Journal of the Korean Statistical Society (2005), 34: 1, pp 73–83

ON COMPARISON OF PERFORMANCES OF SYNTHETIC AND NON-SYNTHETIC GENERALIZED REGRESSION ESTIMATIONS FOR ESTIMATING LOCALIZED ELEMENTS

AMITAVA SAHA[1]

ABSTRACT

Thompson's (1990) adaptive cluster sampling is a promising sampling technique to ensure effective representation of rare or localized population units in the sample. We consider the problem of simultaneous estimation of the numbers of earners through a number of rural unorganized industries of which some are concentrated in specific geographic locations and demonstrate how the performance of a conventional Rao-Hartley-Cochran (RHC, 1962) estimator can be improved upon by using auxiliary information in the form of generalized regression (greg) estimators and then how further improvements are also possible to achieve by adopting adaptive cluster sampling.

AMS 2000 subject classifications. Primary 62D05.

Keywords. Adaptive cluster sampling; Generalized regression estimator; Mean square error; Synthetic estimator; Rao-Hartley-Cochran sampling scheme.

1. INTRODUCTION

Thompson (1990) introduced the concept of adaptive cluster sampling to ensure effective representation of rare or localized population elements in a sample. A population is described as rare or localized in the sense that the number of elements having non-zero values of the variable of interest, i.e. the study variable is very small or such units are concentrated in specified geographical region or pockets of the population. The concept of adaptive sampling was further studied and developed by Thompson and Seber (1996), Salehi and Seber (1997, 2002),

Received August 2004; accepted December 2004.

[1]Directorate General of Mines Safety, Dhanbad, Jharkhand - 826001, India (email : saha_amitava@hotmail.com)

Chaudhuri (2000), Chaudhuri, Bose and Ghosh (2004), Christman (2003), Brown (2003) and Brown and Manly (1998) among others.

Large-scale annual socio-economic surveys are conducted in India to provide estimate for the Gross Domestic Product (GDP). But due to heavy localization of the rural unorganized industries having significant contribution towards the nation's GDP, it has not been possible so far, to provide an accurate GDP estimate as the traditional sampling schemes adopted in these surveys fails to ensure effective capture of earners through the rural small-scale unorganized industries.

India has a traditional system of conducting decennial population censuses and of late, it has become a practice to carry out economic censuses (EC) to keep an account of the number of earners engaged in non-agricultural enterprises in the rural and urban sectors of the economy. The last EC was conducted in the year 1998 considering the rural and urban administrative units as the first stage units (fsu's) and the villages within the fsu's as the second stage units (ssu's). Keeping in mind the fact that EC data may be of use to identify the rural localized industries, it might be possible to improve substantially the precision of the GDP estimates through augmentation of a traditional sample by adaptive samples.

In this paper we demonstrate how the precision of a conventional estimator of the population total proposed by Rao, Hartley and Cochran (RHC, 1962) improves through the use of an auxiliary variable in the form of Cassel, S扼愼ndal, and Wretman's (CSW, 1976) generalized regression (greg) estimators and then through its adaptive versions by effective utilization of EC and population census data.

We present Table 1.1 depicting the distribution of six rural small-scale industries in 30 rural administrative blocks and a total of 2056 villages based on EC - 98 data of a particular district of West Bengal, India.

TABLE 1.1 *Distribution, village- and block-wise, of earners by industries in the district*

industry type	code	number of earners	number of blocks	number of villages	minimum (> 0) in a block	maximum in a block
Running poultry	0250	3163	30	555	20	390
Coastal fishing	0600	9132	12	85	0	2886
Cotton khadi	2320	2010	12	53	0	1569
Brass-work	3341	1721	4	31	0	1642
Eateries	6900	14082	30	1318	53	1173
Priesthood	9400	7580	30	1098	14	717

A close look at Table 1.1 above reveals that the industries 'Coastal fishing', 'Cotton khadi' and 'Brass work', coded respectively by '0600', '2320' and '3341' are concentrated in some specific geographical locations and unless some special efforts are made, it would not be easy to capture the earners in abundance through these three industries.

Our proposed methodology of estimation and sampling are presented in Section 2. The numerical findings using live data are given in Section 3 and comments and recommendations are made in Section 4.

