An application of adaptive sampling to estimate highly localized population segments

Arijit Chaudhuri*,[1], Mausumi Bose, J.K. Ghosh

Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India

Received 2 January 2002; accepted 3 December 2002

Abstract

It is difficult to enumerate the people in India who are engaged in various small-scale industries in the unorganized sector because they are concentrated in small regional pockets. In estimating separately the total numbers of workers earning principally through ten respective single-industries in the unorganized small-scale sector in a specific district in rural India, through numerical illustrations we have two observations to report: (1) A traditional stratified two-stage sampling scheme is ineffective for some of the industries because of failures to capture the earners concentrated in priorly unknown locations. (2) An adaptive sampling scheme extending the initial sample by appropriate 'network' formations based on well-defined 'neighbourhoods' brings about dramatic improvements exploiting clustering tendencies of earners by different industries.

MSC: 62 D05

Keywords: Adaptive sampling; Rao–Hartley–Cochran scheme; Stratified two-stage sampling; Varying probabilities

1. Introduction

The problem of estimating the total numbers of earners through specific small-scale industries in the Indian villages in the unorganized sector in various districts is considered to be a hard task. There are many reasons for this. These numbers do not remain stable across regions and over time industry-wise. The people in this sector change their occupations too frequently. The birth and death rates of such small units are too

* Corresponding author. Tel.: +91-33-25778085; fax: +91-33-25773104.

E-mail address: achau@isical.ac.in (A. Chaudhuri).

[1] His research is partially supported by the Grant No: 21(0539)/02/EMR-II of CSIR, India.

high. They are often concentrated in small regional pockets and only in a few establishments. Moreover, they frequently change occupations and locations. However, even if the earners' individual earnings be meagre, their collective contribution to the nation's gross domestic product (GDP) is believed to be substantial and hence is of national interest. Standard sampling designs using readily available frames are felt inadequate to capture them to throw up good estimates. To facilitate preparation of suitable frames for effective application of sampling schemes, a few nation-wide Economic Censuses have been held in India. In spite of that, a follow-up survey may still fail to capture the sparsely scattered industry-wise rural earners in the unorganized sector.

It is shown in this paper that good results can be obtained for this problem by using an appropriate adaptive sampling design based on suitably constructed 'networks', which is initiated by a traditional stratified two-stage scheme with varying probabilities in the first stage. The principle is illustrated by a numerical study using data from the nation-wide Economic Census held in India in 1990–1991 and the importance of the choice of 'networks' is highlighted. The key idea is that for estimating the 'domain total' for a highly-localized industry, i.e., the total numbers of people engaged in it in a specific place, instead of constructing networks based on this same industry, it is more appropriate to define networks on the basis of one or more less-localized industries which 'coexist' and thus are 'well associated' with it. The same technique may be applied for other similar surveys where the objective is to estimate highly localized population segments.

A striking revelation through our illustration here is that by appropriate formation of networks, hopelessly inadequate initial estimates of several localized domain totals can be dramatically improved while retaining the satisfactory levels of quality for all the other domains of interest. Such improvements are not possible with some other network formations. Though we use the Economic Census, 1990–1991 in our illustration here, its use could be avoided in specifying the initial design and estimators—its real use is in the network formation.

An adaptive sampling procedure employed by us leaves the final sample-size as a random variable. An interesting possible modification suggested to us by a referee that requires further exploration is to suitably define neighbourhoods to keep additional sample size within specified limits. We exclude it here as this would demand a significant shift from our aim in this investigation. Our focus is on the application in the present context of a traditional Adaptive Sampling technique which is popularly employed in wildlife surveys and exploration of mineral deposits.

In recommending fruitful network formations we have relied here upon the empirical observations presented by the Economic Census. However, there is another alternative attractive possibility pointed to us by a referee. As suggested by him, Besag's (1974) CAR model may be quite relevant.

