ERRORS AND BIASES IN THE ESTIMATION OF YIELD BASED ON CROP CUTTING FROM SAMPLE PLOTS

BY

ELIJAH S.K. MUWANGA-ZAKE

B.Sc., Makerere University, Kampala (Uganda), 1974

A MASTER'S REPORT

Submitted in partial fulfillment
of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics

Kansas State University

Manhattan, Kansas

1976

Approved by:

A. M. Feyerherm

Major Professor

317

TABLE OF CONTENTS

LIST OF TABLES AND DIAGRAMS

1. INTRODUCTION

The objective method of estimating a yield of a crop consists of selecting a sample or parcels or fields at random, which then are subsampled. The sample units actually harvested are plots of a prescribed dimension located and marked in the parcel (or field) according to clearly defined procedures. The usual method is to make a list of first-stage units, say, villages, in the area (stratum) to be studied. A sample of villages is selected and a list of fields growing the crop in question is prepared for each village in the sample. A sample of fields is taken and a plot is marked at random, in the selected field. The plot is harvested and the produce is weighed after it has been dried. The yield is then estimated from the data collected.

Surveys in different countries have shown that this is a practical method capable of giving yield estimates free from bias, and with a high degree of accuracy. But it is an expensive method, requiring the use of a large number of trained personnel. The method is delicate in the sense that various biases can creep into the results if attention is not paid to details. A number of steps can be taken to keep these biases under control; however, efficient action requires a thorough knowledge of the type of biases and errors encountered.

The main sources of non-sampling errors and biases in the estimation of yield based on crop-cutting from sample plots are:

1. Selection of fields
2. Border bias
3. Plot size and shape
4. Missing crop
5. Date of cutting
6. Cutting procedures and harvesting losses
7. Biases arising in the estimation procedure
8. Other sources

The purpose of this report is to examine these sources of bias to determine under what circumstances each occur; what consequences each has on the yield estimate; and what tools and techniques can be used in improving the estimate.

2.1. SELECTION OF FIELDS

Sample surveys of crop yields are usually carried out crop by crop. The reason is that different crops mature at different times. The sample for a particular crop can only be taken when the crop is mature. This means that there is only a very short period before the actual harvest during which the field is suitable for sampling. Consequently excessive difficulties could be encountered in a rigorous application of the principles of random selection. The main problem with random selection is that it would be very difficult, and hence expensive, to assemble a frame and elicit the required information from the sampled fields. This is why recourse is made to simplification and deviation from strictly random selection. Many different types of alternatives to strictly random sampling schemes are possible but only four will be considered.

(i) With this technique a sample of villages (or Crop Reporting Districts) from a suitable area of units is selected. The selectors then drive along roads traversing agricultural areas near these villages. The car is stopped at equal intervals and the field is selected in the sample which bears the crop concerned and lies nearest to the car. Unfortunately some fields may have been harvested already while others may not be ready for harvest [See sections 4 and 5]. Furthermore, all the fields growing the crop under survey do not have the same probability of selection. Those which are far from the road have no chance of being selected. In many of the "developing" countries, a large proportion of fields will lie far in the interior with no motorable roads. If their yield rates are different from those along the road, the estimates will be biased.

Analysis of the crop-cutting material collected by the Indian Statistical Institute in the three years 1947-48 to 1949-50 on winter paddy relating to 50 Police Stations in West Bengal, did not reveal any relation between yield rate of paddy and distance from the nearest roadsides [19]. Sengupta concluded, however, that this study, based on only three years data from 25 percent of Police Stations of a single Province, namely, West Bengal, with its relatively homogeneous yields contours, did not give any conclusive results. He recommended examination of data on a larger scale, for different crops over a larger number of years, before conclusions can even be tentatively drawn!

(ii) A local enumerator makes a preliminary visit to the selected fields and determines the date of harvest in consultation with the farmers concerned. The farmers are requested not to harvest the selected fields until a sample of the crop has been taken. The enumerator then visits the field on the date agreed on and takes the sample. In this way very few selected fields are missed. The disadvantage with this method is that the fields in the sample are known in advance. For fear of taxation or otherwise, the farmer may try to reduce the yield by stripping ears or by removing shoots, etc. Second, the date fixed may be inappropriate in that the harvest is delayed beyond its proper time; this may lead to loss of yield due to shredding, etc.

(iii) In this method the fields to be sampled are not known in advance to the farmers. Rather a cluster of fields is taken on the spot (say by taking a sample of households and asking them about the fields they operate) and the fields which grow the crop in question are sampled before the crop is harvested. A sample of two fields is taken and the yield is determined by subsampling the fields. The problem with this method is that the crops ready for harvest at different times will be unequally represented in the sample. [See section 5] This cannot be avoided unless the sampling is so adjusted that its distribution over time corresponds to the distribution of the actual harvesting of fields. Secondly, the interval between the time when a crop is considered ready for cutting and the actual time of harvest is not constant. The bias involved cannot be disregarded unless it can be shown that there is no correlation between the yield and this interval. In the third place, the selection of only two fields from all those judged ready for cutting at a given time will result in over-representation of the areas (villages) in which the proportion of the fields under the crop is low. This source of bias can be eliminated by taking a sample from all or a fixed proportion of the fields judged fit for cutting.

(iv) In another procedure, similar to (iii) above, a sample of households or holdings is sometimes selected and those which are in the sample are then visited and asked whether they grow a particular crop. If they do, a field is selected at random for further work. If they do not, the selected household is substituted by the nearest household growing the crop.

The main problem with this procedure is that fields belonging to holders residing in cities or outside the villages concerned have no chance of being selected. If one field is selected from each household growing the crop, unequal probabilities will be introduced into the selection of fields. This calls for weighting the estimates, which is not desirable.

Many different types of deviation from random selection of fields are possible. Unfortunately, the effects of some of these deviations on survey results may be to cause bias in data unless due care is taken.

THIS BOOK CONTAINS NUMEROUS PAGES WITH DIAGRAMS THAT ARE CROOKED COMPARED TO THE REST OF THE INFORMATION ON THE PAGE.

THIS IS AS RECEIVED FROM CUSTOMER.

2.2 BORDER BIAS

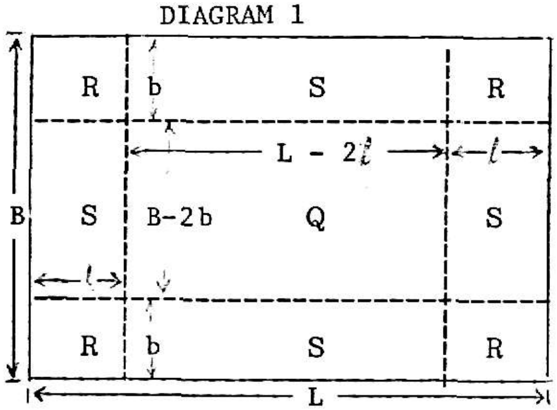
The sample cut is usually located by taking a pair of random numbers, x and y, say. The enumerator walks x paces along the length of the field and y places in the perpendicular direction. The point reached is treated as the southwest corner of the plot to be marked if the plot is a square, a rectangle, or a triangle. In the case of circular plots it is the center of the circle. If the demarcated plot partly lies outside the field it is sometimes rejected. This method does not give an equal chance of selection to the different portions of the field. 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.

