ESTIMATION OF COTTON YIELD AT TEHSIL LEVEL USING DOUBLE SAMPLING APPROACH UNDER STRATIFIED TWO STAGE SAMPLING DESIGN FRAMEWORK

PRAMOD KUMAR MOURY & TAUQUEER AHMAD

Division of Sample Surveys, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

ABSTRACT

General Crop Estimation Surveys (GCES) Scheme is adopted in all the states of country to estimate crop yield at higher level (state, district). With progress in planning in agriculture, especially in case of the cotton crop, we need estimate of cotton crop yield at tehsil/block level with the desired degree of precision. Application of GCES as such, as tehsil/block level with the same number of crop cutting experiments (CCEs) may yield estimates with less degree of precision. If the simple crop-cutting approach is to be adopted directly for this purpose, the present number of crop cutting experiments will have to be increased significantly. In such case, use of information getting from auxiliary variable correlated with variable under study may increase the degree of precision of estimates at tehsil level. In this study, double sampling regression approach under stratified two stage sampling design framework has been proposed for estimation of the average yield of cotton at the tehsil level using the picking having highest correlation with total yield as auxiliary variable.

KEYWORDS: *GCES, Auxiliary Variable, Crop Cutting Experiments, Double Sampling & Two-Stage Sampling*

<div style="text-align:right">***Original Article***</div>

INTRODUCTION

Cotton is often referred to be an important fiber yielding crop of global importance, which is grown in tropical and subtropical regions of more than 80 countries of the world. Cotton crop is harvested in the form of a number of pickings. Harvesting continues for a long time interval. The total number of pickings may vary from state to state. It varies from 2-3 pickings to 10 pickings. In irrigated crop, pickings are over in 3/4 pickings and crop is uprooted to accommodate the second crop whereas, in rain-fed areas the pickings are staggered over a period of time and depending on advent of rain when more flashes appear in the crop.

Presently the estimates of the total yield of cotton are being obtained on the basis of scientifically designed crop cutting experiments (CCEs) conducted under the GCES. The sampling design followed under GCES scheme for crop cutting experiments in the states is stratified three stage random sampling with mandals/taluks/revenue inspector circles/blocks/tehsils as strata, villages within the stratum as first stage units (fsu), fields/survey numbers in the selected village as the second stage units (ssu) and plot of specified size within the selected field as third stage units. Since the third stage units (tsu) i.e. ultimate units of sampling are selected from the selected second stage units, i.e. plot of specified size within the selected field is selected as third stage unit, the sampling design adopted for crop cutting experiments may be considered as stratified two stage random sampling.

To increase in planning specially in the cotton crop, we need yield estimates for small areas. The GCES scheme is used to estimate the yield of cotton crop at higher level (state, district). If we apply GCES for yield estimation at the tehsil level with the same number of crop cutting experiments as in case of GCES then we may not get the desired degree of precision of estimate of the cotton crop. To get the desired degree of precision of estimate of the cotton crop, we have to conduct more number of crop cutting experiments in tehsil as compared to GCES. Thus, we have to incur more cost per tehsil as compared to GCES scheme. By incurring a little extra expenditure involved in collecting data on the auxiliary variable, it is possible to obtain reliable estimates of crop yield at the lowest/block level without disturbing the existing structure of data collection being used in the official statistical system.

Estimation of unknown mean $\overline{Y}$ of the population more efficiently may be done through the use of auxiliary variable (X) which is correlated with variable under study and whose mean $\overline{X}$ is known. In this case, it is estimated using double sampling regression type of estimators. Double sampling is a sampling method which uses auxiliary information for improving the precision of the estimator. In double sampling approach, a large sample of units is drawn to obtain auxiliary information, and then a second sample is drawn where information on auxiliary variable and variable of interest are observed. The regression estimator utilizes information obtained from auxiliary variable.