Another alternative course that might perform well in the present investigation could be the use of a 'model-cum-design' based generalized regression (greg) estimator, and the use of a purely model-based empirical Bayes (EB) estimator as a convex combination of the 'greg' estimator and an estimator involving that of the regression coefficient, postulating a regression model with a zero-intercept and a single regressor variable relevant to the main variable of interest, namely, the number of industrial

Table 1
Showing the distribution of earners in Birbhum

Industry Number and Code	Number of earners by industry	Number of villages with earners	Number of blocks with earners	Ranges of earners industry-wise in blocks	
				Minimum	Maximum
1 (H)	4582	199	21	6	1701
2 (B)	3715	314	21	18	509
3 (HU)	2352	648	21	10	210
4 (P)	2012	146	21	1	227
5 (S)	1543	19	6	6	1177
6 (SB)	3886	36	6	1	1940
7 (BM)	1539	154	21	1	309
8 (IS)	1523	474	21	30	119
9 (C)	1381	372	21	15	123
10 (PC)	1139	75	15	2	351
Total	23672	1286	21		

earners. Both might be tried based on the traditional and the adaptive sampling schemes. We do not pursue with this in order to avoid a shift from our basic motivation.

In Table 1, using data from Economic Census 1990–1991 for a particular district called Birbhum in the state of West Bengal, we show how the earners by 10 specific rural unregistered industries are variously concentrated in the 21 blocks of the district composed together of 1286 villages. The 10 industries in Birbhum district which we shall consider are the following, numbered and coded: 1. Handloom (H), 2. Bamboo (B), 3. Husking (HU), 4. Pottery (P), 5. Silk (S), 6. Stone-breaking (SB), 7. Bidi-manufacturing (BM), 8. Ironsmithy (IS), 9. Carpentry (C) and 10. Paddy-crushing (PC).

Table 1 shows the high degree of disparity in the distribution of the earners through different industries in Birbhum. While 1523 ironsmiths (8) are spread over 474 villages covering all the 21 blocks, the 1543 workers in the silk industry (5) are concentrated in only 19 villages and 6 blocks; 3715 bamboo industry (2) earners are found in 314 villages and all 21 blocks while the 3886 stone-breakers (6) are localized over only 36 villages covering only 6 of the 21 blocks; paddy-crushers (10) are found in only 75 villages in 15 blocks. So, it is not easy to recommend a standard sampling design to catch these earners in sufficient numbers to throw up useful estimates of the 10 domain sizes, namely, the segments of the total numbers of these 23,672 earners by these 10 separate industries.

To effectively sample the villages, we use a traditional stratified two-stage sampling scheme with blocks as the first-stage units (fsu) and villages as the second-stage units (ssu). As the block 'sizes' are different we apply the Rao–Hartley–Cochran (RHC, 1962) scheme in the first stage for the sake of higher efficiency as well as simplicity and SRSWOR is used in the second stage. The details of the sampling scheme are described in Section 2. Some interesting features of association are found in the EC

Table 2
Showing an association of industries in their locations in Birbhum villages

	1 (H)	2 (B)	3 (HU)	4 (P)	5 (S)	6 (SB)	7 (BM)	8 (IS)	9 (C)	10 (PC)
1 (H)	199 (100)	54 (17.20)	121 (18.67)	30 (20.55)	8 (42.11)	2 (5.56)	33 (21.43)	90 (18.99)	74 (19.89)	13 (17.33)
2 (B)	54 (27.14)	314 (100)	166 (25.62)	48 (32.38)	8 (42.11)	9 (25.00)	57 (37.01)	123 (25.95)	113 (30.38)	33 (44.00)
3 (HU)	121 (60.80)	166 (52.87)	648 (100)	76 (52.05)	12 (63.16)	18 (50.00)	88 (57.14)	27 (5.70)	206 (55.38)	33 (44.00)
4 (P)	30 (15.08)	48 (15.29)	76 (11.73)	146 (100)	3 (15.79)	4 (11.11)	21 (13.64)	63 (13.29)	59 (15.86)	13 (17.33)
5 (S)	8 (4.02)	8 (2.55)	12 (1.85)	3 (2.05)	19 (100)	1 (2.78)	5 (3.25)	9 (1.90)	7 (1.88)	4 (5.33)
6 (SB)	2 (1.01)	9 (2.87)	18 (2.78)	4 (2.74)	1 (5.26)	36 (100)	7 (4.55)	20 (4.22)	14 (3.76)	1 (1.33)
7 (BM)	33 (16.58)	57 (18.15)	88 (13.58)	21 (14.38)	5 (26.32)	7 (19.44)	154 (100)	71 (14.98)	72 (19.35)	23 (30.67)
8 (IS)	90 (45.23)	123 (39.17)	272 (41.98)	63 (43.15)	9 (47.37)	20 (55.56)	71 (46.10)	474 (100)	170 (45.70)	26 (34.67)
9 (C)	74 (37.19)	113 (35.99)	206 (31.79)	59 (40.41)	7 (36.84)	14 (38.89)	72 (46.75)	170 (35.86)	372 (100)	26 (34.67)
10 (PC)	13 (6.53)	33 (10.51)	33 (5.09)	13 (8.90)	4 (21.05)	1 (2.78)	23 (14.94)	26 (5.49)	26 (6.99)	75 (100.00)

