Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(2) 2012 321-326

Crop Yield Estimation at District Level using Improvement of Crop Statistics Scheme Data - An Application of Small Area Estimation Technique

U.C. Sud[1*], Hukum Chandra[1] and A.K. Srivastava[2]
[1]*Indian Agricultural Statistics Research Institute, New Delhi, India*
[2]*Field Operations Division, National Sample Survey Office, Faridabad, India*

Received 02 May 2011; Revised 19 August 2011; Accepted 01 February 2012

SUMMARY

In this article we describe an application of small area estimation techniques to derive district level estimates of crop yield for paddy in the State of Uttar Pradesh using the data on crop cutting experiments supervised under Improvement of Crop Statistics (ICS) scheme and the secondary data from Population Census. The results show considerable improvement in the estimates generated by using small area estimation method.

Keywords : Crop cutting experiments, General crop estimation surveys, Improvement of Crop Statistics, District level estimates, Small area estimation, Census.

1. INTRODUCTION

Crop area and crop production forms the backbone of any agricultural statistics system. In India, crop area figures are, by and large, compiled on the basis of complete enumeration while the crop yield is estimated on the basis of sample survey approach. The yield rate estimates are developed on the basis of scientifically designed crop cutting experiments (CCEs) conducted under the scheme of General Crop Estimation Surveys (GCES). A crop cutting experiment consists of randomly identifying a field growing a specific crop, locating and marking, as per specified instructions, a plot of given size and shape in the selected field, harvesting, threshing and winnowing the produce within the plot and weighing the grains obtained. Since the grain on the harvested day contains moisture, it is stored and reweighted after driage to determine the marketable form of produce. The GCES covers 68 crops (52 food and 16 non-food) in 25 States and 4 Union Territories. More than 500,000 CCEs are conducted annually for this purpose. This much sample size is sufficient to provide precise estimates of crop yield (i.e., production per hectare of land) at the district level. Although the CCE technique is an objective method of assessment of crop yield, the procedure of conduct of CCE is tedious and time consuming. Due to this and some other factors, a tendency has been seen that the enumerators do not follow the prescribed procedure for the conduct of CCE in a number of cases. As a result of this, the data quality under the GCES is observed to be below the desirable limit. To improve the quality of data collected under the GCES, a scheme titled 'Improvement of Crop Statistics (ICS)' has been introduced by the Directorate of Economics and Statistics, Ministry of Agriculture, Government of India and implemented by the National Sample Survey Office (NSSO) and the State Agricultural Statistics Authority (SASA) jointly. Under this scheme, quality check on the field operation of GCES is carried out by supervising around 30,000 CCE by NSSO and State Government supervisory officers. The findings of the ICS results reveal that the crop cutting experiments are generally not carried out properly resulting in data which lacks desired quality.

[*]*Corresponding author* : U.C. Sud
E-mail address : ucsud@iasri.res.in

In view of limitation of infrastructure and constraints of resources, there is a need felt to reduce the sample size under GCES drastically so that volume of work of the enumerator is reduced and also better supervision of the operation of CCE becomes possible leading to improvement in data quality. However, reduction in sample size will have a direct bearing on the standard error of the estimator. The reduced sample size is more alarming when used for producing estimates at district level since estimators based on the sample data from any particular district can be unstable. This small sample size problem can be easily resolved provided auxiliary information is available to strengthen the limited sample data from the district. The underlying theory is referred to as the small area estimation (SAE). The SAE techniques aim at producing reliable estimates for such districts/areas with small (or even no) sample sizes by borrowing strength from data of other areas. The SAE techniques are generally based on model-based methods, see for example, Pfeffermann (2002) and Rao (2003). The idea is to use statistical models to link the variable of interest with auxiliary information, e.g. Census and Administrative data, for the small areas to define model-based estimators for these areas. Such small area models can be classified into two broad types:

(i) Area level random effect models, which are used when auxiliary information is available only at area level. They relate small area direct estimates to area-specific covariates (Fay and Herriot 1979) and

(ii) Unit level random effect models, proposed originally by Battese *et al.* (1988). These models relate the unit values of a study variable to unit-specific covariates.

