Method of Area Frame Sampling Using Probability Proportional to Size Sampling Technique for Crops' Surveys: A Case Study in Pakistan

Abdul Qayyum[1*] and H. M. Muddasar Jamil Shera[1]

[1]*Department of Agriculture, Crop Reporting Service, Government of the Punjab, Pakistan.*

Authors' contributions

Article Information

Original Research Article

ABSTRACT

The Crops' estimates have been greatly concerned by the Government of Punjab (Pakistan) all the times. Crop Reporting Service (CRS), Agriculture Department, Punjab, as a unique and the largest statistical organization in Punjab, has been working on agricultural statistics using the sampling technique, List Frame Sampling (LFS), for conducting surveys to gather information regarding crops acreage, cost of production, crops yield and other agricultural items since 1978. The development of the rural economy in Pakistan brings new problems and challenges to the methods of agricultural statistics. The back bone of agricultural statistics is the sampling technique, LFS, in which primary sampling unit is a village and an enumerator has to survey the whole village whatever the size of the village causing an increase in non-sampling error. In spite of the sufficient area coverage, representation of population in the sample is not satisfactory. The solution is Area Frame Sampling Technique in which primary sampling unit is a Segment of a specific acreage covering maximum dimensions of the population land. In this paper a method of Area Frame Sampling (AFS) has been proposed. As most of the research papers focus on the method of AFS through Geographical

_____

*Corresponding author: E-mail: qayyum.ravian@gmail.com;

Information System (GIS) technique. But in this paper two-stage statistical sampling technique has been used to achieve the same objective in an efficient and economical way. In the first stage Probability Proportional to Size Sampling (PPS) has been used. Here size is cropped area of a Union Council (UC), the smallest geographical cluster in Punjab, Pakistan. In the second stage Simple Random Sampling (SRS) has been used. Here Primary Sampling Units are Segments of land of a village. The results show that the problem of non-representation of agricultural land is minimized and, consequently, getting better estimates in terms of precision using less amount of land data. It is recommended that this method can be extended for multiple stages of sampling and for multiple measures of sizes.

1. INTRODUCTION

Estimating accurate and reliable crops' area and yield is the prime duty of Crop Reporting Service (CRS), Agriculture Department, Punjab. Being a department of Statistics, CRS entirely depends on randomly selected sample of villages from where data are collected. Currently CRS is using LFS design for collection of data. However, the development of rural economy brings problems and new challenges of non-representation to this method. To overcome these problems, AFS method has been proposed as a replacement of LFS. To focus the importance of AFS technique, Punjab's administrative break up is given in the Appendix at the end.

A Frame or Sampling Frame is any device used to obtain observational access to the finite population of interest [1]. Frames are generally of two types: area frames and list frames. The most forthright form of frame is a list of elements of the population with appropriate contact information [2]. List frame Sampling is the conquered sampling method before 1990's using the administration units as sampling survey units. However, list frames are subject to rapid degeneration over time, which may lead to problems of undercover age and obsolete information if they are not properly maintained [3].

Area frames are the backbone to the agricultural statistics [4]. Area frames are usually made up of a hierarchy of geographical units which in turn contain units in the survey population; that is, the frame units at one level can be subdivided to form the units at the next level. An area frame is a special kind of list frame where the units on the frame are geographical areas [5]. Area frames have the advantage of providing full coverage of the target population, are duplication-free and remain up-to-date for a long time. In addition, they are ideally suited to the generation of estimates of parameters relating to land areas, such as a total cultivated area, as they enable objective measures to be taken on the ground. On the other hand, although their costs are falling, they remain expensive to use for drawing samples [4]. In addition that the presence of outliers in samples from area frames has a considerable impact on estimates [6]. If list frames are not sufficiently updated or have completeness problems, area frames provide a very useful alternative [7].

Food and Agriculture organization (FAO) conduct the agriculture survey programs in many countries by using Area-frame sampling in the Period From 1970 to 1990s To Improve Agriculture and Rural Statistics, many developed countries have established their area-frame sampling survey system based on remote sensing since 2000 [8]. It was reviewed that area-frame sampling method was most efficient and Suitable for Chinese current cropland survey [9]. An area frame used to collect data for multiple variables in one survey and provided unbiased estimates with measurable precision. Once an area frame is constructed, it can be used year after year without having to update the sampling units [10].