1990–1991 data of Birbhum and this motivates us in using ‘adaptive sampling’ to tackle the present issue. Since the villages are the ‘lowest stage units’ for us, we propose to use adaptive sampling to capture more villages with no restriction within the selected blocks on exploiting their associations with the industries of our interest. In Table 2, this association pattern is shown. Labeling the rows and columns of Table 2 as i and j respectively, $i, j = 1, \ldots, 10$ being the codes of the 10 industries, the entries in cell (i, j) of Table 2 are as follows:

An entry in cell (i, i) is the number of villages where there are earners by industry i, $= f_{ii}$, say. An entry in cell (i, j), $i \neq j$ is the number of villages where there are earners by both industries i and j, $= f_{ij}$, say. The parentheses show $100 \times f_{ij}/f_{jj}$, $i, j = 1, \ldots, 10$.

Table 2 shows, for example, that if a number out of the 199 villages which abound with handloom experts (1) are chosen, then huskers (3), ironsmiths (8) and carpenters (9) should also be found there. Again, if any of the 19 villages with silk workers

(5) are selected, then huskers (3), ironsmiths (8), handloom weavers (1) and bamboo artisans (7) are likely to be present there too. These possibilities encourage us to explore the possibilities of adaptive sampling as an improvement upon the original stratified two-stage scheme we start with. Thompson (1992) and Thompson and Seber (1996) give a good account of 'network' and 'adaptive sampling' techniques. Chaudhuri (2000) extended their results by showing how unbiased estimators of totals and variance estimator can be constructed for stratified multi-stage sampling with varying selection-probabilities. In the rest of the paper, we emphasize the applications of these ideas.

We shall explain and show below that in order to apply adaptive sampling one needs to suitably (a) define certain 'neighbourhoods' and (b) construct 'networks' thereby for every population unit. If one starts with any sample, a corresponding adaptive sample just adds to it the units in the networks of all the sampled units. Corresponding to an estimator for a population total based on the initial sample using any sample 'weights', the adaptive sample employs the same estimator only replacing each sample observation by the simple average of all the observations in its network.

It is easy to check that adaptive sampling gives a better estimator than the sample mean for the population mean based on simple random sampling without replacement. But such general results are not easy to claim for arbitrary sampling and estimation methods with arbitrary network formations. By a numerical exercise we illustrate below that an effective way, to achieve higher efficacy in estimating simultaneously the totals of several variables, is to form the networks 'suitably' on taking account of their mutual associations. This is done by defining networks in terms of one or more industries which are well distributed over Birbhum and which are well associated with the localized industries. As explained after Table 2, this makes it much more likely that the workers in the associated localized industries will be captured by the adaptive sample based on such networks. This method will be generally applicable whenever one can identify networks which are suitable in this sense. For this, one needs an association table similar to Table 2, which need not however be very accurate since the exact numbers of this table are not used directly in the computations. We utilize data from the EC to design our initial sample appropriately and for adaptive sampling. With our illustrations we may claim that accurate occasional follow-up surveys adopting our proposed method may justify delaying the expensive Economic Censuses over intervals of not less than 10 years.

In Section 2 we describe the sampling schemes and the estimation procedures. In Section 3 we give the performance criteria used for comparison and in Section 4 we show some of the performances of the estimators by numerical illustrations based on data from EC 1990–1991.