2. SAMPLE SELECTION AND ESTIMATION

The 30 administrative blocks of the district are considered as the N fsu's and the villages within the blocks are treated as the ssu's, M_i denoting the number of ssu's in the ith block. The distribution of the villages within the blocks is given in Table 2.1.

The total population of the blocks according to 1991 population census is taken as the size-measures for the blocks and for the villages the total numbers of enterprises in the villages according to EC 1998 are treated as the size-measures. Let us denote by p_i the normed size-measure for the ith fsu and by p_{ij} that of the jth ssu in the ith fsu ($j = 1, \cdots, M_{ij}; i = 1, \cdots\cdots, n$). For selecting a sample of $n = 10$ blocks and then from the ith selected block a sample of $m_i = [M_i/5], i = 1, \cdots, n$ villages we propose to employ the sampling scheme due to Rao, Hartley and Cochran (RHC, 1962) in both the stages. To select a sample of n blocks from N blocks by RHC strategy, first certain positive integers N_i's are chosen subject to $\sum_n N_i = N$, $\sum_n$ denoting the sum over the n random

TABLE 2.1 *Showing distribution of the villages within the blocks*

block no.	no. of villages (M_i)	block no.	no. of villages (M_i)	block no.	no. of villages (M_i)
01	130	11	43	21	39
02	83	12	59	22	36
03	59	13	60	23	70
04	86	14	65	24	90
05	68	15	8	25	93
06	74	16	17	26	27
07	16	17	41	27	87
08	60	18	131	28	108
09	70	19	84	29	167
10	49	20	86	30	50

groups into which the N blocks are randomly split-up, the ith group containing N_i units. Then, writing $Q_i = p_{i_1} + \cdots + p_{i_{N_i}}$ for the ith group, one unit is chosen from the ith group with a probability proportional to its p-value divided by Q_i and this is repeated independently for each of the n groups to get a sample of n blocks. Suppose that the ith block with M_i villages is eventually selected. For obtaining a sample of m_i villages from this block following RHC scheme as applied in sampling the blocks we write $\sum_{m_i}$, M_{i_j}, Q_{i_j} paralleling the notations $\sum_n$, N_i, Q_i where m_i denotes the numbers of random disjoint groups into which the M_i villages of the ith block are randomly divided. Denoting y_{ij} as the value of a variable of interest corresponding to the jth selected village of the ith chosen block, an unbiased estimator for the ith block total by RHC's strategy is given by

$$\hat{y}_{ij} = \Sigma_{m_i} y_{ij} \frac{Q_{ij}}{p_{ij}}$$

along with an unbiased variance-estimator

$$v(\hat{y}_i) = C_i \sum_{m_i} \sum_{m_i} Q_{ij} Q_{ik} \left(\frac{y_{ij}}{p_{ij}} - \frac{y_{ik}}{p_{ik}}\right)^2$$

writing $C_i = (\Sigma_{m_i} M_{ij}^2 - M_i)/(M_i^2 - \Sigma_{m_i} M_{ij}^2)$, $\sum_{m_i} \sum_{m_i}$ to denote sum over non-duplicated and non-overlapping distinct pairs of the groups, $j = 1, \cdots\cdots, m_i$ $(j \neq i)$. Thus an over-all estimator for $p_{i=1}^N y_i$, i.e., the total of y for the entire district is

$$t = \sum_n \hat{y}_i \frac{Q_i}{p_i} = \sum_n \frac{Q_i}{p_i} \sum_{m_i} y_{ij} \frac{Q_{ij}}{p_{ij}}.$$

Following Chaudhuri, Adhikary and Dihidar (2000) an unbiased variance estimator for t is given by

$$v(t) = C\sum_n \sum_n Q_i Q_j (\frac{\hat{y}_i}{p_i} - \frac{\hat{y}_j}{p_j})^2 + \sum_n \frac{Q_i}{p_i} v(\hat{y}_i)$$

writing $C = (\Sigma_n N_i^2 - N)/(N^2 - \Sigma_n N_i^2)$.

