ON RELATIVE SAMPLING OF VARIOUS REGIONS OF THE FIELD

BY G. C. SHALIGRAM

Department of Agriculture, Maharashtra State

AND

M. B. GOLHAR AND M. N. GHOSH

Institute of Agricultural Research Statistics, I.C.A.R.

1. INTRODUCTION

IN yield surveys a field is sample-harvested by locating a plot of definite dimensions in a random manner within the field. The method used for locating a plot randomly is to locate the corner of the plot in the field by selecting its co-ordinates from a table of random numbers and then to mark the plot of given dimensions in the positive directions with the chosen point as its corner. This method of random selection of a plot suffers from a serious drawback in that it does not give an equal chance of selection to the different portions of the field, and a portion nearer the borders of the field has less chance of being included in the harvested plot than a corresponding portion in the central part of the field. This point was noted by Mahalanobis (1944, 1946), who referred to this as an argument against the use of large size plots. Thus the usual method of selection of plots gives a biased sample and consequently the yield estimate may be biased if portions nearer the border have different yield per unit area than the corresponding portion in the central region, *i.e.*, if there is a border effect in the yield. The bias, if any, would depend on the relative shapes and sizes of the field and the plot and on the border effect in the yield.

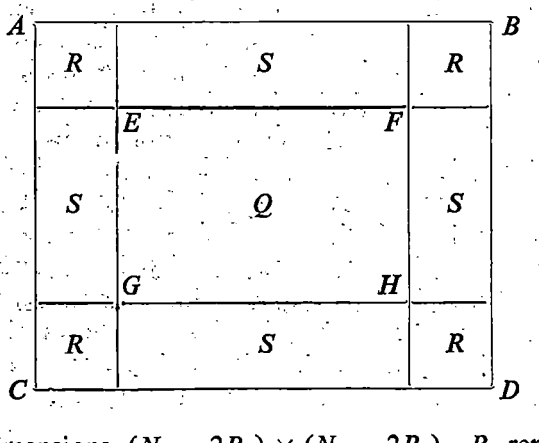
A procedure for making a strictly random selection in this case is to divide the whole field into the plots of given dimensions and then to draw one out of them randomly. But this is not practically feasible because it is not usually possible to divide the field into portions of given size. Mahalanobis (1946) has suggested the use of small plot size to reduce the bias; but to be effective the plot has to be very small which would, however, increase the variance and/or cost of survey. Choosing a number of very small plots may be considered as an

alternative procedure, but here again we come across another kind of bias noticed by Panse (1946) and Sukhatme (1946) for small size plots.

We shall consider here the probabilities of the different portions of the field to be included in the harvested plot and the averages of the probabilities over specified regions of the field, *i.e.*, the corners, sides and the central part. For a suitable definition of these regions the average probabilities have been shown to have definite proportions. Also, a new method of selection of the plot is described which will approximately equalise these average probabilities, so that bias is likely to be small, even when there is a border effect in the yields.

2. Probabilities of Inclusion in Various Regions of the Field

Consider a rectangular field of dimensions $N_L \times N_B$ (units of length) from which a plot of dimensions $P_L \times P_B$ is to be sample-harvested randomly. We shall suppose $N_L > 2P_L$ and $N_B > 2P_B$, *i.e.*, the field is relatively large compared to the plot. The field is now divided into portions as indicated in the diagram. There are four corner regions, four side regions and one central region represented by the symbols R, S and Q respectively. Q represents the central



region of dimensions $(N_L - 2P_L) \times (N_B - 2P_B)$, R represents the corner regions of dimensions $P_L \times P_B$ and S represents the side regions of $(N_L - 2P_L) \times P_B$ and $(N_B - 2P_B) \times P_L$ dimensions.

Since the corner regions have the same dimensions as those of the plot, according to the usual method of locating the plot by locating its left-hand upper corner (say), we chose random numbers between 0 and $(N_L - P_L)$ and 0 and $(N_B - P_B)$ in the rectangle with corners (A, H). Thus it is approximately the same as consider-

ing the co-ordinates (x, y) of the corner of the plot to be chosen with an uniform distribution in the rectangle (A, H). If (ξ, η) is any point in the corner region (A, E), then it is included in the plot if only $0 < x \leqslant \xi$, $0 < y \leqslant \eta$ so that the probability that the point (ξ, η) is included in the plot is

$$\frac{\xi}{N_L - P_L} \cdot \frac{\eta}{N_B - P_B}.$$

It may be seen that it is the same for other corners also when instead of ξ and η we consider distances from the nearest edges of the field.