The main objective of the study is to develop a method of AFS at maximum economical and statistical way without initially using the GIS technique (at present shape files of UCs of Punjab are not available and we need sampling at UC level) that increases the sampling spots and better representation of crops' population in the sample covering all cultivated area pockets of a district. It is also needed in case of Crop Insurance Project based on the concept of AYII (Area Yield Index Insurance) to narrow down the unit Area of insurance from tehsil to UC to increase homogeneity in the average yield of a crop within the unit area of insurance.

2. PROPOSED METHOD

As discussed earlier List Frame (LF) and Area Frame (AF) are two available frames for the selection of a sample of villages / segments for crops' surveys using suitable sampling techniques. There are two approaches remote sensing and statistical sampling techniques of sample selection using AFS. But here we discussed only statistical sampling techniques.

2.1 Segment's Frequency in a UC

There is only ONE segment selection from a village as there is a lot of homogeneity in cropping pattern and other agro based parameters with in a village so it is redundant to select more than one unit from a village. Each UC comprises a specific number of villages and selection of number of segments is subject to the total cropped area (acres) of the UC. Number of segments to be selected from a UC is subject to the available resources of a district i.e. number of

sanctioned strength of Crop Reporters (enumerators), financial constraints, severity of district temperature (summer temperature goes to 50ºC in many districts of Punjab) and average survey working potential of a Crop Reporter per day,
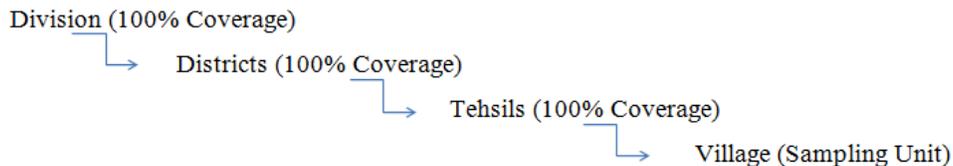
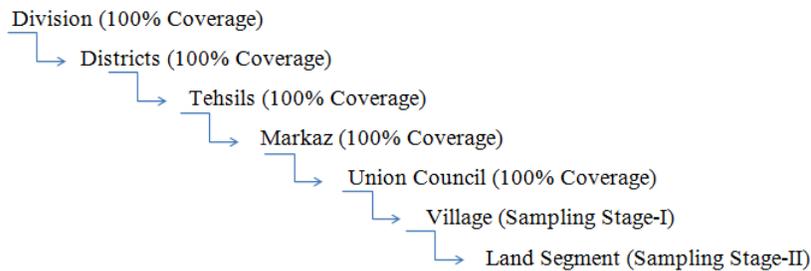| Cropped area (acres)	No of Segment in a UC
Less than 5000	1
Less than 5000	2
Less than 5000	3
Above 15,000	4

(1 acre = 198×220 ft²)

This distribution leads to assignment of maximum 5 segments per crop reporter.

2.2 Sampling Designs Using Area Frame

For better understanding and suitability of sampling design, the population's layers shown in List Frame Sampling is as

Division (100% Coverage)
→ Districts (100% Coverage)
→ Tehsils (100% Coverage)
→ Village (Sampling Unit)

On the other hand the population's layers in proposed Area Frame sampling technique are as

Division (100% Coverage)
→ Districts (100% Coverage)
→ Tehsils (100% Coverage)
→ Markaz (100% Coverage)
→ Union Council (100% Coverage)
→ Village (Sampling Stage-I)
→ Land Segment (Sampling Stage-II)

Two stage sampling techniques are being used in Area Frame Sampling. In this section two stage sampling techniques have been discussed. There is no sampling up to UC level i.e. representation of each UC is must in the sample. Sampling techniques have been used at village level and within a village at Land Segment level.