A possible improvement on t can be attained by introducing a regressor variable in the form of Cassel, Särndal and Wretman's (CSW, 1976) generalized regression estimator, which is both 'asymptotically design unbiased' (ADU) and 'asymptotically design consistent' (ADC). We consider the total number of all workers in a village as per EC-98 as the regressor, denoting by x_{ij} the value of the regressor, say, x for the jth village in the ith block; x_i as the total for the ith block and X as the total of x for the entire district. Then a revised estimator of y_i can be obtained by using synthetic and non-synthetic greg estimators in the following two ways,

$$\hat{y}_{g1i} = \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij} + \hat{b}(x_i - \sum_{m_i} \frac{Q_{ij}}{p_{ij}} x_{ij}),$$
$$\hat{y}_{g2i} = \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij} + \hat{b}_i(x_i - \sum_{m_i} \frac{Q_{ij}}{p_{ij}} x_{ij}),$$

writing $\hat{b} = (\sum_n Q_i/p_i \sum_{m_i} y_{ij}x_{ij}R_{ij})/(\sum_n Q_i/p_i \sum_{m_i} x_{ij}^2 R_{ij})$, $\hat{b}_i = (\sum_{m_i} y_{ij}x_{ij}R_{ij})/(\sum_{m_i} x_{ij}^2 R_{ij})$ and R_{ij} as a suitably chosen positive constant. The values of R_{ij} are $1/x_{ij}$, $1/x_{ij}^2$, $1/((p_{ij}x_{ij})/Q_{ij})$, $(1 - p_{ij}/q_{ij})/((p_{ij}x_{ij})/Q_{ij})$ similar to $1/x_i$, $1/x_i^2$, $1/(\pi_i x_i)$, $(1 - \pi_i)/(\pi_i x_i)$ in CSW's (1976) original greg estimator. Here we consider $R_{ij} = (1 - p_{ij}/q_{ij})/((p_{ij}x_{ij})/Q_{ij})$. Following Särndal (1982) an estimator for mean square error of $\hat{y}_{gki}$, $k = 1, 2$ is given by

$$m_k(\hat{y}_{gki}) = C_i \sum_{m_i} \sum_{m_i} Q_{ij}Q_{il}(\frac{e_{kij}}{p_{ij}} - \frac{e_{kil}}{p_{il}})^2$$

where $e_{1ij} = y_{ij} - \hat{b}x_{ij}$ and $e_{2ij} = y_{ij} - \hat{b}_i x_{ij}$. The over-all greg estimators for the population total $p_{i=1}^N p_{j=1}^{M_i} Y_{ij}$ are given by

$$t_{g1} = \sum_n \hat{y}_{g1i} \frac{Q_i}{p_i} \text{ and } t_{g2} = \sum_n \hat{y}_{g2i} \frac{Q_i}{p_i}$$

with the corresponding MSE estimators as

$$m_k(t_{gk}) = C\sum_n \sum_n Q_i Q_i (\frac{\hat{y}_{gki}}{p_i} - \frac{\hat{y}_{gkj}}{p_j})^2 + \sum_n \frac{Q_i}{p_i} m_k(\hat{y}_{gki}), k = 1, 2.$$

Among the six industries considered here, three are highly concentrated in unknown geographical locations and it is a common feature in rural India to have a

highly skewed distribution of earners through rural unorganized industries. So, for the sake of improving the performance of greg estimators discussed above we apply adaptive cluster sampling following Thompson (1990) and Thompson and Seber (1996).

In order to implement adaptive cluster sampling, for each sampled unit a 'neighbourhood' consisting of the initially selected unit and one or more units is defined. In our case, all the villages with a common boundary with the initially selected village are assumed to constitute a neighbourhood of the initially chosen village. For each observed unit in the initial sample the neighbouring units are sampled if they satisfy certain pre-defined condition C and this process is repeated till such a unit not satisfying C is found. The collection of all such units is called a 'cluster' of the initial unit. The units not satisfying C in a cluster are termed as 'edge-units' and all the units satisfying C in a cluster constitutes a 'network' for the initial unit. Now if the edge-units are treated as 'singleton' networks then it follows that all the networks for a population are disjoint and they together exhaust the population. Then writing $A(ij)$ as the network for the jth village in the ith block and by denoting C_{ij}, its cardinality it follows that