If the point (ξ, η) is in the central portion ϕ, then it is included in the plot if

$$\xi - P_L < x \leqslant \xi, \quad \text{and} \quad \eta - P_B < y \leqslant \eta.$$

Thus the probability that a point in the central region Q is included in a plot, is

$$\frac{P_L}{N_L - P_L} \cdot \frac{P_B}{N_B - P_B}.$$

Similarly the probability that a point in the sides is included in the plot is

$$\frac{P_L}{N_L - P_L} \cdot \frac{\eta}{N_B - P_B} \quad \text{if it is along the length of the field}$$

and

$$\frac{\xi}{N_L - P_L} \cdot \frac{P_B}{N_B - P_B} \quad \text{if it is along the breadth of the field.}$$

The probability of inclusion for a point in the central region is thus constant while the probability decreases as the edge is approached in the corner and side regions and it becomes zero on the edge itself. A comparison of these probabilities may be made by averaging the probabilities over the various regions, *i.e.*, corners, sides and the central parts. These averages, of course, are not probabilities but only used for the purpose of comparison. The average probability for the corner regions is given by

$$\frac{1}{P_L \times P_B} \int_0^{P_L} \int_0^{P_B} \frac{\xi}{N_L - P_L} \cdot \frac{\eta}{N_B - P_B} \, d\xi d\eta$$

$$= \frac{P_L}{4(N_L - P_L)} \cdot \frac{P_B}{(N_B - P_B)}.$$

Similarly the average probability of the side regions is

$$\frac{P_L}{2(N_L - P_L)} \cdot \frac{P_B}{(N_B - P_B)}.$$

Thus, when the regions are defined as before by considering the plot size, the average probabilities of the central, side and corner regions are in the definite proportion of $1 : \frac{1}{2} : \frac{1}{4}$.

3. Probabilities with Defined Border Regions

We thus find that the probability of a point in the border being included in a plot is less than the corresponding probability of the central region. If there is no definite pattern in the yield per unit area in different parts of the field, *i.e.*, the field can be regarded as a random one, then such varying probabilities of selection would not give rise to a biased estimate of yield. It is likely, however, that near the border the yield per unit area may be different from the central part of the field, because the border regions are more exposed to wind and other factors of weather than the central regions; while in irrigated fields the border regions may be more effectively irrigated. The competition from neighbouring plants may also be less for the border regions. Thus yield per unit area in the border region may be significantly different from the central region. A proper definition of the border region can therefore be made according to the distance from the borders upto which such factors operate. Suppose the above operational definition of the border region gives strips of breadth b along the length of the field and strips of breadth a along the breadth of the field then we consider corner regions of dimensions $a \times b$ instead of $P_L \times P_B$ and central region of dimensions $(N_L - 2a) \times (N_B - 2b)$. The average probabilities in all regions for the points being included in the plot are calculated below. The estimate of the yield from the field may be adjusted, if possible, for such unequal probability of selection of different points in different regions.

Assume $N_L - 2a > P_L$ and $N_B - 2b > P_B$. Then the following are the probabilities that any point (ξ, η) in the field is included in the randomly chosen plot

$$= \frac{\xi}{N_L - P_L} \cdot \frac{\eta}{N_B - P_B} \quad \text{if } \xi \leqslant P_L, \ \eta \leqslant P_B$$

$$= \frac{P_L}{N_L - P_L} \cdot \frac{\eta}{N_B - P_B} \quad \text{if } \xi > P_L, \ \eta \leqslant P_B$$

$$= \frac{\xi}{N_L - P_L} \cdot \frac{P_B}{N_B - P_B} \quad \text{if } \xi \leqslant P_L, \quad \eta > P_B$$

$$= \frac{P_L}{N_L - P_L} \cdot \frac{P_B}{N_B - P_B} \quad \text{if } \xi > P_L, \quad \eta > P_B$$