In this article we explore an application of SAE techniques to derive model-based estimates of average yield for paddy crop at small area levels in the State of Uttar Pradesh in India by linking data generated under ICS scheme by NSSO (data collected with much reduced sample size, however, the quality of data is very high) and the Population Census 2001**.** Small areas are defined as the districts of State of Uttar Pradesh in India. It is noteworthy that we adopt the area level model since covariates for our study are available only at the area level. The paper illustrates how the ICS data and Census data can be combined to derive reliable district level estimates of crop yield. The rest of the paper is organised as follows. Section 2 introduces the data used for the analysis and Section 3 describes the methodology applied for the analysis. In Section 4 we present the diagnostic procedures for examining the model assumptions and validating the small area estimates and discuss the results. Section 5 finally sets out the main conclusions.

2. DATA DESCRIPTION

In this study we use data pertaining to supervised CCE on paddy crop under ICS scheme for kharif season for the State of Uttar Pradesh in India collected during the year 2009-10. The variable of interest for which small area estimates are required is yield for paddy crop. We are interested in estimating the average yield at the district level. In the State of Uttar Pradesh there are 70 districts, however supervision, on a sub-sample, of crop cutting experiments work under ICS scheme is carried out in 58 districts only and there is no sample data for the remaining 12 districts. In what follows, we refer these 12 districts as the out of sample districts. These 70 (58 in sample and 12 out of sample) districts are the small areas for which we are interested in producing the estimates. The area specific sample sizes for these 58 sample districts range from minimum of 4 to maximum of 28 CCE with average of 11 CCE (see Fig. 1). A total of 655 CCE were supervised for recording yield data in the State of Uttar Pradesh for paddy crop for the year 2009-10. We see that in a few districts the sample size is small so the traditional sample survey estimation approaches lead to unstable estimate. In addition, in 12 districts due to non

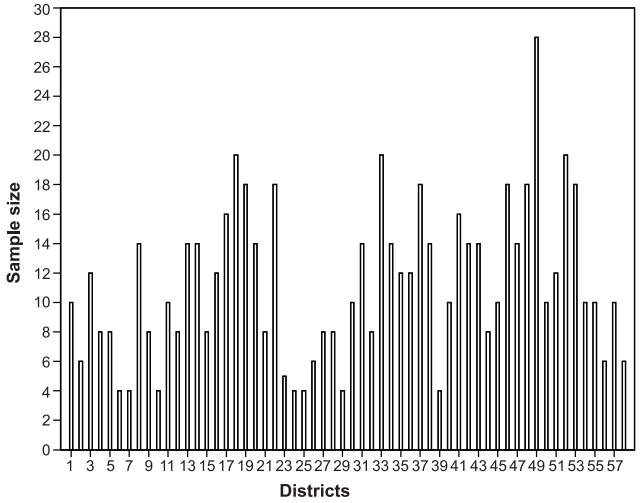

Fig. 1. Distribution of district-specific sample sizes in sample districts.

availability of sample under ICS, we can not estimate paddy yield**.** Indeed, there is no design based solution to provide estimates for these 12 out of sample districts (Pfeffermann 2002). The SAE is an obvious choice for such cases. The covariates (auxiliary variables) known for the population are drawn from the Population Census 2001. Note that use of covariates from the 2001 Population Census to model yield data of paddy crop from the 2009-10 ICS scheme data may raise issues of comparability. However, the covariates used in this study are not expected to change significantly over a short period of time. There were 121 covariates available from these sources to consider for modelling. However, we did some exploratory data analysis, for example, first we segregated group of covariates with significant correlation with target variable and subsequently we implemented step wise regression analysis. Finally we choose model with two significant variables, average household size (HH_SIZE) and female population of marginal household (MARG_HH_F) with 26 per cent R^2. The residual diagnostic plots in Fig. 2 indicate that fitted model is reasonable. For SAE analysis we therefore used these two covariates. Note that for SAE of 12 out of sampled districts we used the same two covariates since we assume that the underlying model for sample areas also holds for out of sample districts.