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.

It is possible that near the border the yield per unit area is 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 ones. In irrigated fields the border regions may be more effectively irrigated. Competition from neighboring plants may also be less for the border regions.

Shaligram, Golhar and Ghosh (1963) [20] give probabilities of inclusion in various regions of the field. Consider a rectangular field of dimensions L x B (units of length) from which a plot of dimensions l x b is to be sample - harvested randomly. Assume that the field is relatively large compared to

DIAGRAM 1



the plot. The field is now divided into portions as indicated in Diagram 1. There are four corner regions, four side regions and one central region represented by the symbols R, S, and Q respectively. They show that average probabilities of inclusion of the central, side, and corner regions are in the proportion of 1:½:¼. To give approximately the same average probability for inclusion in the plot for different regions, they show that the probability for points from 0 to l and from $L - 2l$ to $L - l$ should be twice the probability for points from l to $L - 2l$ to be chosen for x- cordinates of the corner point of the plot. Similarly probability for points from 0 to b and from B - 2b to B - b should be twice the probability for points from b to B -2b to be chosen for y- cordinates of the corner point of the plot. Suppose $L = 100$ and $l = 10$. The x- cordinates 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 compared to those between 10 and 80. From all numbers (0 to 110) if we randomly pick the number 0 or 1 it will fix $x = 0$ of the corner plot. If a random number comes out to be 2 or 3, it will fix $x = 1$ and so on up to $x = 9$ when a 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

Random Number	(x =)
(0,1)	1
(2,3)	2
(4,5)	3
(18,19)	9
20	10
21	11
90	80
(91,92)	81
(109,110)	90

The y-cordinate of the corner point of the plot can be determined in a similar manner.

Another procedure for making a strictly random selection would be to divide the whole field into the plots of given dimensions and then draw from them at random instead of selecting the location by random coordinates. In an ideal situation, i.e.,if the size of the field is such that it is composed of an integral number of plots, the border bias will be eliminated. If the field is rectangular, but not entirely divisible into plots, the border bias will be eliminated at least along two sides of the field. But it is not practically feasible to divide the field into portions of given size.

These two procedures can only be applied if the field is a rectangle. Fields, however, are often irregular. If the plots picked on the basis of the sampled cordinates do not lie entirely within the field, the difficulties start. Namely, if they are kept in the sample all the plots will not have the same area and weighting will be needed. Since this is unacceptable on the grounds of computational complications, such plots are either rejected or pulled inside the field so that their whole area is located within the boundaries of the field. Rejection means that the border area of the field has no chance of being included in the sample. On the other hand, if plots are pulled inside, a belt of the area along the borders has a higher probability of selection than the rest of the field.

Mahalanobis (1946) suggested the use of small plot size to reduce the bias, but to be effective the plot has to be very small which would increase the variance and/or cost of survey and introduces another kind of bias [See section 3]. On the other hand, the larger the size of the sample plot, the larger the part of the field which cannot be sampled.

It may be considered, therefore, that in normal survey practice equal probability of selection of all areas within a field cannot be achieved. This leads to the question: what is the danger of the border bias? Or is the yield along the border different from the yield inside the fields?

According to Panse [2] no danger of border bias exists in India. On the other hand, in 1941 Mahalanobis conducted a yield survey of jute where data were collected on the yield of different distances from the edge of the fields selected. His results, given in Table 1 (taken from Zarkovich [6]) indicate that the yield increases going from the border of fields toward central points. G. Baptista [6] conducted experiments in Sudan and found no significant correlation between the yield and the distance of the sample plots from the border. From experiments in the eastern half of North Dakota (USA) just prior to the 1938 harvest of wheat, King and Jebe [11] wrote that "It was evident from observation that yields adjacent to the roads were lower than yields occurring farther back in the fields. This was especially noticeable in the areas where there was a heavy infestation of grasshoppers. It appeared that grasshoppers were doing more damage around the border of the field than in the center". In taking the field sample, the first 20 paces from the road were excluded!

Thus, the importance of border effects seems controversial. Further experience from other countries for different crops and years would be of great help in clarifying it. The investigator will have to be guided by his own experiments conducted under the agricultural practices prevalent in a particular country.

TABLE 1 - RATE OF YIELD OF JUTE AT DIFFERENT DISTANCES FROM THE BORDER OF THE FIELD

(Jute Yield Survey, Bengal, 1941)

Size of Cuts: 2.5 ft x 2.5 ft			Size of Cuts: 3.5 ft x 3.5 ft			Size of Cuts: 5.0 ft x 5.0 ft		
Distance in feet from the border	Average yield in maunds per acre[1]	Index	Distance in feet from the border	Average yield in maunds per acre	Index	Distance in feet from the border	Average yield in maunds per acre	Index
1.25	218	100	1.75	168	100	2.50	162	100
3.75	267	124	5.25	207	123	7.50	183	113
6.25	291	134	8.75	227	135	12.50	196	122
8.75	305	140	12.25	220	131	17.50	194	120

Footnote: [1] One maund = 37.32 kilograms or 82.28 pounds.

2.3 PLOT SIZE AND SHAPE

The tendency towards over-estimation of yield in small cuts is ascribed to a psychological bias on the part of the investigator to include unduly some of the bordering plants or tillers inside a cut. In small plots the relative proportion of plants on the perimeter is greater than in larger ones. It is, therefore, to be expected that the inclusion or exclusion of even a few plants should influence the results materially.

The effect of the plot size and shape on the estimated yield has been studied more systematically in India than in any other country. Two different groups of statisticans were mainly involved in this work, one working for the Indian Statistical Institute under the leadership of P.C. Mahalanobis and the other for the Indian Council of Agricultural Research under the guidance of P.V. Sukhatme. Some of the results of these investigations (reported in 6) are summarized in Table 2 from work on Indian Statistical Institute and Tables 3, 4, and 5 of the Indian Council of Agricultural Research (reported in 21).

Results in Table 2 do not seem to be conclusive. On the other hand, results in Tables 3, 4, and 5 show that there is a definite risk of obtaining over-estimates of the average yield with small plots. An ideal test of bias would be to compare the results of the different size plots with those obtained from harvesting the whole field. However, for Table 3 the percentages may be taken to represent biases because in earlier investigations the average yield as estimated from the largest triangle (area: 1/100th acre or 435.6 sq.ft.) when compared with that obtained from harvesting the whole field were found to give unbiased estimates [21].

An investigation by Panse [16] also supports the conclusion that the small plots give a higher yield estimate than large ones. Possibility of bias due to size of plot was recognized in the United States by King, and Jebe, [11] and again by King, McCarty and McPeek [12]. Hendricks found that yields estimated from small plots were 10-15 percent greater than the corresponding estimates derived from harvesting whole fields [9].

Comparisons to test the yield estimates from plots of different sizes and shapes made in Southern Rhodesia in 1951 and 1952 [2] showed that small circular plots with a diameter of 5 feet 3 inches (area: 1/2,000th acre or 21.78 sq.ft.), as also square plots with a side of 6 feet 7 inches (area: 1/1,000th acre or 43.56 sq. ft.) overestimated the yield seriously, the degree of overestimation approaching 100 percent, taking the yield obtained by harvesting the entire field as the basis.

