seen to be cost optimal when (17) holds. The cost optimality of $\mathcal{P}_2$ under (18) is similarly established.

If $\bar{\xi} < \xi_2$ the Mitra-Pathak algorithm shifts from S_2^* a mass ξ_1 to S_1 and a mass $\bar{\xi}-\xi_1$ to S_2 which contradicts the cost optimality of $\mathcal{P}$ unless $\xi_1 = 0$. If $\xi_1 = 0$, that is $P(S_1^*) = \theta_1$, $P(S_2^*) = \theta_2-\theta_1$, $P(S_3^*) = 1-\theta_2$ since $P(S_2^*) \neq \theta_2-\theta_1$ contradicts either the cost optimality of $\mathcal{P}$ or the optimality of the plan derived through the Mitra-Pathak algorithm. This shows that the plan $\mathcal{P}$ is itself optimal in the sense of Mitra and Pathak (1984) and the argument given in the preceding paragraph is valid. Q.E.D.

Open problem. It may be of interest to characterize the situation where the plan for an optimal integrated survey is unique. Table 1 illustrates one such case.

For references to earlier work of Keyfitz, Lahiri and Des Raj on the optimal integration of two surveys see any one of the two papers listed below.

References

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

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

Paper received : August, 1985.

Revised : February, 1986.