ON THE METHOD OF OVERLAPPING MAPS IN SAMPLE SURVEYS

By DES RAJ

Indian Statistical Institute, Calcutta

1. Introduction

In multipurpose surveys involving the estimation of several characters, it is usually found desirable to select the units with one set of probabilities for estimating one group of characters and with a different set of probabilities for estimating another group of characters. For example, in the National Sample Survey (NSS) of India, population is made the basis for selection for the household enquiry and area for the land utilisation survey. An important problem arising in such a situation is that of designing a suitable selection procedure, so that the sample units (villages in case of the NSS) for the two types are almost identical or near to one another. Such a procedure will greatly reduce the cost of operations in the field.

Lahiri (1954) has given two methods of selection for the purpose, called the 'serpentine' method and the 'two dimensional' method. He has not, however, entered into the mathematics of the problem. The object of this paper is to present the problem mathematically and offer general solutions.

2. Formulation of the problem

Suppose a tract contains n villages. We are required to select a pair of sample villages, one with probabilities

$$\frac{a_1}{G}, \frac{a_2}{G}, \ldots, \frac{a_n}{G}$$

proportional to area and the other with probabilities

$$\frac{b_1}{G}, \frac{b_2}{G}, \ldots, \frac{b_n}{G}$$

proportional to population in such a manner that the two villages are close to one another, if not identical. Let c_{ij} be the distance (in some sense) between the i-th area village and the j-th population village.

TABLE 1. AMOUNTS OF PROBABILITY MASS TO BE DISTRIBUTED

	village no.	by population						total
		1	2	...	j	...	n	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	1	x_{11}	x_{12}	...	x_{1j}	...	x_{1n}	a_1
	2	x_{21}	x_{22}	...	x_{2j}	...	x_{2n}	a_2
by area	⋮							⋮
	i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{in}	a_i
	⋮							⋮
	n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{nn}	a_n
	total	b_1	b_2	...	b_j	...	b_n	a

Let x_{ij}/G $(i, j = 1, \ldots, n)$ be the probability with which the corresponding pair of villages is selected. The problem is to find x_{ij} such that

$$\sum_{j=1}^{n} x_{ij} = a_i, \sum_{i=1}^{n} x_{ij} = b_j; \quad \sum_{1}^{n} a_i = \sum_{1}^{n} b_j = G; \quad x_{ij} \geqslant 0$$

and $$Z = \sum \sum c_{ij} x_{ij} \text{ is minimised.}$$

Stated thus, this is the familiar 'transportation problem' (Koopmans, 1951) in linear programming. The solution by the simplex method due to Dantzig is given in the book referred to above.

Starting with the arbitrary basic solution of Dantzig, one gets the final solution in a finite number of stages. If the villages are arranged in a serpentine fashion and the same arrangement is used for area as well as for population, it is interesting to see that Lahiri's 'serpentine method' is the same as Dantzig's arbitrary basic solution with which the iterative process starts.

3. AN ILLUSTRATION

As an illustration of the method, we consider the following ten villages, the map of which is given in Fig. 1 below:

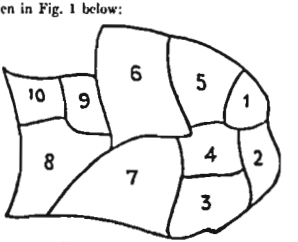

Fig 1. Map of ten villages

The villages are numbered in a serpentine fashion and their areas and populations are given in the table below:

TABLE 2. AREA AND POPULATION

village no.	area	population	village no.	area	population
(1)	(2)	(3)	(1)	(2)	(3)
1	3	8	6	15	20
2	4	5	7	13	10
3	6	10	8	11	5
4	5	5	9	5	10
5	11	6	10	3	1

Measuring the actual geographical distances (from the centres) between the villages, the cost matrix is obtained as the following:

TABLE 3. COST MATRIX c_{ij}

	village no.	by population 1	2	3	4	5	6	7	8	9	10
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	1	0	5	10	5	5	11	14	20	15	21
	2	5	0	7	5	8	13	13	21	18	23
	3	10	7	0	6	11	14	8	17	16	20
	4	5	5	6	0	6	9	9	16	13	18
	5	5	8	11	6	0	6	11	16	12	17
by area	6	11	13	14	9	6	0	10	11	5	10
	7	14	13	8	9	11	10	0	9	9	12
	8	20	21	17	16	16	11	9	0	6	5
	9	15	18	16	13	12	5	9	6	0	5
	10	21	23	20	18	17	10	12	5	5	0

The tables below give the arbitrary basic solution (of Dantzig), the final solution and salient features of the iterative process.

TABLE 4. ARBITRARY BASIC SOLUTION

	village no.	by population 1	2	3	4	5	6	7	8	9	10	total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	1	3										3
	2	4										4
	3	1	5	0								6
	4			5								5
	5			5	5	1						11
by area												
	6					5	10					15
	7						10	5				15
	8							5	5	1		11
	9									5		5
	10									4	1	5
	total	8	5	10	5	6	20	10	5	10	1	80

TABLE 5. FINAL SOLUTION

village no.	by population									
	1	2	3	4	5	6	7	8	9	10
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	2									
2		4								
3			6							
4		0		5						
5	5	0			6					
by area										
6						15				
7		1	4			0	10			
8						5		5	1	
9									5	
10									4	1

TABLE 6. BRIEF DETAILS OF THE ITERATIVE PROCESS

step (i)	$Z_i = Z_{i-1} - M_{i-1}\,\theta_{i-1}$	$M_i = \max.(\bar{c}_{ij} - c_{ij})$	θ_i
(1)	(2)	(3)	(4)
1	381	20	1
2	355	28	4
3	253	22	0
4	243	23	0
5	243	13	1
6	230	10	4
7	190	8	4
8	158	16	0
9	158	13	0
10	158	7	1
11	151	0	

It is found by the author that in problems of this type where the cost matrix contains all zeros in the diagonal, instead of starting with Dantzig's solution, which is very inefficient in this case, one should start with any solution with the maximum mass in the diagonal (obtained by putting $\min(a_i, b_i)$ in the diagonal). In fact, in this example, seven steps were necessary to get at a solution with the maximum mass in the diagonal and then only four further steps were required to get the optimum solution.

It is of interest to note in this connection that Lahiri's two-dimensional solution (in table 7 below), obtained by inspection, is almost as good as the optimum since $Z = 152$ in this case.

TABLE 7. LAHIRI'S TWO-DIMENSIONAL SOLUTION

village no.	by population									
	1	2	3	4	5	6	7	8	9	10
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	3									
2		4								
3			6							
4		1		4						
5	8				5					
by area 6						15				
7			4	1			10			
8						4		5	2	
9									5	
10						1			3	1

4. The optium character of the serpentine method

In the illustration given above the distance between two villages is defined as the actual geographical distance between the centres of the villages. To simplify the problem we may define 'distance' as the difference between the serial numbers of the villages when the villages have been numbered in a serpentine fashion. We shall now show that, with this definition of distance, Dantzig's arbitrary basic solution (which is the same as Lahiri's serpentine method) gives the optimum solution in the sense that it minimises the expected 'distance'. Without any loss of generality we shall take the case of $n = 5$ villages with areas and populations given in Table 1 before.

Let Dantzig's arbitrary solution be given by Table 8. The essential feature of this solution is that if one travels from the first to the last cell containing the basis, the path is a continuous one and is always parallel to the sides.