Ghosh (1948) gave linear regression estimator in double sampling using many auxiliary variables. Sukhatme and Panse (1951) investigated the magnitude of the bias from the data collected in district crop cutting surveys and had shown that bias is negligible. Utilizing the information on eye estimated yields and using the double sampling technique, ratio and regression estimates were obtained for the mean yield per block. Olkin (1958) was the first to consider the multi-auxiliary variables in building ratio estimator. Method of estimation suggested by Olkin was extended by Rao and Madholkar (1967) to a case where some of the auxiliary variables are positively correlated with the character under study and some are negatively correlated with the character under study. Sukhatme and Koshal (1959) discussed the ratio method of estimation in case of a single auxiliary variable with unknown population mean for multi-stage sampling design. Goswami and Sukhatme (1965) extended their results to several auxiliary variables with unknown mean for a three-stage sampling design. Panse *et al*. (1966) adopted the double sampling approach for estimation of block yield by treating farmer's eye estimate as auxiliary character and crop cut estimate as character under study. Srivastava (1971) studied the case when some of auxiliary variables are positively correlated and some are negatively correlated. The ICAR-IASRI initiated a pilot sample survey for estimating yield of cotton in Hissar (Haryana) during 1976-77 with the objective of developing sampling methodology for estimating the yield of cotton and suggesting a procedure for building up advance estimates of the yield of cotton on the basis of few pickings. Ahmad and Kathuria (2010) adopted double sampling approach to obtain the most reliable estimates of crop yield at the block level using farmer's eye estimate and area of the fields as auxiliary variables. Isabel Molina and J.N.K. Rao (2010) has shown some aspects of poverty estimation at small area level. Ahmad *et al.* (2013) developed an alternative sampling methodology for estimation of the average yield of cotton using double sampling technique under stratified two stage sampling design framework under the study entitled "Study to develop an alternative methodology for estimation of Cotton production of India" funded by DES, Ministry of Agriculture, Govt. of India. Saegusa (2015) developed the variance estimation procedure using the bootstrap under two-phase stratified sampling without replacement. Asghar *et al.* (2017) proposed a multivariate regression-cum-exponential type estimator for estimating a vector of population variance using multi-auxiliary variables in two-phase sampling.

In this paper, an attempt has been made to obtain the reliable estimates at tehsil level using double sampling

regression approach under stratified two stage sampling design framework treating revenue circles as a strata, villages as first stage units, fields growing cotton crop within the selected village as second stage units and a randomly located plot of standard size in the selected field as the ultimate sampling unit and considering the picking having highest correlation with total yield of cotton crop as auxiliary variable.

DATA

Estimation of yield of cotton at tehsil level along with percentage standard error (%S.E.) using double sampling regression approach under stratified two stage sampling design framework has been carried out for tehsils of two districts namely, Amravati and Aurangabad districts of Maharashtra State for the year 2012-2013. The data available in the division of Sample Surveys, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, under the project entitled "Study to develop an alternative methodology for estimation of cotton production" has been used for this study.

METHODOLOGY FOR ESTIMATION OF AVERAGE YIELD OF COTTON AT TEHSIL LEVEL

The proposed estimation procedure for estimation of the average yield of cotton at the tehsil level is double sampling regression approach under stratified two stage sampling design framework using the picking having highest correlation with total yield as auxiliary variable. It has been observed during the previous study carried out by Ahmad *et al.* (2013) that third picking of the cotton crop in Maharashtra State has the highest correlation with total yield of cotton crop. Thus, by exploiting correlation between the data on picking having the highest correlation with total yield and the available CCE data, it is possible to obtain a precise estimate of crop production using less number crop cutting experiments than GCES. Therefore, to overcome these problems, we used third picking of cotton crop as an auxiliary variable to obtain estimates at the tehsil level of cotton crop with an acceptable degree of precision. Thus, by incurring a little extra expenditure involved in collecting data on the auxiliary variable it is possible to obtain reliable estimates of crop yield at the lower/block/tehsil level without disturbing the existing structure of data collection being used in the official statistical system. This study aims at generating precise estimates of crop yield at the block/tehsil level by exploiting the correlation between the total crop yield and yield of the picking (auxiliary variable) having the highest correlation with total yield of cotton crop.

Under the proposed approach, from each stratum, a large sample of villages has been selected by Simple Random Sampling without Replacement (SRSWOR) for observing the yield of picking having the highest correlation with total yield. This preliminary sample of villages under the proposed double sampling approach under stratified two stage sampling design framework is the same number of villages selected under GCES scheme. From each selected village, four fields growing cotton have been selected by SRSWOR for observing yield pertaining to one picking (auxiliary variable, i.e. third picking) with the help of the existing field agency of State Government., if possible, or through full time ad-hoc staff. The proposed design requires CCE data from two additional fields from the selected CCE village besides CCE data from two fields of the same selected village. Out of these preliminary sample villages, a sub-sample of villages has been selected for observing yield pertaining to the remaining pickings. Out of the four fields selected for one picking, in each of these sub-sample villages, two fields have been selected for observing yield pertaining to the remaining pickings from CCE plot. The data on sub-sample pertaining to remaining pickings was collected with the help of the existing field agency of State Government as the data were to be collected from the significantly reduced number of fields under GCES set up.