1. Average probability for points in corner region being included in plot

$$= \frac{ab}{4(N_L - P_L)(N_B - P_B)} \quad \text{if } a \leqslant P_L, \quad b \leqslant P_B$$

$$= \frac{bP_L^2}{4a(N_L - P_L)(N_B - P_B)} + \frac{bP_L(a - P_L)}{2a(N_L - P_L)(N_B - P_B)}$$

$$\text{if } a > P_L, \quad b \leqslant P_B.$$

$$= \frac{\left[\frac{P_L^2 P_B^2}{4} + \frac{P_L(a-P_L)P_B^2 + P_L^2 P_B(b-P_B)}{2} + P_L(a-P_L)P_B(b-P_B)\right]}{ab(N_L - P_L)(N_B - P_B)}$$

$$\text{if } a > P_L, \quad b > P_B.$$

Average probability in side region (along the length)

$$= \frac{1}{N_L - 2a}\left[\frac{(P_L^2 - a^2)b}{2} + \frac{P_L(N_L - 2P_L)b}{2}\right]\frac{1}{(N_L - P_L)(N_B - P_B)}$$

$$\text{if } a \leqslant P_L, \quad b \leqslant P_B$$

$$= \frac{1}{N_L - 2a}\left[\frac{P_L(N_L - 2a)}{(N_L - P_L)} \cdot \frac{b}{2(N_B - P_B)}\right] \quad \text{if } a > P_L, \quad b \leqslant P_B$$

$$= \frac{1}{(N_L - 2a)b}\left[\frac{P_L(N_L - 2a)}{(N_L - P_L)} \cdot \frac{P_B^2}{2(N_B - P_B)} + \frac{P_L(N_L - 2a)}{(N_L - P_L)} \times \frac{P_B(b - P_B)}{(N_B - P_B)}\right] \quad \text{if } a > P_L, \quad b > P_B.$$

Average probability in the side region (along the breadth)

$$= \frac{1}{a(N_B - 2b)}\left[\frac{a^2}{(N_L - P_L)} \cdot \frac{P_B^2 - b^2}{2(N_B - P_B)} + \frac{a^2}{2(N_L - P_L)} \times \frac{P_B(N_B - 2P_B)}{(N_B - P_B)}\right] \quad \text{if } a \leqslant P_L, \quad b \leqslant P_B.$$

$$= \frac{1}{(N_B - 2b)}\left[\frac{P_L^2}{N_L - P_L} \cdot \frac{P_B^2 - b^2}{2(N_B - P_B)} + \frac{P_L(a - P_L)}{N_L - P_L} \cdot \frac{P_B^2 - b^2}{N_B - P_B} + \frac{P_L(a - P_L)}{N_L - P_L} \cdot \frac{P_B(N_B - 2P_B)}{N_B - P_B} + \frac{P_L^2}{2(N_L - P_L)} \cdot \frac{P_B(N_B - 2P_B)}{(N_B - P_B)}\right]$$

$$\text{if } a > P_L, \quad b \leqslant P_B$$

$$= \frac{1}{a(N_B - 2b)}\left[\frac{P_L^2}{2(N_L - P_L)} \cdot \frac{P_B(N_B - 2b)}{N_B - P_B} + \frac{P_L(a - P_L)}{(N_L - P_L)} \times \frac{P_B(N_B - 2b)}{(N_B - P_B)}\right] \quad \text{if } a > P_L, \quad b > P_B.$$

Average probability in central region

$$= \frac{1}{(N_L - 2a)(N_B - 2b)}\left[\frac{P_L^2 - a^2}{(N_L - P_L)} \cdot \frac{P_B^2 - b^2}{(N_B - P_B)} + \frac{P_L(N_L - 2P_L)}{(N_L - P_L)} \times \frac{P_B^2 - b^2}{N_B - P_B} + \frac{P_L^2 - a^2}{(N_L - P_L)} \cdot \frac{P_B(N_B - 2P_B)}{(N_B - P_B)} + \frac{P_L P_B(N_L - 2P_L)(N_B - 2P_B)}{(N_L - P_L)(N_B - P_B)}\right] \quad \text{if } a \leqslant P_L, \quad b \leqslant P_B$$

$$= \frac{1}{(N_L - 2a)(N_B - 2b)}\left[\frac{P_L(N_L - 2a)}{(N_L - P_L)} \cdot \frac{P_B^2 - b^2}{(N_B - P_B)} + \frac{P_L P_B(N_L - 2a)(N_B - 2P_B)}{(N_L - P_L)(N_B - P_B)}\right] \quad \text{if } a > P_L, \quad b \leqslant P_B$$

$$= \frac{1}{(N_L - 2a)(N_B - 2b)}\left[\frac{P_L P_B(N_L - 2a)(N_B - 2b)}{(N_L - P_L)(N_B - P_B)}\right] \quad \text{if } a > P_L, \quad b > P_B$$