3. SMALL AREA ESTIMATION METHODOLOGY

In this Section we describe the underlining theory of SAE used in the paper. In particular, we elaborate SAE based on the area level model (Fay and Herriot 1979). It was proposed to estimate the per-capita income of small places with population size less than 1000. This model relates small area direct survey estimates to area-specific covariates. The SAE under this model is one of the most popular methods used by private and public agencies because of its flexibility in combining different sources of information and explaining different sources of errors. To start with, we first fix our notation. Throughout, we use a subscript d to index the quantities belonging to small area or district d (d = 1, ..., D) , where D is the number of small areas (or districts) in the population. Let $\hat{\theta}_d$ denotes the direct survey estimate of unobservable population value θ_d for area d (d = 1, ..., D). Let $\mathbf{x}_d$ be the p-vector of known auxiliary variable, often obtained from various

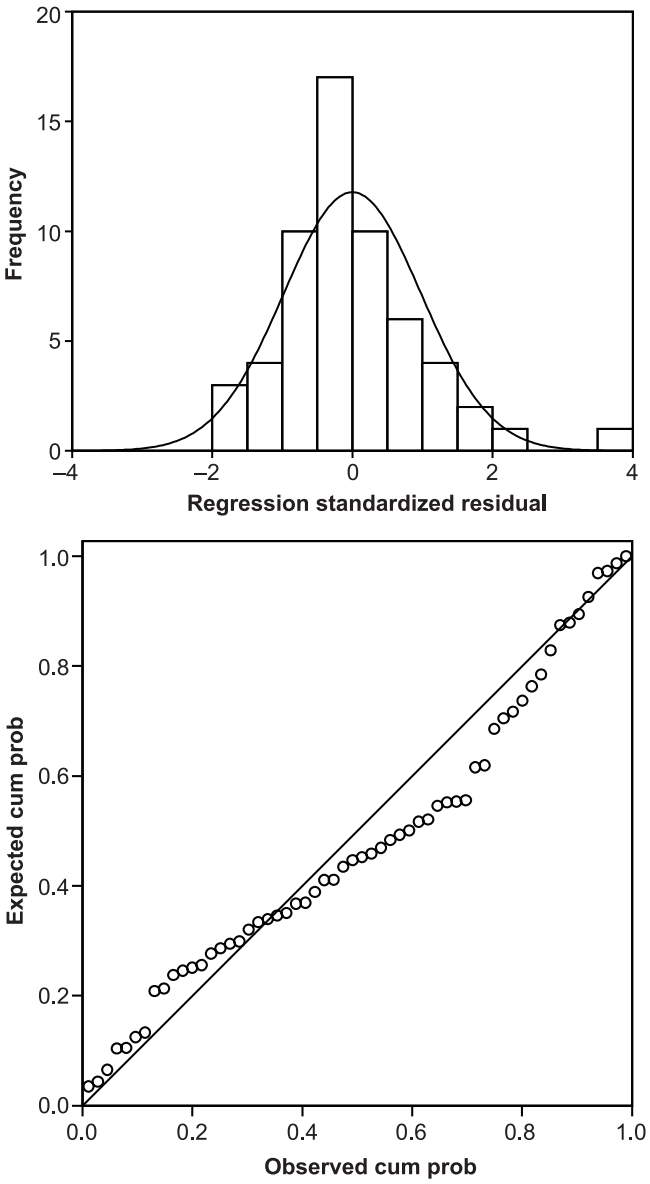

Fig. 2. Histogram and normal P-P plot of regression standardized residual.

administrative and census records, related to the population mean θ_d. The simple area specific two stage model suggested by Fay and Herriot (1979) has the form

$$\hat{\theta}_d = \theta_d + e_d \text{ and } \theta_d = \mathbf{x}_d^T\boldsymbol{\beta} + u_d,\ d = 1, ..., D. \quad (1)$$