TABLE 2 - THE YIELD OF AMAN AND AUS PADDY AS OBTAINED IN SAMPLE SURVEYS USING DIFFERENT SHAPES AND DIFFERENT SIZES OF PLOTS.[1]

Shape of the sample cut	Area of the cut in square feet	Percentage Yield				
		Aman paddy		Aus paddy		Jute
		1945-46	1948-49	1944-45	1945-46	1947-48
Circle Radius 2 ft.	12.6	98	106	110	99	99
Circle Radius 4 ft.	50.3	98	106	103	98	99
Circle Radius 5 ft 8 in.	100.9	99	106	100	100	99
Circle Radius 8 ft.	201.1	100	106	-	100	98
Rectangle 33 x 16½ ft.	544.5	-	100	-	-	100
Rectangle 33 x 35 ft.	1155.0	-	-	-	-	100

Footnote 1: The Yield is expressed as percentage of the average yield of the Largest Plot.

TABLE 3 - OVERESTIMATION OF YIELD WITH SMALL PLOTS

(Moradabad District, India, 1944-1945)

Size and shape of plot		Area of plot in square feet	Irrigated Wheat		Unirrigated Wheat	
			Average yield in pounds per acre	Percentage over estimation	Average yield in pounds per acre	Percentage over estimation
Equilateral Triangle	33 ft.	471.55	831.1	-	539.0	-
	16½ ft.	117.89	870.6	4.8	598.2	11.0
	8¼ ft.	29.47	961.9	15.7	664.9	23.4
Circle Radius	3 ft.	28.29	954.5	14.9	618.8	14.8
	2 ft.	12.57	1183.3	42.4	767.7	42.4

TABLE 4 – AVERAGE YIELD OF PADDY IN POUNDS PER ACRE FOR PLOTS OF DIFFERENT SIZES

Name of Subdivision		Gaya		Aurangabad		Jebanabad		Nawadab		District (Gaya)			
Size and shape of plot	Area in sq.ft.	No. of plots	Average yield lbs.	No. of plots	Average yield lbs.	No. of plots	Average yield lbs.	No. of plots	Average yield lbs.	No. of plots	Weighted average yield lbs.	St. error	Percent of over estimation
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Rectangle 33' by 16½'	544.5	70	891.98	48	1074.65	48	1096.05	40	963.57	206	991.54	49.37	-
Equilateral triangle of side 9.9'	42.4	140	941.35	96	1087.82	96	1459.75	80	924.89	412	1078.77	60.07	8.6
Isosceles right-angled triangle equal sides 5'	12.5	140	1178.33	96	1181.62	96	1472.08	80	1092.75	412	1221.12	65.01	23.1

ILLEGIBLE DOCUMENT

THE FOLLOWING DOCUMENT(S) IS OF POOR LEGIBILITY IN THE ORIGINAL

THIS IS THE BEST COPY AVAILABLE

P-474

TABLE 5 - AVERAGE YIELD OF PADDY IN LBS/ACRE FOR PLOTS OF DIFFERENT SIZES

Name of Taluka		Bander		Bezwada		Divi		Gannavaram		Gudivada		Kaikalur		District (Kistna)			
Size and shape of plot	Area in sq.ft.	No. of plots	Average yield in lbs/acre	No. of plots	Average yield in lbs/acre	No. of plots	Average yield of lbs/acre	No. of plots	Average yield in lbs/acre	No. of plots	Average yield in lbs/acre	No. of plots	Average yield in lbs/acre	No. of plots	Average yield in lbs/acre	Standard error of the average yield in lbs/acre	Percentage over estimation
Whole field		18	1314.3	18	2665.0	18	2137.7	18	2404.7	18	1855.2	18	1533.1	108	1939.2	107.3	
50x20(Links)	435.60	18	1225.7	18	2677.4	18	2202.8	18	2437.8	18	1890.3	18	1637.5	108	1954.1	105.0	.8
3' Circle	28.29	36	1179.8	36	2784.6	36	2156.5	36	2682.3	36	1927.2	36	1686.8	216	2025.9	125.8	4.5
2' Circle	12.57	36	1274.6	36	2967.6	36	2280.1	36	2961.3	36	1941.6	36	1601.3	216	2113/2	129.1	9.0
5' Equilateral triangle	11.12	36	1135.3	36	3016.0	36	2191.6	36	4271.1	36	1806.4	36	2506.3	216	2433.4	161.7	25.5

2.3.1 Analysis of Variance

Tables 6, 7 and 8, based on results reported by Sukhatme [21], give estimates of the components of variance between villages (V), between fields within villages (F) and between plots within fields (P) for different sizes and shapes of plots. The results have been expressed on a pound per acre basis to facilitate comparison between the different sizes of plots. Table 6 shows that P increases as the plot size decreases, as is to be expected. The values of V and F components, which are interpreted as measuring, respectively, the village to village and the field to field variations and which are independent of the plot size, change with the size of plot. Thus for irrigated wheat F shows a steady increase from the 33 foot triangle to the 8¼ foot triangle, thereafter apparently remaining constant. For unirrigated wheat F shows a somewhat irregular behavior with change in the size of plot, but increases on the whole as the plot size decreases. V is relatively more steady but assumes an unusually large value for the 2 foot circle in the case of unirrigated wheat. The variation in V and F is more than can be expected on the ground of sampling error. Again in Tables 7 and 8 V, F and P increase as the plot size decreases in agreement with Table 6.

It would, therefore, appear that small-size plots suffer not only from a constant bias, but also from a variable bias from field to field which results in inflated values of V and F. Small-size plots thus fail to furnish unbiased estimates of both average yield and the between villages (V) and field/village (F) variance components.

2.3.2. Effect of the Shape of Cuts

Certain standard shapes of plots have been used regularly in yield surveys based on crop cutting. The most common ones are circular, rectangular and

TABLE 6 - ESTIMATES OF VARIANCE COMPONENTS
(For wheat, Moradabad District, India, 1944-45)
(10 lbs/acre)

SOURCE	Equilateral Triangle (sides)						Circle (Radius)			
	33 ft.		16½ ft.		8¼ ft.		3 ft.		2 ft.	
	d.f.	component	d.f.	component	d.f.	component	d.f.	component	d.f.	component
Irrigated										
Villages (V)	15	642	15	505	15	924	15	683	15	684
Fields/Villages (F)	19	405	19	686	19	984	19	992	19	981
Plots/Fields/Villages (P)	39	479	39	834	39	909	78	766	78	1,445
Unirrigated										
Villages (V)	21	592	21	744	21	780	21	529	21	1,150
Fields/Villages (F)	27	308	27	390	27	182	27	561	27	610
Plots/Fields/Villages (P)	53	321	53	360	53	727	108	643	107	1,021

TABLE 7 - ESTIMATES OF VARIANCE COMPONENTS
(paddy (rice) in Gaya District-India, 1944-45)
(10 lbs/acre)

Size and shape of plots	V	F	P
I. Rectangle 33 ft. x 16½ ft. (area: 544.5 sq. ft.)	1706	1544	-
II. Equilateral triangle of side 9.9 ft. (area: 42.5 sq. ft.)	2496	2173	294
III. Isosceles right-angled triangle equal sides 5 ft. (area: 12.5 sq. ft.)	3076	2211	772