The estimation procedure for estimation of the average yield of cotton using double sampling regression approach

under stratified two stage sampling design framework is as under:

Let

L = Number of strata (revenue circles) in a tehsil

N_h = Total number of first stage units (fsu's), villages, in h^{th} stratum (h=1,2,…,L)

n_h' = Number of fsu's selected randomly in h^{th} stratum for observing yield for the pth picking (p=1,2,…,P)

n_h = Size of sub-sample selected randomly in h^{th} stratum for observing yield for the remaining pickings

m' = Number of second stage units (ssu's), fields, selected for observing yield for the pth picking in i^{th} (i=1,2,…, n_h') village of hth stratum

m = size of sub-sample for observing yield for remaining pickings of these m' ssu's

$y_{hij}(p)$ = yield of cotton in j^{th} field (j=1,2,…, m') of ith village in h^{th} stratum corresponding to p^{th} picking

Estimator of average yield corresponding to p^{th} picking, $\bar{y}_{nm}(p)$ for a tehsil is given by

$$\bar{y}_{nm}(p) = \frac{1}{n_0 m} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m} y_{hij}(p) \text{ Where } n_0 = \sum_{h=1}^{L} n_h$$

Estimate of average yield, $\bar{y}_{nm}$, for a tehsil is given by

$$\bar{y}_{nm} = \sum_{p=1}^{P} \bar{y}_{nm}(p)$$

A double sampling regression estimator for estimation of average yield of cotton under the proposed framework can be written using double sampling procedure:

$$\bar{y}_{ld} = \bar{y}_{nm} + \hat{\beta}[\bar{y}_{n'm'}(p) - \bar{y}_{nm}(p)]$$

Where $\bar{y}_{n'm'}(p) = \frac{1}{n_0' m'} \sum_{h=1}^{L} \sum_{i=1}^{n_h'} \sum_{j=1}^{m'} y_{hij}(p)$, $n_0' = \sum_{h=1}^{L} n_h'$,

$$\hat{\beta} = \frac{(\frac{1}{n_0} - \frac{1}{n_0'})s_{by(p)y} + \frac{1}{n_0'}(\frac{1}{m} - \frac{1}{m'})s_{wy(p)y}}{(\frac{1}{n_0} - \frac{1}{n_0'})s^2_{by(p)} + \frac{1}{n_0'}(\frac{1}{m} - \frac{1}{m'})s^2_{wy(p)}}$$

Where

$$s_{by(p)y} = \frac{1}{L} \sum_{h=1}^{L} \frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} (\bar{y}_{hi.(p)} - \bar{y}_{h..(p)})(\bar{y}_{hi.} - \bar{y}_{h..})$$

$$s_{wy(p)y} = \frac{1}{n_O(m-1)} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m} (y_{hij(p)} - \bar{y}_{hi.(p)})(y_{hij} - \bar{y}_{hi.})$$

Where

$$\bar{y}_{hi.}(p) = \frac{1}{m} \sum_{j=1}^{m} y_{hij}(p)$$

$$\bar{y}_{h..}(p) = \frac{1}{n_h} \sum_{i=1}^{n_h} \bar{y}_{hi.}(p)$$

$$s_{by(p)}^2 = \frac{1}{L} \sum_{h=1}^{L} \frac{1}{(n_h-1)} \sum_{i=1}^{n_h} (\bar{y}_{hi.}(p) - \bar{y}_{h..}(p))^2$$

$$s_{wy(p)}^2 = \frac{1}{n_O(m-1)} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m} (y_{hij}(p) - \bar{y}_{hi.}(p))^2$$

Estimator of $MSE(\bar{y}_{ld})$ is given by

$$\hat{MSE}(\bar{y}_{ld}) = \left\{ \left(\frac{1}{n_0} - \frac{1}{N_0} \right) s_{by}^2 + \frac{1}{N_0 m} s_{wy}^2 \right\} (1 - r^{*2}) + r^{*2} \left\{ \left(\frac{1}{n_0'} - \frac{1}{N_0} \right) s_{by}^2 + \left[\frac{1}{N_0 m} - \frac{1}{n_0'} \left(\frac{1}{m} - \frac{1}{m'} \right) s_{wy}^2 \right] \right\}$$