2. Sampling schemes and estimation methods

The original sampling scheme is chosen carefully so that the villages may be effectively sampled. To start with we treat the blocks as the first stage units (fsu) and split them into 3 strata of 7 blocks each. Using data from EC (1990–1991), the total

Table 3
Showing the distribution of earners by blocks and strata

Stratum	The total earners in 7 blocks of respective strata							Total
1	2910	1412	1211	821	683	380	284	7656
2	2879	2040	1189	889	559	397	270	8223
3	2259	2178	1132	1077	510	462	175	7793
Total								23672

numbers of earners by all the above 10 industries together in the respective blocks are used to rank the blocks $1, 2, \ldots, 21$ in decreasing order of these numbers. The blocks ranked $1, 6, 7, 12, 13, 18$ and 19 form the first stratum, those ranked $2, 5, 8, 11, 14, 17$ and 20 form the second stratum and the rest form the third. The numbers of earners through all these 10 industries in each block of the 3 strata are shown in Table 3.

In the first stage, using the entries in Table 3 as the ‘size-measures’, from each stratum we choose samples of 3 blocks independently, applying the Rao–Hartley–Cochran (RHC, 1962) sampling scheme which is described below. In the second stage, from each of the chosen blocks we independently select 20%, 12% and 10% (rounded upwards to the next integer) simple random samples (SRS) of villages without replacement (WOR). Using the samples so selected, we use appropriate estimators for the ‘total numbers of earners’ by the 10 respective industries and also a couple of ‘combinations’ of them and all of the 10 industries taken together.

2.1. Rao–Hartley–Cochran method of sampling and estimation

Suppose $U = (1, \ldots, i, \ldots, N)$ is a finite survey population with y_i as the value of a real variable of interest y. Let $Y = \sum y_i$ be the population total to be estimated, writing $\sum$ as sum over i in U. Let p_i, $(0 < p_i < 1, \sum p_i = 1)$, be known numbers, called the normed size-measures of i in U. To choose a sample of n units from U by the RHC method, one needs to first choose n positive integers $N_1, \ldots, N_n$ with their sum $\sum_n N_i = N$; then, an SRSWOR of N_1 units of U are chosen to form the first group, and then successively SRSWOR’s of sizes $N_2, \ldots, N_n$ are chosen from the remaining units of U each time $(2, \ldots, n)$, thus forming n disjoint random groups. Suppose $p_{i1}, \ldots, p_{ij}, \ldots, p_{iN_i}$ be the p_i values for the units in the ith group and let $Q_i = p_{i1} + \ldots + p_{iN_i}$, $i = 1, \ldots, n$. Then, one unit is chosen from the ith group with probability p_{ij}/Q_i for the ijth unit and this is independently repeated for each group.

For simplicity, writing y_i and p_i as the y- and p_i value of the unit drawn from the ith group, for this RHC scheme of sampling, RHC’s unbiased estimator for Y is:

$$t_{\text{RHC}} = \Sigma_n y_i \frac{Q_i}{p_i} \tag{2.1}$$

and the unbiased estimator of $\mathrm{V}(t_{\mathrm{RHC}})$ is:

$$v(t_{\mathrm{RHC}}) = C\Sigma_n\Sigma_n Q_i Q_{i'}\left[\frac{y_i}{p_i} - \frac{y_{i'}}{p_{i'}}\right]^2 \tag{2.2}$$

where $C = (\Sigma_n N_i^2 - N)/(N^2 - \Sigma_n N_i^2)$; $\Sigma_n\Sigma_n$ denotes sum over distinct pairs of the groups with no duplication and no overlap, $i' = 1, \ldots, n$, $(\neq i)$.