TABLE 8 - ESTIMATES OF VARIANCE COMPONENTS
(paddy in Kistna District - India, 1944-45)
(10 lbs/acre)

Size and shape of plots	V	F	P
I. Whole field harvested	3272	1940	not appl-icable
II. Rectangle 50 links x 20 links (area: 1/100th acre)	3207	1924	-
III. Circle-radius 3 ft.	4321	3190	34
IV. Circle-radius 2 ft.	8770	3634	164
V. Equilateral triangle sides 5 ft.	7297	4639	387

triangular. One of the problems of yield surveys is to determine whether the shape of plots has any effect on the yield estimate. For any given size (area) a circular plot, which inherently has minimum perimeter, should be preferable. From Table 3 it can be seen that circular and triangular plots of approximately the same area gave yields which were very similar for irrigated wheat and considerably different for the unirrigated wheat. However, these differences were not statistically significant. Experiments directed by Bose, R.C. and Roy, S.N. in 1945-46 showed that the bias due to shape becomes negligibly small in the neighborhood of 50 square feet above which no systematic tendency towards over-estimation was noticed.

Table 9, from Sukhatme's paper [21], shows that the difference in the average yield obtained from a triangular plot and a circular plot of the same size is not statistically significant. The results show that size is more important than shape in determining yield estimates.

Mahalanobis and Sengupta [14] have also studied this problem by using differently shaped plots of approximately the same size and comparing the yield estimates. The results they got, given in Table 10, indicate that a triangular shape may produce biased results.

It must be added that comparisons between the various shapes must take into account the method of demarcating the plot. If it is found that the shape of a meter square marked in the field by means of a rigid frame or a corresponding metallic hoop, does not lead to significantly different results, it cannot be concluded that the same will apply to square plots that are marked using pegs, string, measuring tapes and cross-staffs, because of differences regarding the inclusion or exclusion of border plants.

TABLE 9 - COMPARISON OF ESTIMATES OF AVERAGE YIELDS FOR DIFFERENT SIZES OF PLOTS (Moradabad District, India, 1944-45)

Area and Shape of plots compared	Irrigated Wheat				Unirrigated Wheat			
	Difference in average yields lbs/acre	Std. error of the diff.	Degrees of freedom	't'	Difference in average yield lbs./acre	Std. error of the diff.	Degrees of freedom	't'
I. 33ft.[1] vs. 16½ft.[1]	40.3	20.6	77	1.95	60.1	15.6	104	3.88*
II. 16½ft.[1] vs. 8¼ft.[1]	93.8	28.0	77	3.35*	69.9	15.6	104	4.52*
III. 8¼ft.[1] vs. 3ft.[2]	22.2	54.3	38	0.42	48.5	26.3	53	0.18
IV. 3ft.[2] vs. 2 ft.[2]	251.0	42.8	38	5.83*	149.8	32.1	53	4.63*

Notes: 1. Equilateral triangle

2. Circle
(Areas given in Table 3)

* Significant at 1 per cent level

TABLE 10 - DIFFERENCE IN THE YIELD OBTAINED BY USING DIFFERENT SHAPED PLOTS OF APPROXIMATELY THE SAME SIZE
(Aman Paddy, Bengal, India, 1945-1946)

Shape (each 12.5 sq.ft.)	Average yield expressed in percentages of the average yield of the circular cut.			
	Gouripur	Katwa	Sainthia	Combined
Circular	100.0	100.0	100.0	100.0
Triangular	115.8	125.0	123.2	123.5
Square	93.0	109.3	107.9	103.5
"Fork"[1]	91.1	100.4	108.8	103.5

Note: [1]"Fork" is a rigid appliance in the shape of a fork with two parallel prongs, each of size 3 ft. 6 in. spaced 3 ft. 6 in. apart and connected by a cross arm, at the mid-point of which a handle is fixed - this appliance enclosed a square- (area 12.5 sq.ft.)

2.3.3 Yield Surveys in Small Fields

There are crops, such as paddy and peas, which are sometimes grown in very small fields. If yield surveys based on large plots are being conducted it may not always be possible to accommodate sample plots on such fields. As a consequence small fields will be underrepresented in the sample. If there is a difference in yield between small and large fields, a bias will be introduced in the estimated yield.

The importance of this problem obviously depends upon the fragmentation of land and cultivation practices in a country. Where large fields or plantations belong to the wealthy, who have access to the modern ways of cultivation, and small ones to the peasants, it is possible for yields to be quite different.

Mahalanobis and Sengupta [14] and later Panse recommended adjusting the size of plots to the size of the fields sampled. The smaller the size of the plot the lower the proportion of fields that cannot be sampled. With sufficiently small plots the contribution to the total production from these fields that cannot be sampled becomes negligible. Therefore, such fields may be disregarded in yield surveys without any danger of biases.

In the Uganda Census of Agriculture (1965) most of the Correlations between yield and size of fields compared for 9 crops were not significant.

These experiences raise the question of what is the proper size and shape of plots for yield surveys. The problem of size and shape of plots must be considered not only from the error point of view, but also from that of practical convenience. From the operational point of view small cuts have many advantages:

(1) They can be more conveniently and quickly demarcated using a portable ridge frame.

(2) Actual field operations are simpler which is an important consideration in ensuring good quality of work. Difficulties of field operations tend to make the investigator (interviewer) careless or dishonest.

(3) They reduce the amount of undersampling of the border of the plot and the number of fields that cannot be sampled.

The efficiency of the design of a sample survey is determined by the cost in relation to the sampling error of the final result. Sample-cuts of a large size would have, individually, smaller errors of sampling but would be more expensive compared to those of a smaller size. Working with small cuts. it may be possible, for a given cost, to obtain so many more samples that the sampling errors of the final estimates become smaller than those when cuts of a larger size are used. The lower the cost of reaching an assigned level of precision (i.e. any assigned margin of error) with a given size of sample-cuts, the higher its efficiency.

TABLE 11 - RELATIVE EFFICIENCY OF SAMPLE PLOTS OF DIFFERENT SIZES

Shape and size of the plot (sq.ft)	Number of cuts needed for the same sampling error (within plots)	Average number of hours of work needed to cut the crop	Total time needed (= col. (2) x col. (3))	Total time as percentage of the standard
(1)	(2)	(3)	(4)	(5)
Circle				
12.5	3.47	0.3	1.04	67
50.3	2.19	0.6	1.31	84
100.9	1.74	0.9	1.56	100
201.1	1.39	1.4	1.95	125
Rectangle				
544.5	1.00	2.6	2.60	167

Table 11 gives the relative cost of using sample-cuts of different sizes to reach the same error within sample plots in the case of amman paddy computed by Mahalanobis and Sengupta [14]. The plot of 100.9 square feet is taken as standard and sizes of other plots are expressed as percentages. Larger plots require more time while smaller ones require less time but are subject to some bias. In this case we conclude that the plot size of 100.9 square feet is the most economical size leading to unbiased estimates.

In the yield surveys conducted by the Indian Council of Agricultural Research, plots ranging from 1/10th acre (43.56 sq.ft.) to 1/160th acre (272.25 sq.ft.), were used, depending on the crop and size of fields and Sukhatme, P.V. concludes that "... large plots of the order of 1/80th acre appear to be free from bias".

