Three Stage Diminishing Sample Size: A Rotating Sampling Procedure for Estimating Seasonal Mono-Crop Yield

AW Wijeratne[1]* and AD Ampitiyawatta[2]
[1]Department of Agribusiness Management,
[2]Department of Export Agriculture, Faculty of Agricultural Sciences, Sabaragamuwa University of Sri Lanka 70140 Belihuloya, Sri Lanka

Abstract

Rotational sampling procedures are extensively used in national surveys to reduce administrative and respondent burdens and response bias introduced by sampling the same unit in consecutive periods. In this paper, *three stage diminishing sample size* design is introduced to estimate mono-crop yield made in the same location over consecutive time periods. It is suggested that the approach proposed in this paper reduces administrative and respondent burdens and the response bias. The correlation structure of observations in consecutive time periods is used without sacrificing the precession of estimates in order to reduce the sample size in three consecutive sampling stages. The *best combined estimate* of average mono-crop yield at each consecutive period is obtained by weighing the independent estimates inversely as their variances. In this paper the complete estimation procedure is presented on the proposed rotation sampling design taking the case of paddy yield estimation which is recommended for estimating the national average production of other seasonal mono-crop yields.

Keywords: Rotational sampling, Diminishing sample size, Combined estimates, Mono-crop yield, Sample size

****Corresponding author:*** aw.wijeratne@gmail.com

Introduction

It is well comprehended that surveys at National interests in developing countries are conducted in situations of resource constraints and at varying levels of technical capacities. Common types of "sampling through time surveys are repeated, panel and longitudinal surveys, rotating panel surveys, split panel surveys and rolling samples (Steel and McLaren, 2008). There is a dire need in developing a sample survey methodology which yields efficient, unbiased estimation of up-to-date population parameters under stringent resource limitations. Rotation sampling designs have been used in sample surveys for the purpose of reducing administrative and respondent burdens and response bias introduced by sampling the same unit in consecutive periods (Cochran, 1977). According to David (1998) though past information could be routinely used to plan future data collection, expecting that correlations between the two sets continue to be strong, it has been observed that the correlations turned out weaker than anticipated, because of the rigorous changes in agriculture in developing countries and non-sampling errors in massive survey operations.

In practice, there are many other aspects of the design and implementation that must be evaluated before seriously considering a rotation design (Cantwell, 2008). They pointed out that the relative importance of level, change, and averages over time is a major factor in deciding the rotation scheme. When the measuring of change over time is more important than that of level or averages over time, a heavier overlap of the sample units may be required. When averages are desired at a given time, overlap generally is less important. They also warned that the tendency of biased response due to rotation design. It could be due to the behavioral factors of respondents.

Steel and McLaren (2008) reported the choice of rotation in terms of the impact on the estimation of levels and changes They also reviewed composite and other forms of estimators and the interaction between design and the estimation. Furthermore, they emphasized some important aspects where special considerations should be given to aspects such as (1) frequency of sampling (2) the spread and pattern of inclusion of units over time (3) the use of overlapping or non-overlapping samples over time and (4) the precise pattern of overlap.

The factors affecting the design of a sample over time would be the key estimates to be produced, the type and level of analyses to be carried out, cost, data quality and the reporting load. The common frequencies for repeated survey are monthly, quarterly and annual (Steel and McLaren, 2008). When the method of sampling becomes complex, there are many theoretical and practical issues involved in the analysis of data beyond conducting a weighted analysis and correctly estimating variances of estimators (Brogan, 2005).

Methodology

In order to elaborate the sampling design, the estimation procedure of the average paddy yield in Sri Lanka was proposed. This practice is documented in the Operational Manual of the Agriculture and Environmental Statistics Division of Department of Census and Statistics of Sri Lanka. Accordingly, the household survey on paddy is conducted in each Divisional Secretariat (DS) of all the districts of Sri Lanka. Hence we recommend districts and DSs not to be considered as stages of sampling. The sampling is applied within each DS division where stratified random sampling is carried out taking modes of irrigations as strata. The following reasons are taken in to consideration in estimating agricultural statistics at DS level.