TABLE 8. ARBITRARY SOLUTION

village no.	by population					total
	1	2	3	4	5	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	x_{11}					a_1
2	x_{21}	x_{22}	x_{23}			a_2
by area 3			x_{33}	x_{34}		a_3
4				x_{44}		a_4
5				x_{54}	x_{55}	a_5
total	b_1	b_2	b_3	b_4	b_5	O

TABLE 9. COST MATRIX c_{ij}

village no.	by population				
	1	2	3	4	5
(1)	(2)	(3)	(4)	(5)	(6)
1	0	1	2	3	4
2	1	0	1	2	3
by area 3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

With this definition of distance, the (direct) cost matrix is given by Table 9. To obtain the indirect cost matrix $\tilde{c}_{ij}$, we have $\tilde{c}_{ij} = c_{ij}$ for any element (i, j) appearing in the basis. The other elements $\tilde{c}_{ij}$ are given by $\tilde{c}_{ij} = u_i + v_j$ where u_i and v_j are to be determined from the elements in the basis.

Let $u_1 = c_{11}$ and $v_1 = 0$. Then we have in succession

$$u_2 = c_{21} - v_1 = c_{21},$$
$$v_2 = c_{22} - u_2 = c_{22} - c_{21},$$
$$v_3 = c_{23} - u_2 = c_{23} - c_{21},$$
$$u_3 = c_{33} - v_3 = c_{33} + c_{21} - c_{23},$$
$$v_4 = c_{34} - u_3 = c_{34} - c_{33} + c_{23} - c_{21},$$
$$u_4 = c_{44} - v_4 = c_{44} - c_{34} + c_{33} + c_{21} - c_{23},$$
$$u_5 = c_{54} - v_4 = c_{54} - c_{34} + c_{33} - c_{23} + c_{21},$$
$$v_5 = c_{55} - u_5 = c_{55} - c_{54} + c_{24} - c_{33} + c_{23} - c_{21}.$$

TABLE 10. COST MATRIX $\tilde{c}_{ij}$ CORRESPONDING TO BASIS

village no.	by population				
	1	2	3	4	5
(1)	(2)	(3)	(3)	(5)	(6)
1	c_{11}				
2	c_{21}	c_{2}	c_{23}		
by area 3			c_{33}	c_{34}	
4				c_{44}	
5				c_{54}	c_{55}

Substituting $c_{ij} = |i-j|$, we have

$$u_1 = 0,\ u_2 = 1,\ u_3 = 0,\ u_4 = -1,\ u_5 = 0$$

and

$$v_1 = 0,\ v_2 = -1,\ v_3 = 0,\ v_4 = 1,\ v_5 = 0.$$

Hence the indirect cost matrix $\bar{c}_{ij}$ is as given in Table 11.

TABLE 11. INDIRECT COST MATRIX $\bar{c}_{ij}$

village no.	by population 1	2	3	4	5
(1)	(2)	(3)	(4)	(5)	(6)
1	0	−1	0	1	0
2	1	0	1	2	1
3	0	−1	0	1	0
by area 4	−1	−2	−1	0	−1
5	0	−1	0	1	0

Comparing this with the direct cost matrix c_{ij} of Table 9, we see that

$$M = \max (\bar{c}_{ij} - c_{ij}) = 0$$

so that the solution obtained is optimum.

5. Some further remarks on the serpentine method

A number of interesting conclusions emerge from the result proved above. We have proved that if the villages are arranged in a serpentine fashion, the serpentine method gives the optimum solution. But it is possible to arrange the villages in a serpentine fashion in a number of ways. Since for any serpentine arrangement the marginal totals a_i's and b_js' would remain unchanged (they may occur in a different order), the same minimum value of the expected 'distance' is achieved, in whatever manner the villages be arranged provided the arrangement is serpentine. This proves that the method given by Lahiri (1954) namely 'when the serpentine arrangement is more or less at our choice we should endeavour to arrange the villages in such a manner that the cumulative density fluctuates as frequently as possible about the tehsil density', cannot improve the situation. Any serpentine arrangement will do as well.