In European countries, the United States, Japan, USSR, etc. it appears very small plots, of size around one square meter (10.76 sq. ft.) are used.

Sengupta [18] points out that the bias, in addition to the size of the plot depends upon:

(1) lack of precision in the instruments used to demarcate the crop to be cut;

(2) bias in taking very small weights on a balance scale - there is a tendency to overestimate weight when they are very small;

(3) ambiguities in the definition of the plots to be cut-whether to include border crop or not, and

(4) errors of judgement on the part of the investigator (enumerator).

He concludes that circular cuts of radius 4 feet (50.3 sq.ft.) are adequate provided the field staff are reasonably trained and the instruments are accurate and dependable.

In the investigations by Hendricks, [9] (1948), provision was made in the survey to control the various biases resulting from the use of small plots by harvesting a subsample of the selected fields completely and to provide correction factors for yields obtained from small plots.

It is clear from all this that practice varies from crop to crop and country to country with respect to the size and shape of plots used in yield surveys. With the present knowledge one cannot generalize. The differences in cultivation may account for the variability in the experiences with plot size. If the crop is cultivated in rows, as is normally the case in modern agriculture, it is probably not difficult to ascertain whether each individual plant belongs to the plot or not. With a crop planted by broadcasting, the individual plants are spread all around and some of those on the border line (perimeter) will present difficulties.

As stated earlier, the overestimating bias is attributed to the general tendency (on the part of the investigator) of an over-inclusion of plants lying on the perimeters of a cut. Special experiments on different methods for the elimination of this perimeter bias carried out by the Indian Statistical Institute, are reported by Sengupta. [18] For one of the methods, called "Balanced" cuts, the investigators (enumerators) were asked to keep a separate count of the tillers that fell on the borders (the circumference of the circle) and to have them harvested and weighed separately. It was expected that an allocation of half of this quantity to the circular cut at the analysis stage would effectively remove any bias accruing on this account. Of the several types of sample cuts studied so far, this method seems to be the best and reasonably free from over-inclusion bias. It has been tried in a number of special experiments at different centers in India, by the Indian Statistical Institute [18], on paddy and wheat. The West Bengal (India - 1953-54) surveys on Jute and Aman paddy indicate that this technique can be advantageously used even in large scale operations without much additional trouble. "Balanced" cuts of radius 4 ft (area: 50.3 sq.ft.) may be considered as completely safe for paddy and other cereal crops and most pulses and oilseeds.

2.4 MISSING CROP

In carrying out yield surveys there are two main approaches from the organizational point of view, namely using moving teams or using local staff.

The moving team approach consists of establishing a moving machinery of properly trained staff who are equipped with transport and everything else needed for the work. These moving teams visit the fields selected for the purpose of cutting the crop. The ideal situation is to reach the field selected immediately before the harvest. In some cases, however, the harvest will already be over when the team reaches the selected field. Statistically speaking this is how the problem of missing information or nonresponse arises. If each team is responsible for a large area, it will hardly be possible to arrange the work in such a way that no sample field is harvested before the team comes.

In a survey on wheat in the United States [11], when the team found the selected field already cut and shocked or windrowed, they just went ahead and located the sample plot in the same way as it would be located in standing grain. The number of heads was then determined by a "stubble count" of the sample plot. The same number of heads was chosen from the nearby bundle or windrow (a deducation had to be made for the number of heads that lay on the ground in the sample plot stripped by grasshoppers. Although this method of sampling was followed in 1938 and proved practical, it was much more time-consuming than sampling of the uncut fields. Besides, it is not applicable to crops where harvesting is done by uprooting the whole plant and it is easy to introduce biases in the choice of the same number of heads.

The problem of missing crops is solved using a second approach based on the cooperation of local staff who get in touch with farmers and find out the harvesting day. In this case the crop-cutting is always done immediately before the harvest and the problem of missing information does not arise [See section on Selection of Fields].

The extent of bias due to missing crop depends upon the differential in yield between fields with an early harvest date and those with a late harvest date.

In an attempt to avoid this problem in yield surveys carried out in Sweden [6] it was decided to proceed in two stages. In the first stage, the field was selected and in agreement with the farmers concerned, the position of the plot marked on it. Afterward, the field staff remained in contact with the farmers to get information about the day of the harvest. On the basis of this information the plots selected were cut immediately before harvest. This arrangement was possible because each team was responsible for a small area.

2.5. DATE OF CUTTING

As pointed out in sections on missing crop, the ideal way of conducting yield surveys would be to cut the crop for survey purposes immediately before the harvest takes place. However, the survey team will reach some fields when the crop is not yet ripe and will be obliged to cut it as it is because it is not normally possible to go back to the same fields several times. The time lapse between the survey and the harvest may introduce a bias since the crop will be constantly growing and changing.

The results obtained in experimental yield surveys conducted in Sweden show that for some crops the difference of a couple of days between the survey cutting and the harvest may cause considerable bias. As an illustration consider the data obtained in the hay yield survey carried out in 1958. In this, and other surveys, the time of the survey cut and of the harvest was recorded and the difference between the two was found to vary between 0.5 and 2.0 days. In this particular survey the difference was 1.3 days. In order to evaluate its effect on survey data a special sample of fields was selected and a study of the growth of the crop was made in the course of the week preceding the harvest. It was found that the growth varied from one area to another. The average for the country as a whole was 3 percent per day. Bearing in mind the above time difference of 1.3 days between the sample cutting and harvest, survey data underestimated the yield of hay by approximately 4 percent [6].

A similar study was conducted in 1959. In spite of all the precautions taken to reduce the time difference between the survey and the farm haymaking to a minimum, the difference turned out to be 0.9 day. A growth study conducted along similar lines to the one in the previous year showed a growth of 1.9 percent per day. This means that the survey data underestimated the yield by 1.7 percent. The same type of data were collected in 1960 with a time difference of 1.5 days and a daily growth of 2.1 percent, making survey estimates off by 4.0 percent.

In 1938 in the Eastern half of North Dakota (U.S.A.) [11] it was found that within a single county the fields did not differ more than 7 days in date of maturity. About two-thirds of the fields did not vary more than 4 days in maturity. The greatest deviation in date of maturity within a county was due largely to varietal differences. There was a marked gradation in the date of harvest from the southern to the northern part of the state. In some areas the fields were cut when the grain was in the dough stage in order to avoid grasshopper damage. In these areas it was decided to sample fields that would otherwise have been eliminated from the sample because of immaturity. Ordinarily, if an immature field was selected for sampling, it was discarded, and a sample was taken from the nearest mature field along the route. However, immature fields were selected no more than once or twice per 100 fields sampled. Thus, the substitution of mature fields for these immature selections would make the bias from this source very slight. Data on other crops are not available. Therefore, it is not known how serious this problem is.

2.6 CUTTING PROCEDURE AND HARVESTING LOSSES

If the harvesting procedures in the survey and in the actual harvest are not strictly comparable, biases will appear in the survey data. The yield obtained from a sample cut is usually obtained under controlled conditions. This is different from the conditions under which the farmer will harvest his fields. If the crop is cut and the produce preserved so that no losses occur, the survey estimates refer to what is often called the "biological yield". The usual aim is to establish the "economic yield" i.e., the usuable part of the biological yield, after allowing for harvesting wastage and losses. Therefore, if waste is disregarded, the estimate of the yield obtained from the survey will not be comparable with the yield actually obtained by the farmers, i.e., the survey will overestimate the "economic yield".