Note that (2.1) and (2.2) are applicable only if the y_i's are ascertainable for a sample. But, suppose $y_i = \sum_{j=1}^{M_i} y_{ij}$, is the sum of M_i second stage units (ssu) contained in the fsu i of U and is not ascertainable. Then, one may draw an SRSWOR of size m_i from these M_i ssu's and use

$$\hat{y}_i = \frac{M_i}{m_i}\Sigma' y_{ij} \tag{2.3}$$

as an unbiased estimator of y_i, writing Σ' as the sum over the sampled ssu's. An unbiased estimator for $\mathrm{V}(\hat{y}_i)$ is given by

$$v_i = M_i^2\left(1 - \frac{m_i}{M_i}\right)\frac{1}{m_i(m_i-1)}\Sigma'(y_{ij} - \bar{y}_i)^2, \tag{2.4}$$

where $\bar{y_i} = \Sigma' y_{ij}/m_i$. Now, following arguments as in Chaudhuri et al. (2000), an unbiased estimator of the variance of the unbiased estimator

$$e_{\mathrm{RHC}} = \Sigma_n \hat{y}_i \frac{Q_i}{p_i} \tag{2.5}$$

for Y in this two-stage sampling 'RHC–SRS' is:

$$v(e_{\mathrm{RHC}}) = C\Sigma_n\Sigma_n Q_i Q_{i'}\left[\frac{\hat{y}_i}{p_i} - \frac{\hat{y}_{i'}}{p_{i'}}\right]^2 + \Sigma_n \frac{Q_i}{p_i} v_i \tag{2.6}$$

where $\hat{y}_i$ and v_i are as in (2.3) and (2.4).

As a better course one may employ the RHC–RHC scheme where at the second stage an RHC sample of m_i ssu's are selected from the M_i ssu's in i, utilizing some known normed size-measures p_{ij} $(0 < p_{ij} < 1,\ \sum_{j=1}^{M_i} p_{ij} = 1)$, say, by forming m_i groups of M_{ij} ssu's in the jth group $(j = 1, \ldots, m_i)$, and with Q_{ij} as the sum of the M_{ij} values of p_{ij}'s falling in the jth group. Then, with analogous notations,

$$y_i^* = \Sigma_{m_i} y_{ij}\frac{Q_{ij}}{p_{ij}} \tag{2.7}$$

is unbiased for y_i and an unbiased estimator for the variance of y_i^* is

$$v_i^* = C_i\Sigma_{m_i}\Sigma_{m_i} Q_{ij}Q_{ij'}\left(\frac{y_{ij}}{p_{ij}} - \frac{y_{ij'}}{p_{ij'}}\right)^2, \tag{2.8}$$

where $C_i = (\Sigma_{m_i} M_{ij}^2 - M_i)/(M_i^2 - \Sigma_{m_i} M_{ij}^2)$. Then, an unbiased estimator for Y is

$$e_{\mathrm{RHC}}^* = \Sigma_n y_i^* \frac{Q_i}{p_i}$$

and an unbiased estimator of $\mathrm{V}(e^*_{\mathrm{RHC}})$ is

$$v(e^*_{\mathrm{RHC}}) = C\Sigma_n\Sigma_n Q_i Q_{i'} \left[\frac{y_i^*}{p_i} - \frac{y_{i'}^*}{p_{i'}}\right]^2 + \Sigma_n \frac{Q_i}{p_i} v_i^*,$$

where y_i^* and v_i^* are as in (2.7) and (2.8). In our present application, as the population is divided into 3 strata, we calculate these estimators for the 3 respective strata and then add the estimators and the variance estimators across the strata to get estimators for the overall population total, (i.e. the total for Birbhum district), along with variance estimators.

2.2. Adaptive sampling

In adaptive sampling, from a given survey population $U = (1,\ldots,i,\ldots,N)$ for which a total $Y = \sum y_i$ is to be estimated, initially a sample s of units is suitably drawn and further units are added to it so as to improve upon an estimator based on the initial sample by capturing more units of relevance in course of the survey utilizing the observations on the sampled units. For this, a 'neighbourhood of every unit' is defined consisting of a unit and further units linked through a well-defined reciprocal relationship. In adaptive sampling, given a unit i in a sample, every unit in its 'neighbourhood' is also to be observed in respect of the value(s) of one (or more) variable(s) to be specified. If these values satisfy certain conditions then units in the neighbourhoods of every such unit are also to be observed and checked for the conditions on their values, continuing this process until a unit of a neighbourhood ceases to satisfy the specified conditions. The set of all such units to be checked starting with a unit i is a cluster $C(i)$ around i. All the units in $C(i)$ for which the specified conditions are satisfied, together constitute a subset of $C(i)$ called a 'Network' $N(i)$ of i. The remaining units of $C(i)$ for which the conditions are not satisfied are called edge units $E(i)$ of $C(i)$. Every edge unit is regarded as a 'singleton network' consisting of a single unit for which the specified conditions are not satisfied. All networks defined corresponding to any i in U are mutually exclusive and the set of all the networks together with the singleton networks exhaust the entire population U. The entire set of units in the clusters of all the units of the original sample constitute an 'adaptive sample'.