We can express model (1) as an area level linear mixed model given by

$$\hat{\theta}_d = \mathbf{x}_d^T\boldsymbol{\beta} + u_d + e_d;\ d = 1, ..., D. \quad (2)$$

Here $\boldsymbol{\beta}$ is a p-vector of unknown fixed effect parameters, u_d's are independent and identically distributed normal random errors with E(u_d) = 0 and

$\text{Var}(u_d) = \sigma_u^2$, and e_d's are independent sampling errors normally distributed with $\text{E}(e_d \mid \theta_d) = 0$, $\text{Var}(e_d \mid \theta_d) = \sigma_d^2$. The two errors are independent of each other within and across areas. Usually, σ_d^2 is known while σ_u^2 is unknown and it has to be estimated from the data. Methods of estimating σ_u^2 include maximum likelihood (ML) and restricted maximum likelihood (REML) under normality, the method of fitting constants without normality assumption, See Rao (2003, Chapter 5). Let $\hat{\sigma}_u^2$ denotes estimate of σ_u^2. Then under model (2), the Empirical Best Linear Unbiased Predictor (EBLUP) of θ_d is given by

$$\hat{\theta}_d^{EBLUP} = \mathbf{x}_d^T\hat{\boldsymbol{\beta}} + \hat{\gamma}_q(\hat{\theta}_d - \mathbf{x}_d^T\hat{\boldsymbol{\beta}}) = \hat{\gamma}_d\hat{\theta}_d + (1-\hat{\gamma}_d)\mathbf{x}_d^T\hat{\boldsymbol{\beta}} \quad (3)$$

where $\hat{\gamma}_d = \hat{\sigma}_u^2/(\sigma_d^2 + \hat{\sigma}_u^2)$ and $\hat{\boldsymbol{\beta}}$ is the generalized least square estimate of $\boldsymbol{\beta}$. It may be noted that $\hat{\theta}_d^{EBLUP}$ is a linear combination of direct estimate $\hat{\theta}_d$ and the model based regression synthetic estimate $\mathbf{x}_d^T\hat{\boldsymbol{\beta}}$, with weight $\hat{\gamma}_d$. Here $\hat{\gamma}_d$ is called "shrinkage factor" since it 'shrinks' the direct estimator, $\hat{\theta}_d$ towards the synthetic estimator, $\mathbf{x}_d^T\hat{\boldsymbol{\beta}}$. For out of sample areas (i.e. areas with $n_d = 0$), the EBLUP predictor (3) leads to synthetic predictor of the form $\hat{\theta}_d^{SYN} = \mathbf{x}_d^T\hat{\boldsymbol{\beta}}$.

Prasad and Rao (1990) proposed an approximately model unbiased (i.e. with bias of order o(1/D)) estimate of mean squared error (MSE) of the EBLUP (3) given by

$$\hat{MSE}(\hat{\theta}_d^{EBLUP}) = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2)\,\hat{\text{Var}}(\hat{\sigma}_u^2), \quad (4)$$

where

$$g_{1d}(\hat{\sigma}_u^2) = \hat{\gamma}_d\sigma_d^2,$$
$$g_{2d}(\hat{\sigma}_u^2) = (1-\hat{\gamma}_d)^2\mathbf{x}_d^T\,\hat{Var}\,(\hat{\boldsymbol{\beta}})\mathbf{x}_d, \text{ and}$$
$$g_{3d}(\hat{\sigma}_u^2) = \left\{\sigma_d^4/(\sigma_d^2 + \hat{\sigma}_u^2)^3\right\}\hat{Var}\,(\hat{\sigma}_u^2)$$

with $\hat{Var}(\hat{\sigma}_u^2) \approx 2D^{-2}\sum_{d=1}^{D}\left(\sigma_d^2 + \hat{\sigma}_u^2\right)^2$ when estimating $\hat{\sigma}_u^2$ by method of fitting constants. See Rao (2003, Chapter 5) for details about various theoretical developments. Under model (2), the MSE estimate for the synthetic predictor $\hat{\theta}_d^{SYN}$ is given by $\hat{MSE}\left(\hat{\theta}_d^{SYN}\right) = \mathbf{x}_d^T\,\hat{Var}\,(\hat{\boldsymbol{\beta}})\mathbf{x}_d + \hat{\sigma}_u^2$.