In experimental crop-cutting surveys conducted in Sweden a study was made in order to estimate the effect of the difference in hay cutting for survey purposes and the regular hay making. This is a typical case where cutting higher or lower than the actual farmers' practice might produce considerable bias in yield estimates. The problem was studied experimentally in 1958 [6] and it was found that hay-cutting machines cut higher than did the survey field staff using shears. Owing to this difference the survey overestimated the yield of hay by 0.6 percent. A similar study made in 1959 [6] showed a bias of 4.1 percent.

Other differences in handling the crop cause similar effects. For example, the hay cut in the 1960 experimental survey [6] was dried indoors, i.e., not as the farmers dry it. It was found that this difference accounted for an overestimation of the yield of 12.8 percent.

Investigations were carried out by the Agricultural Estimates Division of Agricultural Marketing Services (United States) [26] to find to what extent the differences in estimates of yield per acre, derived from weighing small samples of corn just before harvest could be reconciled with yields reported by farmers and the official yield estimates derived from such reports. The harvesting loss was of the order of 10 percent of the estimate of yield per acre derived from sample cuts.

A number of approaches have been tried in order to estimate the "economic yield". In one approach the survey procedures are made to resemble as closely as possible the procedures actually used by farmers. Thus the enumerators are asked to harvest and dry the crop as the farmers do. The correction involved is expected to be smaller in this case, but investigations are certainly needed to determine the amount of correction. The reason is that farmers' practices are not uniform over the entire country.

In the Uganda Census of Agriculture-1965, [25] the enumerators visited the selected households approximately daily for one year. On these visits any crop harvested by the holder from previously measured sample fields since the previous visit by the enumerator would be weighted.

In experimental surveys conducted in Sweden to estimate potato losses [6], a sample of potato fields was selected and after the harvest was over, plots of 1 square meter were located at random on these fields. Within these plots potatoes left in the ground were measured and an estimate was made regarding the proportion of potatoes left by the farmers' method of harvesting and that used in statistical surveys where no crop is left. This method is concerned only with harvesting losses. There are additional losses in handling the harvested potatoes, during transport, storage, etc., which are not included here.

Another approach commonly used is to ask the farmers to harvest the fields by using the tools and machinery that they would normally use for the purpose. This is done on a sample of fields from which cuts have been taken. The difference between the two derived estimates provides an estimate of the waste. As an illustration, consider the data presented in Table 12 based on two studies in Sweden [4]. It is clear from this table that the loss may vary from crop to crop and from year to year.

TABLE 12 - ESTIMATED LOSSES IN YIELD SURVEYS
(Sweden)

	Estimated percent loss	
Crop	1958	1959
Winter Wheat	2.9	6.9
Winter Rye	6.1	4.9
Spring Wheat	3.7	8.0
Barley	6.1	6.1
Oats	6.6	11.2

Sometimes commercial yields are available from figures on marketing for a subsample of the sampled fields. These figures can be used for making a comparison with the corresponding sample yields in order to determine the correction factor or verify their accuracy.

2.7 BIASES ARISING IN THE ESTIMATION PROCEDURE

Biases arising at the estimation stage are introduced in different ways The most important of which consists in the use of an estimation procedure that does not correspond to the selection procedure.

Consider a two-stage sampling procedure with first-stage units of unequal size and simple random sampling employed at each stage, with villages (or Crop Reporting Districts) as the first-stage and crop growing fields as the second stage. Let

M_i = the number of second-stage units in the i-th first-stage unit (i = 1, 2, ---, N),

$M_o = \sum_{i=1}^{N} M_i$, the total number of second-stage units in the population,

m_i = the number of second-stage units to be selected from the i-th first-stage unit, if it is in the sample,

$m_o = \sum_{i=1}^{m} m_i$, the total number of second-stage units in the sample,

Y_{ij} = the value of the j-th second-stage unit in the i-th first-stage unit (j = 1, 2, ---, M_i; i = 1, 2, ---, N);

n = number of first-stage units in the sample,

$\bar{Y}_{i.} = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$ = the mean per second-stage unit in the i-th first-stage unit in the population (i = 1, 2, ---, N),

$$\bar{\bar{Y}}_{N.} = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_{i.}$$

$$\bar{Y}_{..} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{Y}_{i.}}{N\bar{M}} \quad \text{where} \quad N\bar{M} = \sum_{i=1}^{N} M_i$$

Several estimators of the population mean $\overline{Y}_{..}$ can be formed. Assuming one plot per field, the simplest is the unweighted arithmetic mean of the first-stage unit means in the sample within each stratum, given by

$$\overline{Y}_s = \frac{1}{n} \sum^{n} \overline{Y}_{i(m_i)}$$

where the summation runs over the first-stage units in the sample, and $\overline{Y}_{i(m_i)}$ represents the arithmetic mean of the m_i selected second-stage units in the i-th selected first-stage unit.

We have $E(\overline{Y}_s) = E\left(\frac{1}{n} \sum^{n} \overline{Y}_{i(m_i)}\right)$

$$= E\left(\frac{1}{n} \sum^{n} \overline{Y}_{i.}\right) = \frac{1}{N} \sum_{i=1}^{N} \overline{Y}_{i.} = \overline{\overline{Y}}_{N.} \neq \overline{Y}_{..}$$

The estimate $\overline{Y}_s$ is therefore a biased estimate. The bias being given by

$$\overline{\overline{Y}}_{N.} - \overline{Y}_{..} = \frac{1}{N} \sum_{i=1}^{N} \overline{Y}_{i.} - \frac{1}{N\overline{\overline{M}}} \sum_{i=1}^{N} M_i \overline{Y}_{i.}$$

$$= \frac{-1}{N\overline{\overline{M}}} \left[\sum^{N} M_i \overline{Y}_{i.} - \frac{1}{N} \left(\sum_{i=1}^{N} \overline{Y}_{i.} \right) \left(\sum_{i=1}^{N} M_i \right) \right]$$

$$= \frac{-1}{N\overline{\overline{M}}} \sum_{i=1}^{N} (M_i - \overline{M}) \left(\overline{Y}_{i.} - \overline{\overline{Y}}_{N.} \right)$$

An unbiased estimate of the bias is provided by

$$\text{Est.(Bias in } \overline{Y}_s) = \frac{-(N-1)}{N\overline{\overline{M}}(n-1)} \sum^{n} (M_i - \overline{M}_n)(\overline{Y}_{i(m_i)} - \overline{Y}_s)$$

It follows that an unbiased estimate of the population mean is given by

$$\overline{Y}_s + \text{Est. (Bias in } \overline{Y}_s)$$

The bias arises because the inequality of sizes of the first-stage units causes the probabilities of selection of the second-stage units to vary from one first-stage unit to another. Unless the M_i vary considerably and the characteristic under study is correlated with M_i, the bias may not be serious. In other words, when villages and fields are selected with equal probability by simple random sampling, as would generally be the case, the unweighted arithmetic mean ($\overline{Y}_s$) of plot yields does not give a strictly unbiased estimate of the stratum value of the yield rate, since, this process of selection gives smaller villages and smaller fields relatively a larger change of selection per unit area under the crop. The unweighted arithmetic mean would in this situation provide an acceptable estimate of the yield rate only if it is known that there is no correlation between the size of villages and of fields on one hand and yield rate on the other. Indian [2] and Ugandan [26] experience indicate that the correlation is absent, but the position needs to be examined in each individual circumstances. Where there is evidence to suggest that the yield rate is correlated with the size of the sample units, both the first-stage and second-stage units may be selected with probability proportional to their crop areas, in which case the unweighted arithmetic mean is an unbiased estimate of the yield rate. The justification for the use of the unweighted arithmetic mean in either case is the extreme simplicity of its computation and its high efficiency [See Table 13].