$$p_{i=1}^{N} p_{j=1}^{M_i} u_{ij} = p_{i=1}^{N} p_{j=1}^{M_i} y_{ij} = Y \quad \text{and} \quad p_{i=1}^{N} p_{j=1}^{M_i} v_{ij} = p_{i=1}^{N} p_{j=1}^{M_i} x_{ij} = X$$

where $u_{ij} = (C_{ij})^{-1} \sum\limits_{(ij)\epsilon A(ij)} y_{ij}$ and $v_{ij} = (C_{ij})^{-1} \sum\limits_{(ij)\epsilon A(ij)} x_{ij}$.

Following Thompson and Seber (1996), Chaudhuri (2000) demonstrated how adaptive cluster sampling may be effectively used for estimating a finite population total of a rare or localized population when the initial sample is chosen by adopting any arbitrary unequal probability sampling design. Since $p_{i=1}^{N} p_{j=1}^{M_i} u_{ij} = p_{i=1}^{N} p_{j=1}^{M_i} y_{ij} = Y$ and $p_{i=1}^{N} p_{j=1}^{M_i} v_{ij} = p_{i=1}^{N} p_{j=1}^{M_i} x_{ij} = X$, estimating Y and X is equivalent to estimating $p_{i=1}^{N} p_{j=1}^{M_i} u_{ij}$ and $p_{i=1}^{N} p_{j=1}^{M_i} v_{ij}$ respectively.

Now based on the adaptive sample reached through the original two-stage RHC sampling scheme the revised formulae corresponding to $\hat{y}_{gki}$ and $m_k(\hat{y}_{gki})$, $k = 1, 2$ are derived simply on replacing y_{ij} by u_{ij} and x_{ij} by v_{ij} in these formulae.

The set, say, $A(s)$ of units in the networks of all the initially chosen units in the sample s is called an adaptive sample and the process of extending s to $A(s)$ is called adaptive cluster sampling. We present below a table showing the association among the various rural industries based on EC-98 data.

TABLE 2.2 *Giving distribution of the earners by various industries village-wise in the district*

industry code	0250	0600	2320	3341	6900	9400
0250	555	43	23	2	432	350
	(100.00)	(50.59)	(43.40)	(6.45)	(32.78)	(31.88)
0600	43	85	0	0	72	73
	(7.75)	(100.00)	(0.00)	(0.00)	(5.46)	(6.65)
2320	23	0	53	0	37	40
	(4.14)	(0.00)	(100.00)	(0.00)	(2.81)	(3.64)
3341	2	0	0	31	21	15
	(0.36)	(0.00)	(0.00)	(100.00)	(1.59)	(1.37)
6900	432	72	37	21	1318	808
	(77.84)	(84.71)	(69.81)	(64.74)	(100.00)	(73.59)
9400	350	73	40	15	808	1098
	(63.06)	(85.88)	(75.47)	(48.39)	(61.31)	(100.00)

In the above table, the diagonal entries, say, a_{ii} present the numbers of villages where there are earners by the industry i and the off-diagonal entries, say, a_{ij} present the numbers of villages with earners by both the industries i and j; i,j being the six industries 0250, 0600, 2320, 3341, 6900 and 9400. Then the formula $\eta_{ij} = 100a_{ij}/a_{jj}$ derives the entries in the parentheses.
It may be observed from Table 2.2 that out of the 85 villages having workers engaged in 'coastal fishing', 85% and 86% villages also have workers engaged in both 'coastal fishing and restaurant' and 'coastal fishing and priesthood'. Thus noting the association among the different industries from Table 2.2 it is pretty simple to construct networks for selection of adaptive samples.

3. NUMERICAL ILLUSTRATIONS

The relative performances of the alternative estimators are judged in the light of the following two criteria, viz.,

(i) ACV, the average coefficient of variation,

(ii) ACP, the actual coverage percentage.