2.2.1 Sampling technique at stage – I

At Stage-I, a set of all villages of a UC is the population from which we have to select a sample of villages by using a suitable sampling technique. Each UC comprises 1 to maximum 25 villages with different areas of acreage. Some villages are significantly large and some are small. Also some villages have more area but less being used for cultivation i.e. comparatively small cropped area of a village. This fact leads to the need of implementing PPS sampling i.e. villages have more cultivated area are more likely to be selected in the sample.

2.2.2 Sampling technique at stage - II

After selection of a village at Stage-I, the available land of village being used for cultivation is the population at Stage-II from which one Segment of size 150 acres is to be sampled or selected i.e. available total land of the village is broken in to all possible Segments (Primary Sampling Unit) and one Segment is selected by using the technique of SRS because there is a lot of homogeneity in all agro based parameters and cropping pattern with in a village. So SRS is the best choice as a sampling technique at Stage-II.

2.3 Implementing Both Sampling Techniques

In this section, we briefly describe the introduction to Probability Proportional to Size sampling at Stage-I, Simple Random Sampling at Stage-II, sequence of steps in the selection of a sample and finally a case study of sample selection in the district Lodhran, province Punjab.

2.3.1 Probability proportional to size sampling (PPS) stage - I

In Probability Proportional to Size sampling technique ith element of population has a specific probability of selection P_i in the sample subject to some particular measure of size Z_i such that [11],

$$P_i = \frac{Z_i}{\sum_{i=1}^{N} Z_i}$$

where,

Z_i is the total cropped area of ith village of a U.C in acres

P_i is the probability of selection of ith village in the sample and

N is the total number of villages in a UC
Obviously for all villages of a UC,

$$\sum_{i=1}^{N} P_i = 1$$

The following steps are applied in selecting a sample using PPS sampling with in a UC for the selection of sampling units i.e. villages,

➢ Calculate P_i for each village of a UC
➢ Multiply P_i by 100 to get a whole number. (P_i×100)
➢ Construct Cumulative Probability Totals for each village.
➢ Construct Ranges for each Cumulative Probability starting from '0' to one less concerned Cumulative Probability Total.
➢ All Ranges fall within 0 to 99.
➢ Use 2-Digit Fisher's Random No. Table [12] to select a sample of required size.

2.3.2 Simple random sampling (SRS) stage - II

In Simple Random Sampling (SRS), each and every element of population has equal and known probability of selection in the sample i.e. no sampling unit is more or less favorable. It is the best technique (in terms of un-biasedness) in case of homogeneity in the variable of interest [11].

The following steps are applied in selecting a sample using SRS with in land of a selected village for the selection of sampling unit i.e. segment,

➢ Split the area of a sample village in to a Grid of Squares (1 square = 25 acres) from North to South starting No1 from North West-Corner of the village.
➢ A map of village's land designed by Board of Revenue (BoR), Government of the Punjab, is already available for each village of the Punjab
➢ Select a number between 1 and Total Area of the sample village (in squares) X (say).
➢ Select a Fisher's Random No. from 1 to X,

Ran#   Village's Area
2-Digit   Max. up to 99   squares
3-Digit   Max. up to 999   squares
4-Digit   Max. up to 9999 squares

➢ The selected Random No. is the selected sample Square and at the NW-Corner of the selected Square, take adjacent SIX SQUARES as a sample Land Segment.
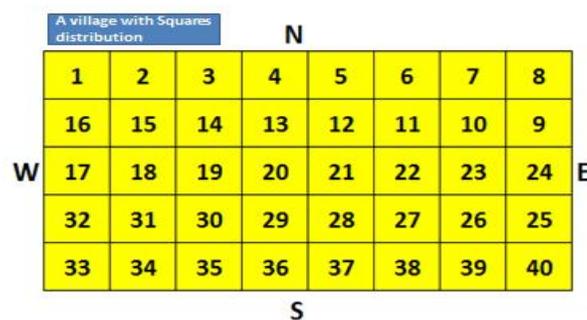
Fig. 1. A village with square distribution