4. EMPIRICAL RESULTS

This Section presents the results from data and theory described in previous Sections. Some diagnostics to examine the reliability of small area estimates are carried out and the bias diagnostics and coefficient of variation are used to validate the reliability of the model-based small area estimates. 95 per cent confidence (CI) intervals for both direct and model-based estimates are also computed.

The bias diagnostics is used to investigate if the model-based estimates are less extreme when compared to the direct survey estimates. In addition, if direct estimates are unbiased, their regression on the true values should be linear and correspond to the identity line. If model-based estimates are close to the true values the regression of the direct estimates on the model-based estimates should be similar (Ambler *et al.* 2001 and Chandra *et al.* 2011). We plot direct estimates on *Y*-axis and model-based estimates on *X*-axis and look for divergence of regression line from $Y = X$ and test for intercept = 0 and slope = 1. The bias scatter plots of the direct estimates against the model-based estimates are given in Fig. 3. From the bias diagnostic it is found that the intercept fails this diagnostic (i.e., intercept is different from zero). The plots show that the model-based estimates are less extreme when compared to the direct estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average. The coefficient of variation (CV) are computed to assess the improved precision of the model-based estimates compared to the direct estimates. The CVs show the sampling variability as a percentage

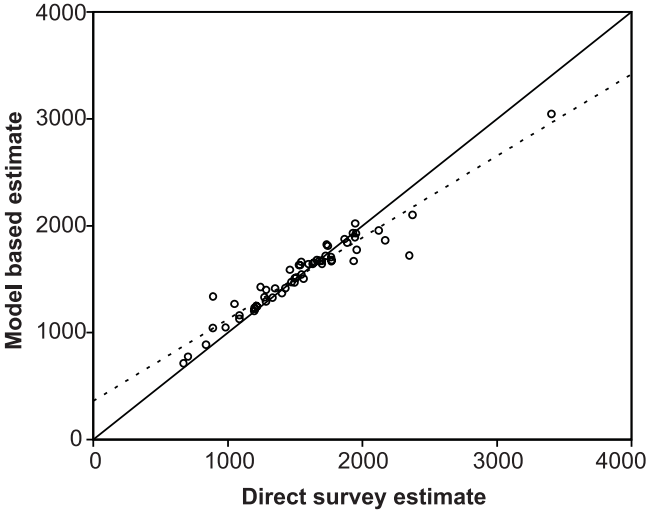

Fig. 3. Bias diagnostic plots for sample districts. Direct estimates versus model based estimates, y=x line (Solid) and linear regression fit line (dash).

of the estimate. Although, there are no internationally acceptable tables for judging what CV is too high, estimates with large CVs are considered unreliable. Fig. 4 shows the CVs for the direct survey estimates and model-based. The figure shows that the estimated CVs for the model-based estimates have a higher degree of reliability when compared to the direct survey estimates. Table 1 presents the district-wise model-

Table 1. Districts wise values of model-based estimate, 95 per cent confidence interval limits and coefficient of variation (CV) for paddy (green) yield (gm/43.3012 sq. mt.)