If, under the method of sampling with equal probability, the unweighted arithmetic mean is found to be biased, an unbiased estimate can be obtained by calculating the weighted mean given by

$$\overline{Y}'_s = \frac{1}{A} \cdot \frac{N}{n} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} A_{ij} \; Y_{ij}$$

where, A is the total area under the crop in the stratum, A_{ij} is the area of the j-th field in the i-th village.

The variance of the weighted mean is larger than that of the unweighted arithmetic mean. Percentage standard errors calculated for the two types of mean from a few yield surveys carried out in India are included in Table 13.

These results show that the unweighted arithmetic mean is far more accurate than the weighted mean calculated from the same sample and is to be preferred provided the bias in the former can be neglected.

The nature and magnitude of this bias is illustrated in Table 14 from a large scale Indian survey. This table gives a comparison of the yield estimate in large samples, based on the unweighted arithmetic mean and the weighted mean. The standard error of the difference between the two estimates is known to be of the order of 6 to 8 per cent; the table shows that not only do the differences change sign from district to district, but also their magnitude is negligible compared to their standard errors. The table is typical of the results usually obtained and shows that bias arising from the use of the unweighted arithmetic mean is negligible for all practical purposes.

Where on account of its bias the unweighted arithmetic mean does not provide a good estimate of the yield rate, an alternative to the weighted mean $\bar{Y}_s'$, which is far less efficient [See Table 13] is the ratio estimate given by

$$\bar{Y}_s'' = \frac{\sum_{i=1}^{n} A_{i.} \left(\frac{\sum_{j=1}^{m_i} A_{ij} Y_{ij}}{\sum_{j=1}^{m_i} A_{ij}} \right)}{\sum_{i=1}^{n} A_{i.}}$$

The estimate is not strictly unbiased, [See Sukhatme (1970)] but the bias diminishes rapidly as the size of the sample is increased, so that in reasonably large sample surveys the bias would be negligible. Its chief advantage lies in its accuracy, which is distinctly greater than that of the weighted mean and often approaches that of the unweighted mean. This will be clear from the percentage standard errors for the ratio estimates which are shown in Table 13.

TABLE 13 - PERCENTAGE STANDARD ERRORS OF DIFFERENT ESTIMATES OF MEAN YIELD

	Simple (unweighted) Arith. Mean	Weighted Mean	Ratio estimate
Wheat (1947-48) (Uttah Pra desh India)	3.7	14.0	4.7
Wheat (1948-49) (Delhi-India)	2.5	10.0	5.7
Cotton (1944-45)	5.5	15.0	11.3
(1945-46) (Madhya-Pradesh, India)	6.9	14.0	13.2

TABLE 14 - MEAN YIELD IN WHEAT SURVEY IN PUNJAB (INDIA) (1943-44)

District	Yield in lb per acre	
	Simple Mean	Weighted Mean
1. Amritsar	1,029	1,041
2. Gurdaspur	829	862
3. Jullundur	839	881
4. Hoshiarpur	804	796
5. Ludhiana	1,247	1,246
6. Ferozepur	1,052	1,079
7. Ambala	854	820
8. Karnal	839	868
9. Hissar	1,090	1,142
10. Rohtak	1,004	997
11. Gurgaon	766	752
State	920	927

2.8 OTHER SOURCES

Other sources of error still contribute to the error and bias in the final yield estimate. The most important of these is the sampling error, which is beyond the scope of this report.

Various tools are used in the collection of data for the estimation of yield. These include tables of random numbers, questionnaires, measuring and weighing instruments, instructions, etc. Clearly, all of these tools can be biased, which introduces errors in the data collected, and hence, in the yield estimate. For example, the procedure often applied in the selection of the sample consists of reproducing a part of the table of random numbers, which is used later in the field for drawing the sample. It is possible for the table as a whole to satisfy the criteria of randomness but for parts of it not to, so if the selected part is not random, the resulting sample will be biased. Also, for economic reasons, strings are often used as a cheap material for demarcating the plots. Before long tension may stretch them and their lengths will also be affected to some extent if they get wet. Similarly, the spring on balance scales often used for weighing will lose some of its resilience in time and this may also cause some bias.

Errors resulting from the use of biased tools tend to have a systematic character. If a question is not properly worded and gives rise to faulty interpretation, the respondents may give the same type of inaccurate answer and the resulting bias could assume considerable magnitude. This systematic character of errors makes it necessary to check the quality of the tools to be used and ensure that they are as accurate as is reasonably possible. If a table of random numbers is to be satisfactorily used, it has to satisfy the criteria of randomness as a whole as well as in parts. However, better quality of tools will increase the cost of the survey and so the problem is that of striking a balance between the increased cost and the improved quality of the information collected.

Bias may also be introduced if some of the selected farmers are not found at home or if they refuse to cooperate. There can also be errors during the processing of the data, that is when collecting, editing, coding, punching, tabulating, etc.

Although the above sources of bias have the potential of affecting survey results, in actual practice, they are kept well under control and do not contribute very much to the bias in yield estimates, especially when compared to the major sources discussed in the text of this report.

3. CONCLUSION

From a comparison of survey, Yield Estimates versus the Officical Estimates for Kansas and Oklahoma [12] it was found that the total amount of bias found in 1940 appeared to differ from that of 1939. If this is always true, it might indicate that the bias arises from several sources, some of which are more important in certain years than others.

Therefore, to ascertain more accurately the amount of bias likely to occur in a given year it will be necessary to determine the relative contribution of each of the various sources of bias. Then with each year's sampling it may be possible to estimate the relative possibility that each bias might occur and adjust for total bias accordingly.

It is clear from the preceding sections that more investigations on the problem of bias arising from the use of small plots in crop cutting surveys are highly desirable, especially under conditions where the small plot would be adopted for routine sampling on the grounds of operational suitability. In such investigations it is important that the same type of field staff as would normally be entrusted with crop cutting work should carry out the field operations connected with the location marking and harvesting of plots, after suitable training. The investigations should also be sufficiently extensive in their coverage both in regard to area, seasons and crops. In the absence of these precautions the results might prove misleading, as it is not unlikely that in the hands of highly trained investigators, as was the case in the Indian Statistical Institute and Indian Council of Agricultural Research, comparisons made in limited areas might not show any evidence of biased estimation of yield from small plots.