6. The problem of Keyfitz

Mention must be made here of the work of Keyfitz(1951) who considers the case when first stage units within strata are to be selected with different probabilities at two successive occasions, so that the probability of having identical first stage units at the two occasions is maximised. Stated mathematically, the problem is the same as discussed in § 2 where the cost matrix is given in Table 12 below:

TABLE 12. COST MATRIX IN KEYFITZ' PROBLEM

unit no.	1951 1	2	3	...	n
(1)	(2)	(3)	(4)	(5)	(6)
1	0	1	1	...	1
2	1	0	1	...	1
1950 3	1	1	0	...	1
⋮					
n	1	1	1	...	0

By the general theory (and obviously enough) the optimum solution consists in putting in the diagonals as much mass as possible, viz., $\min(a_i, b_i)$ in the (i, i) cell. We shall now show that the somewhat unwieldy selection procedure given by Keyfitz is equivalent to the simple procedure given by us. To take a concrete case, let $n = 5$ in Table 1 and $a_1 > b_1$, $a_2 > b_2$, $a_3 > b_3$, $a_4 < b_4$, $a_5 < b_5$ (there is no loss of generality involved here). Then the probability of getting identical villages in our method is given by

$$\frac{1}{G}\sum_{i=1}^{5}\min(a_i, b_i) = \frac{1}{G}(b_1+b_2+b_3+a_4+a_5).$$

By Keyfitz' method, probability of getting identical villages is equal to

$$1-\frac{1}{G}(b_1+b_2-a_4-a_5)$$

$$= 1-\frac{1}{G}(G-b_1-b_2-b_3-a_4-a_5) = \frac{1}{G}(b_1+b_2+b_3+a_4+a_5)$$

so that the two methods are equivalent i.e. Keyfitz' method is also optimum. It may be noticed here that Keyfitz is thinking of a situation where the selection is to be made at two different occasions viz., a's are the sizes for 1950 while b's are the sizes for 1951. In such a case, we will select a village with probabilities proportional to a's in 1950 (say village 1 is selected in the sample).

Then at the second occasion we will select a village with probabilities proportional to $x_{11}/a_1, x_{12}/a_1, \ldots, x_{15}/a_1$ so that the overall probability that a combination (say) (1, 3) is selected is

$$\frac{a_1}{G}\times\frac{x_{13}}{a_1} = \frac{x_{13}}{G} \text{ as desired in the method.}$$

7. THE PROBLEM OF GOODMAN AND KISH

Another related problem occurs in stratified sampling. In the usual type of stratified sampling, units within one stratum are selected independently of those within another stratum. But more generally, one may select a pair of units, one from each stratum, in a dependent way such that certain preferred types of pairs have a higher probability of selection and consequently others have a lower probability of selection. Such a problem has been considered by Goodman and Kish (1950). Their population consists of two strata, one containing 3 coastal and 3 inland units while the other contains five units of which one is coastal and the others are inland. Two units have to be selected, one from each stratum, with assigned probabilities within strata. The object is to maximise the probability of selecting one coastal and one inland unit. It is obvious that this problem easily reduces to the one considered before. The cost matrix is given in Table 13 below:

TABLE 13. COST MATRIX IN GOODMAN AND KISH'S PROBLEM

			stratum 1					
		units	costal			inland		
			B	C	F	A	D	E
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
stratum 2	inland	a	0	0	0	1	1	1
		b	0	0	0	1	1	1
		c	0	0	0	1	1	1
		e	0	0	0	1	1	1
	costal	d	1	1	1	0	0	0

The optimum solution, giving the probabilities with which pairs of units be selected, is presented in Table 14.

TABLE 14. OPTIMUM SOLUTION IN GOODMAN AND KISH'S PROBLEM

		stratum 1						
	units	B	C	F	A	D	E	total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
stratum 2	a	15						15
	b	0	10	20	0			30
	c				10	0		10
	e					20	5	25
	d						20	20
	total	15	10	20	10	20	25	100

This type of problem, however, raises new issues. In the previous problem we were concerned with the estimation of several characters and a particular scheme of selection did not affect the variances of the estimates of individual characters so long as the units within strata were selected with certain assigned probabilities. But in this case since the same character is under investigation in the two strata, the variance of the final estimate will be affected due to the dependent manner in which pairs of units are selected. To be more specific, let there be two strata of sizes N_1 and N_2 respectively. From the first stratum a unit is to be selected with probabilities $p_i (i = 1, \ldots, N_1)$ and another unit is to be selected from stratum two with probabilities $p_j' (j = 1, 2, \ldots, N_2)$ in such a way that the expected cost (of travel) is minimised. Suppose p_{ij} $(i = 1, \ldots, N_1; j = 1, 2, \ldots, N_2)$ is the optimum set of probabilities for selecting the i-th unit from the first stratum and the j-th unit from the second stratum.