- Primary unit of making administrative decisions,
- Every crop species may not be economically grown in each DS within a district: variance due other sources such as soil, micro climate, etc,
- Subsidy programs such as fertilizer subsidies are administered by DS, and
- Crop losses due to natural hazards are estimated and compensations are administered by DS.

The total number of the hectares to be sampled at the national survey on paddy should be determined based on the resource availability, practical feasibility and precession required. Then the number of the hectares to be sampled for each district is calculated proportionate to the total cultivated paddy extents.

Through the diminishing sample size approach, the correlation of the observation from the same paddy field in consecutive time periods is used to reduce the sample size in consecutive years of the same season (Yala to Yala, and Maha to Maha). It should be noted that the length of rotation is determined based on the correlation structure of the estimates through the time line. Once the correlation coefficient (r) between the initial observations and observations at time t drops below 0.7 (according to Cochran, 1977 the % precision gain would drop as r decreases), then we could regard (t) as the length of rotation.

Here we propose a tree stage diminishing sample size rotation design. In order to elaborate the procedure, let $t = 3$ (Table 1).

Let

$I = 3$; number of irrigation modes (Major, Minor and Rain fed),

N_I = total number of recorded cultivated paddy areas (hectare) under the mode of irrigation I,

$N = \sum_{I=1}^{3} N_I$; total number of recorded cultivated paddy areas in a given DS,

a_{Ij} = area of the j^{th} paddy field surveyed under the mode of irrigation I,

y_{Ij} = per hectare paddy yield recorded at the j^{th} paddy field surveyed under the mode of irrigation I,

p_I = number of paddy fields surveyed under the mode of irrigation I,

$\hat{\sigma}_I^2$ = estimate of the variance of paddy yield under the irrigation mode I in a given DS.

Let P_I = total number of paddy fields in the given DS under mode of irrigation I, out of which assume that $p_I = 8$ is taken at the first occasion of sampling.

In this example, the original sample will be reduced by 25% at the second occasion and 50% at the third occasion.

Then it is required to calculate the actual extents of harvested paddy fields/parcels to be sampled at the second and third occasions of sampling. Here, note that the unit of sampling is the hectare (not the paddy field/parcel). At the second occasion we need to select 25% of the initial sample.

Let

$\bar{a}_I = \frac{\sum_{j=1}^{p_I} a_{Ij}}{p_I}$ be the per acre surveyed paddy field, then we have

$\bar{a}_I p_I = \sum_{j=1}^{p_I} a_{Ij} = n_I$. But we need 25% of the initial sample, hence

$$n_{I(25\%)} = \frac{\bar{a}_I p_I}{4}.$$

Next task is to define the criteria to select some paddy fields/parcels to match the 25% initial sampled paddy harvested areas. At field conditions, it may not be perfectly possible to find paddy parcels to meet the 25% of the initial sample size. Hence we propose the following condition,

$$\sum_{j=ran\#}^{p_I k} a_{Ij} \le n_{I(25\%)}$$

Where the areas of few randomly selected paddy parcels are summed until the given inequality is satisfied. In general case the percentage reduction in consecutive sampling should be determined by examining the correlation structure through the sampling process and the available sample size for the reduction. It should be noted that, with too small initial sample size, the percentage reduction in consecutive sampling becomes restricted. The primary purpose of the reduction in sample size is to cut-down the survey cost.

After the 1st occasion of sampling (say Maha season of Year t) only 75% of the original sample size is drawn at Maha season of Year $t+1$ and 50% at Maha season of Year $t+2$. According to Cochran (1977) the maximum matched percentage never exceeds 50%. Therefore, it would be advisable to start a new rotation after the third occasion of sampling. Hence this diminishing approach is readily applicable only for three (03) stages.

According to this design, at the second stage of sampling, only 75% of the original sample size is used, viz, 25% from the previous sample and 50% of new sampling units are introduced. At third stage only 25% of new sampling units are introduced. We propose an alternative too in order to reduce the cost of survey. As Cantwell (2008) reported, in household surveys that use a rotation design, the first interview is conducted in person, which is generally more expensive, so that the interviewer can develop a rapport with the respondent. Given the respondents are familiar with the questionnaires; later interviews may be conducted over the telephone or by a local interviewer or someone in a Centralized telephone facility.