Table 1. Detail of District Lodhran, Punjab

Tehsils	No. of Markaz	No. of UCs	No. of Villages	Total Area (acres)	Total Cropped Area (acres)
Lodhran	2	24	146	267,053	221,868
Dunypur	2	14	192	239,847	210,626
Karorpacka	2	20	102	184,482	160,468
Total	6	58	440	691,382	592,962

Table 2. Villages, total area (acres), cropped area (acres) of district Lodhran

District			Lodhran		
Tehsil			Karorpacka		
Markaz	UC	S.No	Village	Total Area	Cropped Area
Karorpacka	Gahi Mummar	1	Gahi Nummar	792	679
		2	Jalil Pur	446	414
		3	Darwaish Wala	265	237
		4	Mahaar	381	361
		5	Dhora Mahaar	252	224
		6	Choki Sobay Khan	439	345
		7	Jamboo Jat	871	770
		8	Chak Choki Kohna	443	411
		9	Choki Kohna	389	344
		10	Gulhar	509	464
		11	Kamboo	642	614
		12	Rindjada	3,029	2,663

3. RESULTS AND DISCUSSION

3.1 Experimental Area Frame Sample Selection

3.1.1 Sample selection of final segment

In this section a complete example of selection of a Land Segment in the tehsil Karorpacka of District Lodhran, Punjab is discussed in detail. District Lodhran consists of three Tehsils. Details are shown in the Table 1,

Taking an example of Tehsil Karorpacka, Markaz Karorpacka, this Markaz consists of 10 UCs and 54 villages. Taking an example of Gahi Mummar, one of the 10 UCs of the Markaz, this UC consists of 12 villages. As UC is our last population layer, we implement two-stage sampling on this UC level. Details of all villages of the UC, their total area (acres) and cropped area (acres) are shown in the Fig. 2.

As it is shown in the Table 2, there are 12 villages in the UC, Gahi Mummar with different cropped areas. The village, Rindjada, has the maximum cropped area i.e. 2663 acres and the most prominent village of the UC. It leads the need of PPS sampling i.e. prominent villages of a UC must have higher probability of selection in the sample.

We have to select 4 Segments from this UC i.e. we have to select 4 villages (one Segment per village) subject to the size (cropped area) of the village. For this purpose Probability of Selection P_i is calculated for each village of the UC and implementing all steps mentioned above, a sample of 4 villages (mentioned in red color) is selected as shown in the Table 2.

For selection of Random Numbers, Fisher's Random Numbers Table has been used and column selection is subject to the number of letters in the Tehsil under consideration. Like

'Karorpacka' has ten letters and column No.10 of 2-Digit Random Nos. has been used.

Because of PPS sampling, villages have more cropped area have been selected in the sample. It is a completion of Stage-I sampling i.e. sampling of villages in a UC.

At the Stage-II, Land Segments (150 acres) from the selected villages (at the Stage-I) are selected. Implementing all steps mentioned in above, a sample of a Land Segment is selected from each sample village.

Markaz	Union Council	S.No	Village	Total Area	Cropped Area	P_i	100xPi	Cumulative Totals	Range	Random No
District	Lodhran									
Tehsil	Karorpacka									
Karorpacka	Gahi Mummar	1	Gahi Mummar	792	679	0.09	9	9	00-08	5
		2	Jalil Pur	446	414	0.06	6	15	09-14	
		3	Darwaish wala	265	237	0.03	3	18	15-17	
		4	Mahaar	381	361	0.05	5	23	18-22	
		5	Dhora Mahaar	252	224	0.03	3	26	23-24	
		6	Choki Sobay Khan	439	345	0.05	5	31	25-30	
		7	Jamboo Jat	871	770	0.10	10	41	31-40	
		8	Chak Choki Kohna	443	411	0.05	5	46	41-45	
		9	Choki Kohna	389	344	0.05	5	51	46-50	
		10	Gulhar	509	464	0.06	6	57	51-56	55
		11	Kamboo	642	614	0.08	8	65	57-64	58
		12	Rindjada	3029	2663	0.35	35	100	65-99	79

Fig. 2. Probability selection of each village at UC (Stage - I)