4. Method for Equalizing Probabilities

We shall now consider a new method of selection of the plot which will considerably reduce the difference between the probabilities of inclusion of different points into the plot. This method also will make average probabilities for the points of the inclusion in a plot nearly equal in different regions, where the border regions are defined by $a = P_L$, $b = P_B$. In this method we consider different probabilities of choosing the point (x, y) which is the left-hand upper corner of the plot depending on where the point (x, y) is located in the field as given below.

Let

$$N_L > 4P_L \quad \text{and} \quad N_B > 4P_B.$$

Put

$$\text{Prob.}\,(0 \leqslant x < P_L) = \alpha P_L \cdot d,$$

where α and d are constants to be determined.

$$\text{Prob.}\,(P_L \leqslant x \leqslant N_L - 2P_L) = (N_L - 3P_L)\,d$$

$$\text{Prob.}\,(N_L - 2P_L < x \leqslant N_L - P_L) = \alpha \cdot P_L \cdot d.$$

We now impose the condition that

$$\text{Prob.}\,(0 \leqslant x \leqslant N_L - P_L) = 1,$$

i.e.,

$$2\alpha P_L \cdot d + (N_L - 3P_L)\,d = 1,$$

i.e.,

$$d = \frac{1}{N_L + P_L(2\alpha - 3)}.$$

Similarly

$$\text{Prob.}\,(0 \leqslant y < P_B) = \beta P_B \cdot d',$$

where β and d' are constants to be determined.

$$\text{Prob.}\,(P_B \leqslant y \leqslant N_B - 2P_B) = (N_B - 3P_B)\,d'$$

$$\text{Prob.}\,(N_B - 2P_B < y \leqslant N_B - P_B) = \beta \cdot P_B\,d'.$$

From the condition

$$\text{Prob.}\,(0 \leqslant y \leqslant N_B - P_B) = 1$$

we get

$$2\beta P_B \cdot d' + (N_B - 3P_B)\,d' = 1,$$

i.e.,

$$d' = \frac{1}{N_B + P_B(2\beta - 3)}.$$

Thus the probability for any point (ξ, η) in the corner region to be included in the plot is

$$\frac{\alpha\xi}{N_L + P_L(2\alpha - 3)} \cdot \frac{\beta\eta}{N_B + P_B(2\beta - 3)}$$

$$0 \leqslant \xi < P_L, \quad 0 \leqslant \eta \leqslant P_B.$$

Average probability that a point in the corner region is included in the plot is

$$= \frac{\alpha\beta P_L P_B}{4\,[N_L + P_L(2\alpha - 3)]\,[N_B + P_B(2\beta - 3)]}.$$

Probability in side region (along the length) for a point $(\xi; \eta)$ to be included in the plot

$$= \frac{\alpha(2P_L - \xi)\beta\eta + (\xi - P_L)\beta \cdot \eta}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)]}$$

when $P_L \leqslant \xi < 2P_L$, $0 \leqslant \eta < P_B$

$$= \frac{P_L\beta\eta}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)]}$$

when $2P_L \leqslant \xi \leqslant N_L - 2P_L$, $0 \leqslant \eta \leqslant P_B$, etc.

Average probability for side region (along the length)

$$= \beta \cdot \frac{P_L P_B}{2} \frac{[N_L + P_L(\alpha - 3)]}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)][N_L - 2P_L]}.$$

Similarly average probability for side region (along the breadth)

$$= \alpha \frac{P_L P_B}{2} \frac{[N_B + P_B(\beta - 3)]}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)][N_B - 2P_B]}.$$