Districts	Estimate	Lower	Upper	CV, %	Districts	Estimate	Lower	Upper	CV, %
Saharanpur	17759	13667	21851	11.52	Ambedkar Nagar	16667	13652	19681	9.04
Muzaffarnagar	17208	11735	22681	15.90	Sultanpur	16793	13899	19688	8.62
Bijnor	18927	16306	21547	6.92	Bahraich	14735	13606	15865	3.83
Moradabad	16781	12329	21232	13.26	Shrawasti	15168	10783	19553	14.46
Rampur	17174	16148	18200	2.99	Balrampur	12338	9206	15470	12.69
Jyotiba Phule Nagar	11622	8894	14351	11.74	Gonda	16708	14611	18805	6.28
Ghaziabad	16726	11101	22351	16.82	Siddharthnagar	12921	9808	16033	12.05
Bulandshahar	18116	14555	21677	9.83	Basti	14165	10331	17999	13.53
Aligarh	14278	10277	18280	14.01	Sant Kabir Nagar	13273	11626	14920	6.20
Mathura	12688	8322	17054	17.20	Mahrajganj	18640	14465	22815	11.20
Etah	12508	10274	14742	8.93	Gorakhpur	12437	9608	15266	11.37
Mainpuri	13711	9065	18357	16.94	Kushinagar	16699	12301	21096	13.17
Budaun	13307	9961	16652	12.57	Deoria	8866	6143	11588	15.35
Bareilly	14140	10976	17305	11.19	Azamgarh	12033	10073	13993	8.14
Pilibhit	14687	10207	19166	15.25	Mau	10489	7090	13888	16.20
Shahjahanpur	18411	16184	20638	6.05	Ballia	7763	5056	10470	17.44
Kheri	15079	12023	18135	10.13	Jaunpur	16418	13286	19549	9.54
Sitapur	16422	12836	20007	10.92	Ghazipur	11279	8606	13953	11.85
Hardoi	19315	16665	21965	6.86	Chandauli	12229	8333	16125	15.93
Unnao	14005	11188	16821	10.05	Varanasi	17063	12659	21468	12.91
Lucknow	18242	13196	23289	13.83	Sant Ravidas Nagar	7133	2939	11327	29.40
Rae Bareli	19287	16128	22446	8.19	Mirzapur	15052	11815	18290	10.76
Farrukhabad	10446	7420	13471	14.48	Sonbhadra	16328	11079	21578	16.08
Kannauj	30450	27119	33782	5.47	Meerut[#]	14984	8898	21069	20.31
Etawah	15431	13899	16964	4.97	Baghpat[#]	12442	6182	18702	25.16
Auraiya	21021	17121	24922	9.28	Gautam Buddha Nr[#]	16704	10436	22973	18.76
Kanpur Dehat	19547	15717	23378	9.80	Hathras[#]	15258	9158	21357	19.99
Kanpur Nr	16315	12090	20539	12.95	Agra[#]	14803	8716	20890	20.56
Banda	13375	8039	18711	19.95	Firozabad[#]	14391	8289	20492	21.20
Fatehpur	15881	11406	20355	14.09	Jalaun[#]	15186	9048	21325	20.21
Pratapgarh	16437	12543	20331	11.84	Jhansi[#]	17378	11209	23547	17.75
Kaushambi	16624	11363	21884	15.82	Lalitpur[#]	16928	10684	23172	18.44
Allahabad	20218	16164	24272	10.03	Hamirpur[#]	16520	10273	22767	18.91
Barabanki	18756	15176	22336	9.54	Mahoba[#]	16285	10030	22540	19.21
Faizabad	16556	12690	20422	11.68	Chitrakoot[#]	14948	8773	21122	20.65