Markaz	Union Council	S.No	Village	Total Area	Cropped Area	Random No	Total Area Squares	Selected Square No.
District	Lodhran							
Tehsil	Karorpacka							
Karorpacka	Gahi Mummar	1	Gahi Mummar	792	679	5	32	15
		2	Jalil Pur	446	414		18	
		3	Darwaish wala	265	237		11	
		4	Mahaar	381	361		15	
		5	Dhora Mahaar	252	224		10	
		6	Choki Sobay Khan	439	345		18	
		7	Jamboo Jat	871	770		35	
		8	Chak Choki Kohna	443	411		18	
		9	Choki Kohna	389	344		16	
		10	Gulhar	509	464	55	20	18
		11	Kamboo	642	614	58	26	9
		12	Rindjada	3029	2663	79	121	66

Fig. 3. Probability selection of each village at UC (Stage-II)

Taking an example of first village Gahi Mummar, It is a village of 32 squares. Selecting a Random No. from 1 to 32, Random No. 15 is selected (using 10th column of Fisher Random Numbers Table) as shown in the Fig. 3 Same process is repeated for other three villages Gulhar, Kamboo and Rindjada.

At the NW-Corner of square No.15, take upper & lower sides two adjacent squares as a sample Land Segment as shown in the Fig. 3.

The following Fig. 4 shows different expected locations of a Land Segment in case of selection of different Random Nos. between 1 and 40 (e.g. Random No. 40, 9, 6, 37).

The following Fig. 4 shows different expected locations of a Land Segment in case of selection

of different Random Nos. between 1 and 40 (e.g. Random No. 40, 9, 6, 37).

Segment of 6 squares is not allowed in a line as mentioned in the Fig. 5 i.e. Segment in a rectangular shape is compulsory as mentioned in the other cases as shown in the said Figure. Sampling at both stages reveals that there is no discretion of researcher in the selection of the sample rather it is totally randomized and, obviously, returns an unbiased picture of population. The same method is repeated for each UC of the Punjab and ultimately a random sample of required size is selected on the provincial level using AFS, which, obviously, returns a better, unbiased and more representative picture of population at the least cost.
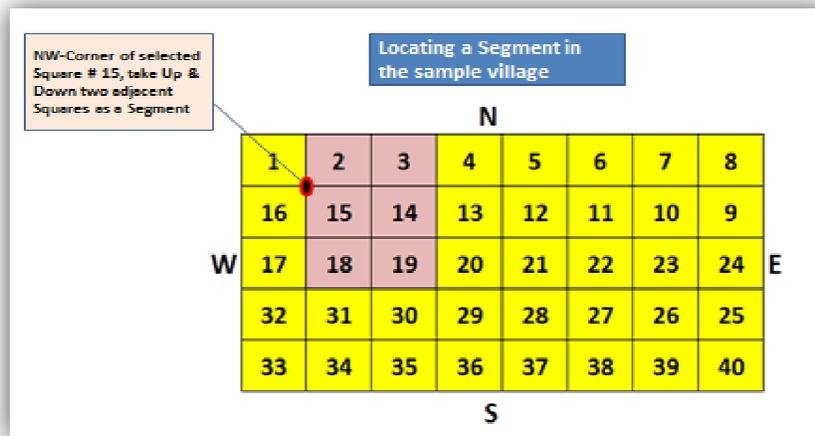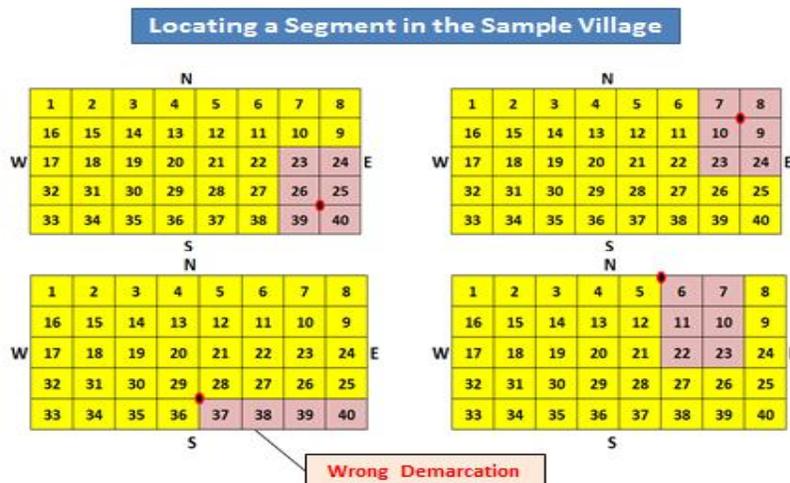