The estimation of yield based on crop cutting from sample plots is, therefore not a simple operation but a complex procedure with many elements that may be subject to error. To make it a success the fields must be selected at random, the sample plot must be demarcated accurately, its location should not be influenced by the appearance of the crop, and the time of crop sampling must be very near to the time of harvest of the crop. The size of the sample plot should be large enough so that the danger of bias arising from errors of demarcation is appreciably reduced. Objective tests should be made from time to time against full-scale harvestings. The method of double sampling may be used for increasing the precision of the estimate. In this method eye estimates are made on a large sample of fields. Sample harvestings are done on a subsample to calibrate the eye estimates. The procedure, however, needs adequate testing.

BIBLIOGRAPHY

A. Books

1. Cochran, W. G. Sampling Techniques. Second Edition, John Wiley and Sons, Inc., 1963.

2. Panse, V. G. Estimation of Crop Yields. United Nations Food and Agriculture Organization, Rome, 1954.

3. Raj, Des. Sampling Theory. McGraw-Hill Book Company, New York, 1968.

4. Raj, Des. The Design of Sample Surveys. McGraw-Hill Book Company, New York, 1972.

5. Sukhatme, P. V. and Sukhatme, B. V. Sampling Theory of Surveys with Applications. Asia Publishing House, London, 1970.

6. Zarkovich, S. S. Quality of Statistical Data. United Nations Food and Agriculture Organization, Rome, 1966.

B. Periodicals, Journals, Publications of Professional Organizations

7. Hansen, M. H.; Hurwitz, W. N. and Bershad, M. A. "Measurement Errors in Censuses and Surveys", Bulletin of the International Statistical Institute, Vol. 38, No. 2, 1961, pp. 359-374.

8. Hendricks, W. A. "Research on Sample Survey Procedures and Crop Estimating Methods", Estadistica, Vol. 15, 1957, pp. 346-355.

9. Hendricks, W. S. "Objective Yield Estimates of Winter Wheat, Winter Rye and Potatoes in the UK/US Zone in Germany", Frankfurt Working Paper of the Bureau of Agricultural Economics, U.S. Department of Agriculture, 1948.

10. Jessen, R. J. "Statistical Investigation of a Sample Survey for Obtaining Farm Facts", Iowa Agricultural Experiment Station Research Bulletin, No. 304, 1942, pp. 1-104.

11. King, A. J. and Jebe, E. H. "An Experiment in Pre-Harvest Sampling of Wheat", Iowa Agricultural Experiment Station Research Bulletin, No. 273, 1940.

12. King, A. J.: McCarty, D. E. and McPeek, M. "An Objective Method of Sampling Wheat Fields to Estimate Production and Quality of Wheat", Washington, D.C., U.S. Department of Agriculture Technical Bulletin, No. 814, 1942.

13. Mahalanobis, P. C. and Lahiri, D. B., "Analysis of Errors in Censuses and Surveys with Special Reference to Experience in India", Bulletin of the International Statistical Institute, Vol. 38, Part 2, 1961, pp. 401-433

14. Mahalanobis, P. C. and Sengupta, J. M. "On the Size of Sample Cuts in Crop-Cutting Experiments in the Indian Statistical Institute 1930-50", Bulletin of the International Statistical Institute, Vol. 33, Part 2, 1951, pp. 359-404.

15. Murthy, M. N. "Assessment and Control of Non-Sampling Errors in Censuses and Surveys", Sankhya, Vol. 25, Series B, 1963, pp. 263-282.

16. Panse, V. G. "Plot Size Again", Journal of the Indian Society of Agricultural Statistics, Vol. 15, 1953, pp. 151-159.

17. Panse, V. G. and Sukhatme, P. V. "Crop Surveys in India I", Journal of the Indian Society of Agricultural Statistics, Vol. 1, 1948, pp. 34-58.

18. Sengupta, J. M. "On Perimeter Bias in Sample Cuts of Small Size", Sankhya, Series B, Vol. 26, 1964, pp. 53-68.

19. Sengupta, J. M. "On the Use of Motor Vehicles in Crop-Cutting Survey and a Sampling of the Roadside Plots Only", Sankhya, Series B, Vol. 27, Parts 1 and 2, 1965, pp. 159-174.

20. Shaligram, G. C., Golhar, B. and Ghosh, M. N. "On Relative Sampling of Various Regions of the Fields", Journal of the Indian Society of Agricultural Statistics, Vol. 15, 1963, pp. 137-146.

21. Sukhatme, P. V. "The Problem of Plot Size in Large Sample Yield Surveys", Journal of American Statistical Association, Vol. 42, 1947, pp. 299-310.

22. Sukhatme, P. V. "Measurement of Observational Errors in Surveys", Review of the International Statistical Institute, Vol. 20, 1952, pp. 121-134.

23. Sukhatme, P. V. "Recent Experiences in Sample Censuses of Agriculture", Monthly Bulletin of Agricultural Economics and Statistics, Vol. 17, 1968, pp. 1-7.

24. Sukhatme, P. V. and Seth, G. R. "Non-Sampling Errors in Surveys", Journal of the Indian Society of Agricultural Statistics, Vol. 4, 1952, pp. 5-41.

C. GOVERNMENT PUBLICATIONS

25. Uganda Government, Ministry of Agriculture, Entebbe. Uganda Census of Agriculture. Vol. IV, 1966.

26. Sarle, C. F. "Adequacy and Reliability of Crop-Yield Estimates", U.S. Department of Agriculture, Technical Bulletin, 321, 1932

27. Statistical Reporting Service of the U.S. Department of Agriculture-Scope-Methods. Washington, D.C., U.S. Department of Agriculture, 1964. (U.S. Department of Agriculture Statistical Reporting Service Miscellaneous Publications No. 967)

ACKNOWLEDGEMENTS

The author wishes to express his sincere gratitude to Professor A. Feyerherm for serving as his major advisor. The constant cooperation of Dr. K. Kemp and Dr. S. K. Perng who served as members of his advisory committee and reviewed the manuscript is greatly appreciated.

I would also like to express my appreciation to the AFGRAD Secretariat of the African-American Institute, New York, New York, which funded my studies in the United States.

ERRORS AND BIASES IN THE ESTIMATION OF YIELD BASED ON CROP CUTTING FROM SAMPLE PLOTS

BY

ELIJAH S.K. MUWANGA-ZAKE

B.Sc., Makerere University, Kampala (Uganda), 1974

AN ABSTRACT OF A MASTER'S REPORT

Submitted in partial fulfillment

of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics

Kansas State University

Manhattan, Kansas

1976

ABSTRACT

In this report, the main nonsampling errors and biases that could creep into the estimation of yield based on crop cutting from sample plots, have been identified. The circumstances under which each occurs have been determined and their effects on the yield estimate examined. A number of tools and techniques that can be used to keep these biases under control and improve the data have been suggested. Results from a number of investigations on the subject have been presented.

To make the estimation of yield based on crop cutting from sample plots a success, the fields must be selected at random; the sample plot must be demarcated accurately; its location should not be influenced by the appearance of the crop; the time of crop sampling must be very near to the time of harvest of the crop; the size of the sample plot should be large enough so that inclusion or exclusion of the border plants or tiller does not greatly influence the results; the harvesting procedures in the survey and in the actual harvest must be comparable and the estimation procedure must correspond to the selection procedure.

Adjustment for total bias in the yield estimate requires the estimation of the relative probability that each bias might occur. This in turn requires the determination of the relative contribution of each of the various sources of bias, which seems to change from year to year and crop to crop.