In this design, if information from the unmatched sampling units from the previous sampling occasions is collected without much effort at the second sampling occasion, the total sample size would be 150% in ideal conditions. At each occasion of sampling, if information is gathered from previous samples there would be an increase in actual enumerated sample units at each occasion of sampling. Given a reasonable time frame, initial sampling units may be retired from the design. The final sampling design and the estimation process should be designed taking the entire practical (resources, accessibility, non-sampling errors, etc) and theoretical (level of precession, properties of estimates, variance – covariance structure, etc)

Table 1: Three Stage Diminishing Sample Size for Mode of Irrigation I

Paddy field/parcel no.	**Occasion of sampling**			
	1	**2**	**3**	
1	a_{I1}	**Unmatched units**	**Unmatched units**	
2	a_{I2}			
3	a_{I3}			
4	a_{I4}			
5	a_{I5}			
6	a_{I6}			
7	a_{I7}	a_{I7}	**25% matched units for 2nd occasion**	
8	a_{I8}	a_{I8}		
9	***$p_I = 8$ no. of paddy fields surveyed under the irrigation mode I in a given DS (100% sampling)**	a_{I9}	**Unmatched units**	
10		a_{I10}		
11		a_{I11}	a_{I11}	**25% matched units for 3rd occasion**
12		a_{I12}	a_{I12}	
13		***Only 75% of p_I is sampled**	a_{I13}	***Available 25% matched units for a next occasion of sampling or new rotation**
14			a_{I14}	
. P_I			***Only 50% of p_I is sampled**	
. P_I				

Note: *for the purpose of convenience assume that the areas sampled across all the paddy parcels are equal.

aspects into consideration.

Estimation of Paddy Yield

We report the estimation equations based on the primary design proposed in Table 1. However, there is a room for modifications for a given case in general.

In a given DS, sampled harvested extent for each given mode of irrigation is calculated based on Neyman allocation, where it is assumed that the direct cost to survey a single hectare is equal among different modes of irrigations. Then we have the approximate harvested extent to be sampled due to Neyman allocation as

$n_I = \frac{nN_I\hat{\sigma}_1}{\sum_{I=1}^{3} N_I\hat{\sigma}_1}$ Where $n = \sum_{I=1}^{3} n_I$ is the total number of harvested extents surveyed in a given DS.

The number of paddy fields/parcels (p_I) to be surveyed in a given DS under the mode of irrigation I, will be determined to suite the total number of actual harvested extents to be surveyed then,

$n_I = \sum_{j=1}^{p_I} a_{Ij}$; sampled harvested extent for irrigation mode I in a given DS. At field conditions it may not always be possible to make approximate sample size equal to the actual sample size. However, possible precautions should be taken have a most reasonable sample size to match Neyman allocation.

The total number of harvested extents to be surveyed (n) in a given DS within a given district is calculated based on the proportional allocation of total number of recorded cultivated paddy areas. However, there is a minimum number of harvested extents to be surveyed at DS level (to make estimates at DS level) that will be calculated based on sampling error reported at a pilot survey, and through repeated estimates made.

However, the exact number of the hectares to be sampled in a given DS under the mode of irrigation I would be $n_I = \sum_{j=1}^{p_I} a_{Ij}$, where a_{Ij} is the area of j^{th} paddy field surveyed under the mode of irrigation I in a given DS.

The estimation process at the DS level

Sampling Stage 1

The per hectare yield (y_{Ij}) associated with paddy field/parcel j in the mode of irrigation I is calculated by dividing the total harvested yield by the harvested area of each paddy parcel.