Fig. 4. Corner selection table

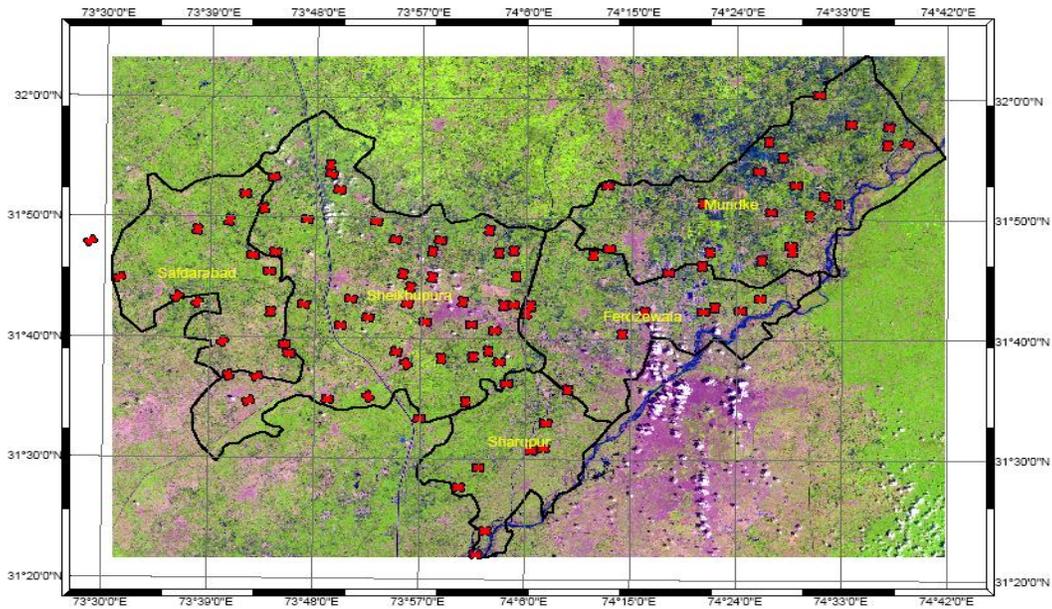**Fig. 5. Locating a segment in the sample village**

Fig. 6. Geo tagged location of all segments in the district Sheikupura, Punjab
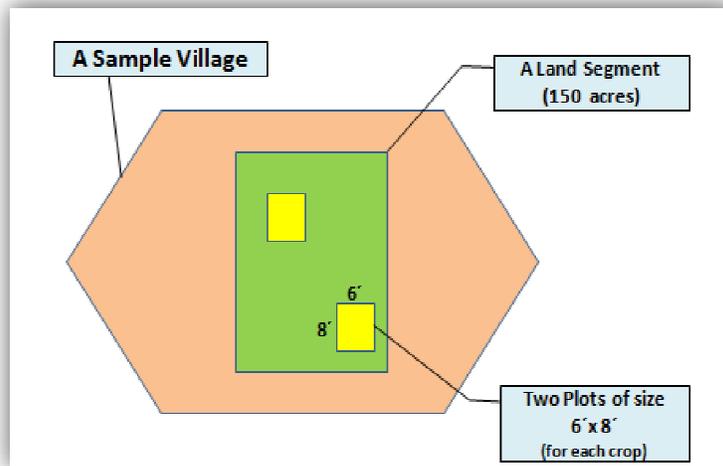

Fig. 7. Land segment

3.2 Yield Estimation in a Land Segment Using AFS

After area assessment of crops another important assessment is of yield of crops, which is included in the main responsibilities of Crop Reporting Service, Punjab.