Where $N_0 = L\bar{N}$ and $\bar{N}$ is the harmonic mean of N_{hs}. This is the usual estimate in sub-sampling design.

$$s_{by}^2 = \frac{1}{L} \sum_{h=1}^{L} \frac{1}{(n_h-1)} \sum_{i=1}^{n_h} (\bar{y}_{hi.} - \bar{y}_{h..})^2$$

$$s_{wy}^2 = \frac{1}{n_0(m-1)} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m} (y_{hij} - \bar{y}_{hi.})^2$$

$$r^{*2} = \frac{q_{y(p)y}^2}{q_{y(p)y(p)} q_{yy}} \text{ With}$$

$$q_{y(p)y} = \left(\frac{1}{n_0} - \frac{1}{n_0'} \right) s_{by(p)y} + \frac{1}{n_0} \left(\frac{1}{m} - \frac{1}{m'} \right) s_{wy(p)y}$$

$$q_{y(p)y(p)} = \left(\frac{1}{n_0} - \frac{1}{n_0'} \right) s_{by(p)}^2 + \frac{1}{n_0} \left(\frac{1}{m} - \frac{1}{m'} \right) s_{wy(p)}^2$$

$$q_{yy} = (\frac{1}{n_0} - \frac{1}{n_0'})s_{by}^2 + \frac{1}{n_0}(\frac{1}{m} - \frac{1}{m'})s_{wy}^2$$

In this investigation, we consider that within each stratum (revenue circle), some villages have been selected and these selected villages are the selected CCE villages. In order to apply double sampling regression approach under stratified two stage sampling design framework under a new setup, the selected CCE villages in revenue circle have been treated as preliminary sample of villages (n') for third picking yield. From each selected village 4 fields growing the cotton crop have been selected by SRSWOR for recording third picking yield. A sub-sample of 2 villages has been selected from preliminary sample villages for observing the remaining pickings. Out of 4 fields (m') selected for third picking yield, in each of these 2 villages, 2 fields have been selected for obtaining harvested yield from a randomly located plot of standard size in each field. The estimates of average yield of cotton along with % S.E. have been obtained for all the tehsils of Amravati and Aurangabad districts of Maharashtra State for the year 2012-2013 using the estimation procedure discussed above with the available data in the division of Sample Surveys, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, under the project entitled "Study to develop an alternative methodology for estimation of cotton production".

RESULTS AND DISCUSSIONS

The data for all the tehsils of Amravati and Aurangabad districts of Maharashtra State for the year 2012-2013 was analyzed as per proposed estimation procedure. The results of tehsil-wise analysis are presented in Table 1 and Table 2.

Table 1: Estimates of Average Yield of Cotton (kg/ha) at Tehsil Level along with Percentage Standard Error for Amravati District of Maharashtra State for the year 2012-13

S. No.	Tehsil	Cotton Yield (kg/ha)	%S.E.
1	Amaravati	691.40	9.63
2	Bhatkuli	589.11	7.81
3	Nangaon	515.02	5.03
4	Chandur railway	807.25	3.24
5	Dhamangaon	657.05	9.07
6	Morshi	760.37	7.18
7	Tiosa	553.75	9.64
8	Warud	652.41	4.51
9	Chandur Bazar	622.80	6.30
10	Daryapur	534.91	4.51
11	Anjangaon	641.80	6.79
12	Achalpur	583.80	7.28
13	Dharni	316.64	4.96
14	Chikhaldara	323.45	5.91

Table 2: Estimates of Average Yield of Cotton (kg/ha) at Tehsil Level along with Percentage Standard Error for Aurangabad District of Maharashtra State for the Year 2012-13

S. No.	Tehsil	Cotton Yield (kg/ha)	% S.E
1	Aurangabad	122.02	4.27
2	Phulambri	207.12	9.91
3	Paithan	172.54	8.06
4	Gangapur	130.88	9.36

Table 2: Contd.,			
5	Vaijapur	183.89	9.84
6	Kannad	261.88	10.18
7	Khuldabad	241.96	4.47
8	Sillod	142.10	8.01
9	Soygaon	268.58	4.079

It can be observed from Table 1 and Table 2 that in the case of the proposed methodology, namely, double sampling regression approach under a stratified two stage sampling design framework using third picking of cotton crop as auxiliary variable, the estimates of the average yield of cotton at the tehsil level has been obtained with less than 10% standard error for all the tissues of Amravati and Aurangabad districts, except canned tehsil of Aurangabad district. The %S.E. obtained for kannad tehsil is 10.18%, which is also around 10% only and is fairly reliable at tehsil level. The estimate with 10% standard error is considered reliable even at district level.