$\bar{y}_{I1} = \frac{\sum_{j=1}^{p_I} a_{Ij}y_{Ij}}{\sum_{j=1}^{p_I} a_{Ij}}$ denotes the average per hectare yield under the mode of irrigation I in the given DS. An estimate of the variance $(\mathrm{var}(y_{Ij1}))$ of paddy yield under the irrigation mode I in a given DS is then given by

$$\hat{\sigma}^2_{I1} = \frac{\sum_{j=1}^{p_I} a_{Ij}y^2_{Ij}}{\sum_{j=1}^{p_I} a_{Ij}} - \left(\frac{\sum_{j=1}^{p_I} a_{Ij}y_{Ij}}{\sum_{j=1}^{p_I} a_{Ij}}\right)^2.$$

Then we can readily define the $\mathrm{var}(\bar{y}_{I1})$ as follows;

$\mathrm{var}(\bar{y}_{I1}) = \left(\frac{N_I - n_I}{N_I}\right)\frac{\mathrm{var}(y_{Ij1})}{n_I}$ which denotes the variance of estimated average paddy yield under the mode of irrigation I in the given DS, which is adjusted by the finite population correction factor. Then we denote the estimated total harvested paddy yield under the mode of irrigation I in the given DS as follows,

$T_{I1} = N_I\bar{y}_{I1}.$

Without loss of generality, we now define an estimate for the variance of the total harvested paddy yield under the mode of irrigation I in the given DS as follows,

$$\mathrm{var}(T_{I1}) = N_I^2\,\mathrm{var}(\bar{y}_{I1}).$$

Then the average paddy yield across three modes of irrigation due to stratified random sampling in a given DSis given by

$$\bar{y}_{D1} = \sum_{I=1}^{3}\frac{n_I}{N_I}\bar{y}_{I1}$$

Then the variance due to stratified random samplingis given by

$$\mathrm{var}(\bar{y}_{D1}) = \sum_{I=1}^{3}\left[\frac{n_I}{N_I}\right]^2\mathrm{var}(\bar{y}_{I1})$$

Then

$T_{D1} = \sum_{I=1}^{3} T_{I1}$ denotes the total paddy production in the given DS. Since the simple random samples are drawn independently across all strata, the variance of total paddy production due to stratified random sampling is obtained by

$\mathrm{var}(T_{D1}) = \sum_{I=1}^{3} \mathrm{var}(T_{I1})$ in a given DS.

Estimations due to Diminishing Sampling

Sampling Stages2, 3

First a liner regression model will be established between the 25% of the matched units and the first sample occasion and the second sample occasion. The correlation coefficient (r) will be used in order to estimate the variance at consecutive sampling occasions. According to Cochran (1977) the best combined estimate of average paddy yield at the second stage of sampling is obtained by weighing the two independent estimates inversely as their variances. Therefore, the variance due to unmatched part of sample size u_{I2} for irrigation mode I in a given DS at the second occasion of sampling is given by

$$\frac{\mathrm{var}(y_{Ij2})}{u_{I2}} = \frac{1}{W_{I2u}}.$$

Then we obtain the variance due to matched part of sample size m_{I2} for irrigation mode I in a given DS at the second occasion of sampling as follows

$$\frac{\mathrm{var}(y_{Ij2})(1-r_1^2\ )}{m_{I2}} + r_1^2 \frac{\mathrm{var}(y_{Ij2})}{0.75 n_I} = \frac{1}{W_{I2m}}.$$

Then the combined estimate of average paddy yields at the sampling occasion 2 under the irrigation mode I is given by

$$\bar{y}_{I2}^{*} = \phi_{I2u}\bar{y}_{I2u} + (1-\phi_{I2u})\bar{y}_{I2m}$$

Where $\bar{y}_{I2u}$ and $\bar{y}_{I2m}$ are average paddy yields due to unmatched and matched parts respectively, and

$$\phi_{I2u} = \frac{W_{I2u}}{W_{I2u} + W_{I2m}}.$$

Then the total paddy yields under the mode of irrigation I at the second occasion of sampling is given by

$$T_{I2} = N_I \bar{y}_{I2}^{*}.$$

The variance of total paddy yields under the mode of irrigation I at the sampling occasion 2 is given by

$$\mathrm{var}(T_{I2}) = N_I^2\ \mathrm{var}(\bar{y}_{I2}^{*}).$$

In order to estimate the yield at the third stage of sampling, we need three (03) sources of variations, variance due to unmatched part of sample size of u_{I3}, and variance due to matched part with sampling stage 2 and the variance due to yield estimate made at second stage of sampling.