Let e be a point estimator for any parameter θ and ν be its MSE estimator. Then by treating $\tau = (e-\theta)/\sqrt{\nu}$ as a $N(0,1)$ variate the 95% confidence interval (CI) for θ is given by $(e-1.96\sqrt{\nu}, e+1.96\sqrt{\nu})$ and the coefficient of variation (CV) for e is $100\sqrt{\nu}/e$. Then the ACV is the average of the CV over $R = 1000$ replicated

TABLE 3.1 *Showing performances of original RHC estimator and Synthetic and Non-synthetic generalized regression estimators*

industry code	RHC estimator		synthetic greg estimator		non-synthetic greg estimator	
	ACV	ACP	ACV	ACP	ACV	ACP
0250	29.1	86.3	26.8	87.6	26.6	87.1
0600	76.0	79.1	72.7	81.4	71.9	83.8
2320	78.2	46.9	74.4	52.4	75.2	52.6
3341	83.7	36.8	81.4	44.8	80.4	44.7
6900	12.2	94.6	14.3	93.9	14.7	94.4
9400	23.6	92.8	24.6	88.6	24.4	87.9
sample size		267				

samples and the ACP is the percentage of cases for which a CI covers θ. The smaller the value of ACV, the better e is as a point estimator and the closer the value of ACP is to 95, the better it is.

The relative performances of the original RHC estimator and synthetic and non-synthetic greg estimators based on $m_i = [M_i/5]$, $i = 1, \cdots, \cdots, n$ villages chosen from $n = 10$ blocks are shown in Table 3.1.

As asserted earlier, the traditional RHC estimator fails to capture effectively the numbers of earners through the industries "Coastal fishing", "Cotton khadi" and "Brass-work". The greg versions of the RHC estimator improve the efficiency slightly, but their performances too are not satisfactory for these three industries. So, in the next stage we apply adaptive cluster sampling through network formation and use the generalized regression method of estimation. We use three criteria for 'network' formation, viz.,

(i) whether or not the number of earners in a village through the industries 0250 or 6900 is 'zero' or 'positive',

(ii) whether or not the number of earners in a village through the industries 6900 or 9400 is 'zero' or 'positive',

(iii) whether or not the number of earners in a village through the industries 0250 or 9400 is 'zero' or 'positive'.

The relative performances of original RHC estimator and synthetic and non-synthetic greg estimators for the above mentioned three networks are shown in Table 3.2.

TABLE 3.2 *Showing relative performances of original RHC estimator and Synthetic and Non-synthetic generalized regression estimators under adaptive cluster sampling*

industry code		original RHC estimator		synthetic greg estimator		non-synthetic greg estimator
	ACV	ACP	ACV	ACP	ACV	ACP
Network: 0250&6900						
0250	26.1	88.4	21.3	93.6	22.6	96.3
0600	71.4	80.3	61.9	80.9	63.4	82.4
2320	68.6	51.1	57.7	55.8	62.5	54.0
3341	80.6	40.6	76.5	48.6	79.4	49.1
6900	11.5	93.1	11.9	92.0	12.0	93.4
9400	20.6	93.5	20.5	92.4	21.3	93.6
sample size		624				
Network: 6900&9400						
0250	26.2	90.1	21.9	92.4	22.5	93.9
0600	68.5	80.0	62.5	80.0	63.0	81.8
2320	68.9	50.4	58.0	50.3	61.7	51.0
3341	79.2	40.8	75.3	41.8	76.4	41.1
6900	12.1	91.0	13.2	94.3	13.4	93.7
9400	19.9	92.1	21.1	90.8	21.9	90.8
sample size		751				
Network: 0250&9400						
0250	24.3	88.0	19.9	88.5	22.8	94.1
0600	67.2	78.8	61.3	80.7	64.5	80.7
2320	68.4	48.4	50.5	49.0	59.0	51.6
3341	79.0	38.1	74.2	42.2	80.2	43.9
6900	12.2	94.5	12.0	89.2	11.4	88.4
9400	19.8	93.2	18.6	90.5	20.0	95.1
sample size		545				