[#]Districts with no sample information under ICS, Nr denotes Nagar

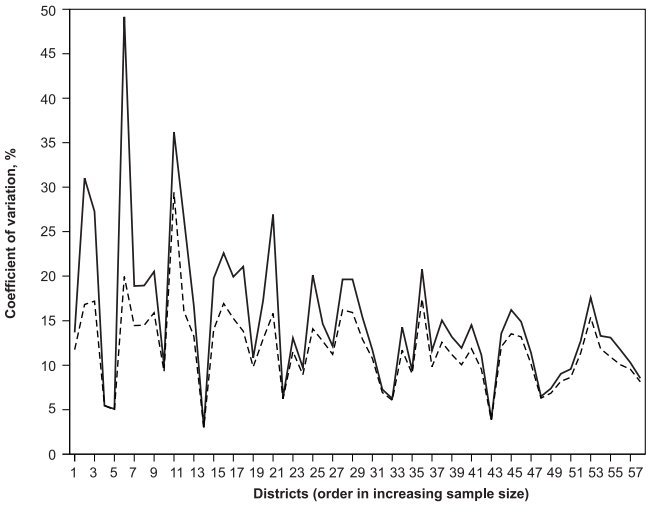

Fig. 4. Coefficient of variations of direct estimates (solid line) and model based estimates (dash line) for sampled districts.

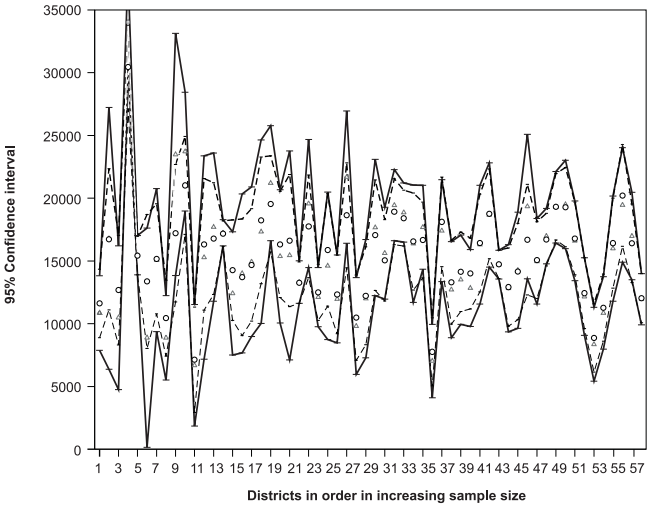

Fig. 5. 95 per cent confidence interval (CI) for direct estimates (Δ) and model based estimates (○) for sampled districts. CI for direct estimates (solid line) and CI for model based estimates (dash line).

based estimates, 95 per cent confidence interval (CI) limits and percentage coefficient of variation for paddy crop yield for all 70 (i.e. both for 58 sample and 12 out of sample) districts. In right hand side part of Table 1, results for last 12 districts correspond to out of sample districts. The CV results in Table 1 reveal that average CV of these out of sample districts is 20.10 per cent. Fig. 5 shows the 95% CI of the model-based and the direct survey estimates. It is apparent that the standard errors of the direct estimates are large and therefore the estimates are unreliable.

5. CONCLUSIONS

This paper illustrates that the small area estimation technique can be satisfactorily applied to produce reliable district level estimates of crop yield using CCE supervised under ICS scheme. Although the ICS supervised crop cutting experiments number only 30,000 in the entire country i.e. the sample size is very low, the collected data is of very high quality. The estimates generated using this data are expected to be relatively free from various sources of non-sampling errors. Further small area estimation technique provides estimates for those districts where there is no sample information under ICS and so direct estimates can not be computed. It is, therefore, recommended that wherever it is not possible to conduct adequate number of crop cutting experiments due to constraints of cost or infrastructure or both, small area estimation technique can be gainfully used to generate reliable estimates of crop yield based on a smaller sample.

REFERENCES

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M. and Wang, S. (2001). Combining unemployment benefits data and LFS data to estimate ILO unemployment for small areas: An application of a modified Fay-Herriot method. *Proceedings of the International Association of Survey Statistician,* Meeting of the ISI, Seoul, August 2001.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.

Census of India (2001). Registrar General and Census Commissioner, New Delhi, India.

Chandra, H., Salvati, N. and Sud, U.C. (2011). Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India-An application of small area estimation technique. *J. Appl. Statist.*, Forthcoming issue.

Fay, R.E. and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74,** 269-277.

Pfeffermann, D. (2002). Small area estimation: New Developments and directions. *Intern. Statist. Rev.*, **70**, 125-143.

Rao, J.N.K. (2003). *Small Area Estimation.* Wiley, New York.