Variance due to unmatched part of sample (u_{I3}) at sampling stage 3

$$\frac{\mathrm{var}(y_{I3})}{u_{I3}} = \frac{1}{W_{I3u}}$$

Variance due to matched part with sampling stage 2

$$\frac{\mathrm{var}(y_{I3})(1-r_2^2)}{m_{I3}} + r_2^2 \frac{\mathrm{var}(y_{I3})}{0.5 n_I} = \frac{1}{W_{I3m}}$$

The variance of $\bar{y}_{I2}^{*}$ (by least square theory in Cochran, 1977; pp346) is calculated by

$$\mathrm{var}(\bar{y}_{I2}^{*}) = \frac{1}{W_{I2u} + W_{I2m}} = \frac{1}{W_{I2}^{*}}.$$

Then we define the weightings to obtain the best combine estimate of $\bar{y}_{I3}^{*}$ as follows

For unmatched yield estimate

$$\phi_{I3U} = \frac{W_{I3u}}{W_{I3u} + W_{I3m} + W_{I2}^{*}}$$

For matched part

$$\phi_{I3m} = \frac{W_{I3m}}{W_{I3u} + W_{I3m} + W_{I2}^{*}}$$

For average yield estimates due to $\bar{y}_{I2}^{*}$

$$\phi_{I2}^{*} = \frac{W_{I2}^{*}}{W_{I3u} + W_{I3m} + W_{I2}^{*}}$$

Then the combined estimate of average paddy yield at the sampling stage 3 under the irrigation mode I is given by

$$\bar{y}_{I3}^{*} = \phi_{I3u}\bar{y}_{I3u} + \phi_{I3m}\bar{y}_{I3m} + \phi_{I2}^{*}\bar{y}_{I2}^{*}.$$

Then the variance of $\bar{Y}_{I3}^{*}$ is obtained by

$$\mathrm{var}(\bar{y}_{I3}^{*}) = \frac{1}{W_{I3u} + W_{I3m} + W_{I2}^{*}} = \frac{1}{W_{I3}^{*}}.$$

Then the total yields for a given occasion is calculated by multiplying the estimated average with the total area harvested under a given mode of irrigation at the given DS division. The total paddy production under the irrigation mode I at sampling occasion 3 in a given DS is obtained by

$$T_{I3} = A_I \bar{y}^{*}_{I3}.$$

The variance of total paddy yields under the mode of irrigation I at the sampling occasion 3 is given by

$$\text{var}(T_{I3}) = N_I^2 \text{var}(\bar{y}^{*}_{I3}).$$

The estimates of average yield, total yield and variances due to stratified random sampling at DS level are obtained through the equations given under sampling stage 1.

The estimation of Districts and National Statistics

The total paddy production in each district under given mode of irrigation is obtained by simply summing up the total yields estimated at DS level.

Then the total paddy production at national level is obtained by summing up district totals across each mode of irrigation. The variance calculations at district totals and national totals are straight forward, which is the summation of variances at each level.

References

Brogan D 2005 Chapter XXI: Sampling error estimation for survey data, Household Sample Surveys in Developing and Transition Countries, United Nations Statistics Division, Department for Economic and Social Affairs.

Cantwell PJ 2008 Rotation Designs and Composite Estimation in Sample Surveys Part 1, Motivating Their Use, Statistical Research Division U.S. Census Bureau Washington, D.C.

Cochran WG 1977 *Sampling Techniques*, third edition. John Wiley and Sons, New York.

David I P 1998 Sampling Strategy for Agriculture Censuses and Surveys in Developing Countries, International Conference on Agriculture Statistics, Washington, D.C.

Steel D and McLaren C 2008 Working Paper: Design and Analysis of Repeated Surveys, Centre for Statistical and Survey Methodology, University of Wollongong.