Then
$$\hat{\bar{y}}_{dep} = \frac{y_i}{p_i} + \frac{y_j'}{p_j'}$$

is an unbiased estimate of the population total.

Also
$$V(\hat{\bar{y}}_{dep}) = \sum \frac{y_i^2}{p_i} + \sum \frac{y_j'^2}{p_j'} + 2 \sum \sum p_{ij} \frac{y_i}{p_i} \frac{y_j'}{p_j'} - Y^2.$$

In case selection within one stratum is independent of that within another, we have

$$\hat{\bar{y}}_{ind} = \frac{y_i}{p_i} + \frac{y_j'}{p_j'},$$

$$V(\hat{\bar{y}}_{ind}) = \sum \frac{y_i^2}{p_i} + \sum \frac{y_j'^2}{p_j'} - (Y_1^2 + Y_2^2)$$

so that

$$V_{dep} - V_{ind} = 2 \sum \sum \frac{p_{ij} - p_i p_j'}{p_i p_j'} y_i y_j'$$

One cannot establish that V_{dep} is always smaller than V_{ind} so that the possibility is there that the new method, though decreasing the travel cost, may increase the variance of the estimate.

Another point worth mentioning is the estimation of error variance in such designs. Obviously enough, it is not possible to get unbiased estimates of the sampling error even in case of independent sampling within strata by selecting only one unit from each stratum. If, however, a duplicate set of units is selected (in the same way as before), it is possible to obtain unbiased estimates of the sampling error. If $\hat{\bar{y}}_{1dep}$ and $\hat{\bar{y}}_{2dep}$ denote the population total estimates based on the first and the second set and $\hat{\bar{y}}'_{dep}$ denotes the pooled estimate, we have

$$\hat{\bar{y}}'_{dep} = \tfrac{1}{2}(\hat{\bar{y}}_{1dep} + \hat{\bar{y}}_{2dep}),$$

$$V(\hat{\bar{y}}'_{dep}) = \tfrac{1}{2} V(\hat{\bar{y}}_{dep}),$$

$$\hat{V}(\hat{\bar{y}}_{dep}) = \left(\frac{\hat{\bar{y}}_{1dep} - \hat{\bar{y}}_{2dep}}{2}\right)^2.$$

In case sampling within one stratum is independent of that within another, we have the comparable quantities:

$$\hat{\bar{y}}'_{ind} = \tfrac{1}{2}(\hat{\bar{y}}_{1ind} + \hat{\bar{y}}_{2ind}),$$

$$V(\hat{\bar{y}}'_{ind}) = \tfrac{1}{2} V(\hat{\bar{y}}_{ind}),$$

$$\hat{V}(\hat{\bar{y}}_{ind}) = \left(\frac{\hat{\bar{y}}_{1ind} - \hat{\bar{y}}_{2ind}}{2}\right)^2.$$

References

1. Goodman, R. and Kish, L. (1950): Controlled selection—a technique in probability sampling. *J. Amer. Stat. Ass.*, 45, 350-72.
2. Keyfitz, N. (1951): Sampling with probabilities proportional to size — adjustment for changes in the probabilities. *J. Amer. Stat. Ass.*, 46, 105-109.
3. Koopmans, T. C. (1951): *Activity Analysis of Production and Allocation*, John Wiley and Sons, New York, 350-373.
4. Lahiri, D. B. (1954): Technical paper on some aspects of the development of the sample design. *The National Sample Survey No. 5, Ministry of Finance*, Government of India.

Paper received ; March, 1955.