CONCLUSIONS

The results of the investigation demonstrate the feasibility of obtaining estimates of cotton crop yield at the tehsil level by adopting double sampling regression approach under stratified two stage sampling design framework using the picking having highest correlation with total yield of cotton crop as auxiliary variable. The method combines the information obtained from the third picking of cotton crop from a large number of villages (CCE villages) with yield of cotton crop obtained from sub-sample of these CCE villages. The estimates of the average yield of cotton at the tehsil level have been obtained with less than 10% standard error for tehsils of Amravati and Aurangabad districts. Therefore, it is recommended that double sampling regression procedure under stratified two stage sampling design framework using the picking having highest correlation with total yield of cotton crop as an auxiliary variable may be adopted for estimation of the average yield of cotton at lower level i.e. tehsil/Mandal/block/talk which will save the cost of the survey significantly.

ACKNOWLEDGMENT

REFERENCES

1. *Ahmad, T., Bhatia, V.K., Sud, U.C., Rai, A., & Sahoo, P.M. (2013). Study to develop an alternative methodology for estimation of cotton production. Project Report, IASRI publication*

2. *Ahmad, T., & Kathuria, O. P. (2010). Estimation of crop yield at block level. Adv. Appl. Res., 2(2), 164-172*

3. *Ghosh, B. (1948). Double sampling with many auxiliary variates. Cal. Statist. Assoc. Bull., 1, 91-93*

4. *Goswami, J.N., & Sukhatme, B.V. (1965). Ratio method of estimation in multi-phase sampling with several auxiliary variables. J. Ind. Soc. Agril. Statist., 17, 83-103*

5. *Raghuvir Singh, Sam A Masih, Vipin Prasad, Sobita Simon & Mercy Devasahayam, Bt Cotton Plant Cultivation at Low Temperature does Not Affect Horizontal Gene Transfer to Soil Bacteria but Decreases Seed Cotton Yield, International Journal of Agricultural Science and Research (IJASR), Volume 5, Issue 3, June 2015, pp. 7-16*

6. *Isabel Molina, & J.N.K. Rao (2010). Small area estimation of poverty indicators, Cand. J. Statist., 38, 369-385*

7. *Olkin, I. (1958). Multivariate ratio estimation for finite populations. Biometrika, 45(1/2), 154-165*

8. *Panse, V.G., Rajagopalan, M., & Pillai, S.S. (1966). Estimation of crop yields for small areas. Biometrics,* **22(2)**, *374-384*

9. *Raheja, S.K., Goel, B.B.P.S., Mehrotra, P.C., & Rustogi, V.S. (1977). Pilot sample survey for estimating yield of cotton in Hissar (Haryana). Project Report, IASRI publication*

10. *Rao, P. S., & Mudholkar, G. S. (1967). Generalized multivariate estimator for the mean of finite populations. J. Amer. Statist. Assoc.,* **62(319)**, *1009-1012*

11. *Saegusa, T. (2015). Variance Estimation under Two-Phase Sampling. Scan. J. Statist.,* **42(4)**, *1078-1091*

12. *Srivastava, S.K. (1971). Generalized estimator for mean of a finite population using multi-auxiliary information. J. Amer. Statist. Assoc.,* **66**, *404-407*

13. *Sukhatme, B.V., & Koshal, R.S. (1959). A contribution to double sampling. J. Ind. Soc. Agril. Statist.,* **9**, *128-144*

14. *Sukhatme, P.V., & Panse, V.G. (1951). Crop surveys in India, II. J. Ind. Soc. Agril. Statist.,* **3**, *97-168*

15. *Asghar, A., Sanaullah, A., & Hanif, M. (2017). A multivariate regression-cum-exponential estimator for population variance vector in two phase sampling. J. King Saud Uni.-Sci. Accepted for publication*