4. COMMENTS AND RECOMMENDATIONS

The conventional RHC scheme as also the sophisticated tool like the generalized regression method of estimation fail to produce serviceable estimates of the number of earners for the heavily localized rural-unorganized industries like "Coastal fishing", "Cotton khadi" and "Brass-work" as is evident from Table 3.1. For these three industries adaptive cluster sampling substantially improves the precision of the greg estimators for all the three methods of network formation. The synthetic greg estimator outperforms the non-synthetic greg estimator in terms of ACV irrespective of the method of network formation and in respect of

all the industries discussed here. The ACP of the non-synthetic greg estimator turns out to be marginally better than that of the synthetic version for some of the industries. However, considering the two criteria for comparison the synthetic greg estimator performs uniformly better than the non-synthetic version in case of adaptive cluster sampling. For the remaining industry groups also the synthetic version induces substantial improvement in efficiency. So our recommendation in similar practical situation is to use adaptive cluster sampling along with the synthetic greg estimator following the methodology discussed here. In spite of having excellent capacity of ensuring effective capture of rare or localized population elements, adaptive cluster sampling suffers from the inherent drawback of increasing the initial sample size excessively. Methods recommending checks on the sample size are discussed in Salehi and Seber (1997, 2002), Brown (2003), Brown and Manly (1998), Christman (2003), Lo, Griffith and Hunter (1997), Lee (1998) and Chao (2003).

ACKNOWLEDGEMENT

The help and guidance received from Prof. Arijit Chaudhuri of Indian Statistical Institute is gratefully acknowledged.

REFERENCES

BROWN, J.A.(2003). "Designing an efficient adaptive cluster sample", *Environmental and Ecological Statistics*, **10**, 95–105.

BROWN, J. A. AND MANLY, B. F. J. (1998). "Restricted adaptive cluster sampling", *Environmental and Ecological Statistics*, **5**, 49–63.

CASSEL, C.M., SÄRNDAL, C.E. AND WRETMAN, J.H. (1976). "Some results on generalized difference estimation and generalized regression estimation for finite population", *Biometrika*, **63**, 615–620.

CHAO, CHANG-TAI (2003). "Markov Chain Monte Carlo on optimal adaptive sampling selections", *Environmental and Ecological Statistics*, **10**, 129–151.

CHAUDHURI, A. (2000). "Network and adaptive sampling with unequal probabilities", *Calcutta Statistical Association Bulletin*, **50**, 237–253.

CHAUDHURI, A., ADHIKARY, A.K. AND DIHIDAR, S. (2000). "Mean square error estimation in multi-stage sampling", *Metrika*, **52** (2), 115–131.

CHAUDHURI, A., BOSE, M. AND GHOSH, J.K. (2004). "An application of adaptive sampling to estimate highly localized population segments", *Journal of Statistical Planning & Inference*, **121**, 175 – 189.

CHRISTMAN, M. C. (2003). "Adaptive two-stage one-per stratum sampling", *Environmental and Ecological Statistics*, **10**, 43–60.

LEE, K. (1998). "Two-phase adaptive cluster sampling with unequal probabilities selection", *Journal of the Korean Statistical Society*, **27**, 265–278.

LO, N. C. H., GRIFFITH, D. AND HUNTER, J. R. (1997). "Using a restricted adaptive cluster sampling to estimate Pacific Hake larval abundance", *CalCOFI Report*, **38**, 103–113.

RAO, J.N.K., HARTLEY, H.O. AND COCHRAN, W.G. (1962). "On a simple procedure of unequal probability sampling without replacement", *Journal of the Royal Statistical Society*, **B24**, 482–491.

SÄRNDAL, C.E. (1982). "Implications of survey design for generalized regression estimation of linear functions", *Journal of Statistical Planning & Inference*, **7**, 155–170.

SALEHI, M.M. AND SEBER, G.A.F. (2002). "Unbiased estimators for restricted adaptive cluster sampling", *Australian & New Zealand Journal of Statistics*, **44**(1), 63–74.

SALEHI, M.M. AND SEBER, G.A.F. (1997). "Adaptive cluster sampling with networks selected without replacement", *Biometrika*, 84(1), 209–219.

THOMPSON, S.K. (1990). "Adaptive cluster sampling", *Journal of the American Statistical Association*, **85**, 1050–1059.

THOMPSON, S.K. (1992). *Sampling*, Wiley & Sons, New York.

THOMPSON, S.K. AND SEBER, G.A.F. (1996). *Adaptive Sampling*, Wiley & Sons, New York.