As mentioned earlier, in each sample village six plots of size 15×20 ft² in three randomly selected fields are harvested for the yield estimation in case of LFS. But in case of AFS, number of Land Segment is significantly higher than the number of sample villages using LFS, so experimental plots of each size 6×8 ft² would be harvested for each crop as shown in the Figs. 3-7.

Obviously, as compared to 15×20 ft² experimental plot, handling of 6×8 ft² plot is easier, gives least non-sampling error and the best results in terms of least Mean Square Error (MSE) [13] among three plot sizes 6×8, 12×16, and 15×20.

4. CONCLUSIONS AND RECOMMENDA-TIONS

Most of the research papers focus on the method of AFS sampling through GIS technique i.e. the required land as a target population is divided in to a grid of Primary Sampling Units (PSUs) of equal sizes then by using SRS technique; a sample of required number of segments is selected. But in this paper, a part of GIS technique, two-stage statistical sampling technique has been used for the land of a UC (the smallest cluster of a district) subject to the condition that a UC has more cultivated area has more number of segments with in itself.

It would also be helpful in case of Crop Insurance Project based on the concept of AYII to narrow down the unit Area of insurance from tehsil to UC to increase homogeneity in the average yield of a crop within the unit area of insurance.

It is recommended that one can use multiple measures of sizes and multiple stages to implement PPS sampling technique in AFS. For instance, in this paper measure of size is only cropped area of a village. But for specific zones of some specific crops / orchards, one may be interested to include the area of a specific crop / orchard in a random sample. For instance, in Sargodha Division, Punjab, villages having more cropped area along with more citrus fruit area should have more probability of selection in the sample. Similarly for Mango zone, Multan Division is specific; for sugarcane zone, district Rahim Yar Khan is specific and for both citrus and sugarcane zone, district Toba Tek Singh is specific. Selection of a land Segment using AFS by implementing conditions of multiple stages and multiple measures of sizes through GIS technique is not as easy and economical as the method presented in this paper.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Sardnal EC, Swensson B, Wretmen J. Model assisted survey sampling, Springer; 2003.
2. Sapsford R, Jupp Victor. Data collection and analysis, 2nd Ed. SAGE; 2006.
3. Gallego FJ, Crop Area Estimation in the MARS Project [J]. Conference on ten years of the MARS Project; 1999.
4. National Agriculture statistics services, USDA, Review Report; 2009.
5. Fellegi P. Survey methods and practices, 3rd Ed. Springer; 2010.
6. Carfagna E. Multiple frame sample surveys: Advantages, disadvantages and requirements. In International Statistical Institute, proceedings – actes – invited papers seoul. ISI Publication: The Hague. 2001;253-270.
7. Javier G. Area sampling frames for agricultural and environmental statistics. Joint Research Centre, EUR 27595 EN; 2015.
8. FAO. Multiple frame agricultural surveys [C]. FAO statistical development series, Agricultural survey programs based on area frame or dual frame (area and list) sample designs, Rome. 2001:2.
9. Yaozhong P, Wang M, Wei G, Wei F, Shi K, Li L, Sun G. Application of area frame sampling for agricultural statistics in China; 2016.
10. Davies C, Cotter J, Nealon J, Roberts R. Area frame design for agricultural surveys. Agricultural Survey Methods. 2010:169-192.
11. Cochran GW. Sampling Techniques, 3rd Ed. Harvard University: John Wiley & Sons; 1977.
12. Fisher RA. Statistical tables for biological, agricultural and medical research. 6th Edition, Longman Group Limited, England. 1982;37(38):134-139.
13. Gujarati DN. Basic econometrics, 4th Ed, McGraw-Hill publishing, New York; 2003.

APPENDIX

An administrative breakup of the Punjab, Pakistan

Unit	Count	Average No. of Villages
Division	9	2920
Districts	36	730
Tehsils	147	181
Markaz	354	74
Union Council	2,797	10
Villages	26,275	-