Probability of inclusion of a point ξ, η in central region

$$= \frac{\alpha(2P_L - \xi)\beta(2P_L - \eta) + \alpha(2P_L - \xi)(\eta - P_B) + (\xi - P_L)\beta(2P_B - \eta) + (\xi - P_L)(\eta - P_B)}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)]}$$

if $P_L \leqslant \xi < 2P_L$, $P_B \leqslant \eta < 2P_B$

$$= \frac{P_L\beta(2P_B - \eta) + P_L(\eta - P_B)}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)]} \quad \text{if } 2P_L \leqslant \xi \leqslant N_L - 2P_L,$$

$P_B \leqslant \eta \leqslant 2P_B$

$$= \frac{\alpha(2P_L - \xi)P_B + (\xi - P_L)P_B}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)]} \quad \text{if } P_L \leqslant \xi < 2P_L,$$

$2P_B \leqslant \eta \leqslant N_B - 2P_B$

$$= \frac{P_L P_B}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)]} \quad \text{if } 2P_L \leqslant \xi \leqslant N_L - 2P_L,$$

$2P_B \leqslant \eta \leqslant N_B - 2P_B$, etc.

Average probability for the central region is thus

$$\frac{P_L P_B [N_L + P_L(\alpha - 3)][N_B + P_B(\beta - 3)]}{[N_L + P_L(2\alpha - 3)][N_B + P_B(2\beta - 3)](N_L - 2P_L)(N_B - 2P_B)}.$$

When

$$\alpha = \frac{2(N_L - 3P_L)}{N_L - 4P_L}, \quad \beta = \frac{2(N_B - 3P_B)}{N_B - 4P_B}$$

the average probabilities of the side and corner regions are found to be equal. With these values of α and β, the average probability in central region also comes out to be equal to that in corner or border regions.

If N_L and N_B are large in comparison with P_L and P_B respectively, then $\alpha \simeq 2$ and $\beta \simeq 2$ approximately.

Thus to give the approximately same average probability for different regions for inclusion in the plot, the above calculations show that probability for points from 0 to P_L and from $N_L - 2P_L$ to $N_L - P_L$ should be twice as the probability for points from P_L to $N_L - 2P_L$ to be chosen for x-co-ordinates of the corner point of the plot. Similarly probability for points from 0 to P_B and from $N_B - 2P_B$ to $N_B - P_B$ should be twice as the probability for points from P_B to $N_B - 2P_B$ to be chosen for y-co-ordinates of the corner point of the plot.

5. Practical Method of Selection

The practical method for choosing the different points with different probabilities as given above is as follows. Suppose $N_L = 100$, $P_L = 10$. The x-co-ordinates of the corner point of the plot will take any value from 0 to 90. Then the numbers between 0 to 9 and 81 to 90 (both inclusive) should have double probability than the numbers between 10 and 80. From all numbers (0 to 110) if we get the number 0 or 1 randomly it will fix $x = 0$ of the corner of plot. If random number comes out to be 2 or 3, it will fix $x = 1$ and so on up to $x = 9$ when random number is 18 or 19. After that the random numbers from 20 to 90 will fix $x = 10$ to 80 respectively. Again random number 91 or 92 will fix $x = 81$; 93 or 94 will fix $x = 82$ and so on up to $x = 90$ when random number is 109 or 110, *i.e.*,

Random numbers	$(x =$ $)$
(0, 1)	0
(2, 3)	1
(4, 5)	2
(18, 19)	9
(20)	10

Random numbers	($=x$)
(21)	11
(90)	80
(91, 92)	81
(93, 94)	82
(109, 110)	90

Similarly y-co-ordinate of the corner point of the plot can be determined.

6. Summary

The problem of over and under-sampling of various regions of the field as a result of the process of random location of the plot followed in yield surveys of crops is dealt in this paper. The average probabilities for points in different regions being included in the plot have been calculated. Comparison of these average probabilities shows the over-sampling of central region. To remove this bias in sampling a new method of choice of random plot in a field is suggested.

7. Acknowledgement

The authors are grateful to Dr. V. G. Panse for the keen interest he has taken during the course of the work and constant encouragement.

8. References

1. Mahalanobis, P. C. .. *Phil. Trans. Roy. Soc., London*, 1944, **231**, 41.
2. ———— .. *Nature*, 1946, **158**, 798.
3. Panse, V. G. .. *Cur. Sci.*, 1946, **15**, 218.
4. Sukhatme, P. V. .. *Ibid.*, 1946, **15**, 119.
5. ———— .. *Jour. Amer. Stat. Assoc.*, 1947, **42**, 297.