In the present context, we start with a given list of all the 1286 villages in Birbhum district and imagine them to be arranged in a circular way. Given any village, two villages ahead of it and two villages following it in the list together with this initial village are supposed to constitute a neighbourhood of five villages. Given a village in an initial sample having say, at least one earner through carpentry (C), a cluster around it is formed by associating with this village all other villages with at least one carpenter found in the successively neighbouring villages. All the villages in such a cluster having at least one carpenter form a network. All such networks, plus the singleton networks of villages with no carpenter at all, exhaust all the villages in Birbhum.

Let $A(i)$ be a network containing a unit i and I_i be its cardinality. Then, for $t_i = (1/I_i)\sum_{j\varepsilon A(i)} y_j$, we have

$$T = \sum_{i\varepsilon U} t_i = \sum y_i = Y.$$

So, estimating Y is equivalent to estimating T.

Moreover, if U is a population of N fsu's i, M_i be the number of ssu's in the ith fsu, the jth ssu in the ith fsu containing L_{ij} third stage units (tsu), then if $Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=1}^{L_{ij}} y_{ijk}$ is to be estimated, then starting with three-stage sampling one may adopt adaptive sampling and proceed equivalently to estimate

$$T = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=1}^{L_{ij}} \frac{1}{I_{ijk}} \sum_{u\varepsilon A(ijk)} y_u,$$

which equals Y, u being any tsu, writing $A(ijk)$ as the network formed with an initial tsu ijk and I_{ijk} as the cardinality of $A(ijk)$. In forming the networks one need not be barred from crossing the limits of the units within the respective stages. Similarly, the stratified sampling case may also be covered with no barriers by strata-formation in defining the networks.

In such cases, since the estimators of $Y = T$ are specified, variance estimation is also a simple matter; in all the formulae one has to replace y_i by $t_i = (1/I_i)\sum_{j\varepsilon A(i)} y_j$, y_{ijk} by $t_{ijk} = (1/I_{ijk})\sum_{u\varepsilon A(ijk)} y_u$ and so on.

In Section 3 we discuss some criteria for evaluation of alternative estimators for Y based on various sampling procedures taking y as the number of earners by separate industries and also by certain groups of industries including the totality of all 10. In constructing the networks, we choose possible alternative industries instead of carpentry (C) alone.

3. Criteria for comparison

Given an estimator t for Y with v as an estimator of V(t), we treat $d = (t - Y)/\sqrt{v}$ as a standard normal deviate and $(t - 1.96\sqrt{v}, t + 1.96\sqrt{v})$ as a 95% confidence interval (CI) for Y calculated from a given sample. Since we are dealing with census findings, we may replicate the sampling and estimation by the same procedure $R = 1000$ times, say.

The percentage of such replicates for which the above CI covers Y is called the Actual Coverage Percentage (ACP). The closer the value of ACP is to 95, the better it is.

The average, over these replicates, of the values of $cv = 100\sqrt{v}/t$ is called the Average Coefficient of Variation (ACV). The smaller the value of ACV, the better is t as a point estimator and also the smaller is the width of the CI.

Moreover, for individual samples, the values of cv are compared in respect of various $(t, v)'s$. In the next section we present some numerical findings in quest of appropriate construction of networks to end up with apposite recommendation for policymakers.

Table 4
ACP, ACV-values based on estimated numbers of earners industry-wise from stratified RHC–SRS samples of 20%, 12% and 10% villages per selected block

Industry no. & Code	ACP (20%, 12%, 10% sampling)	ACV (20%, 12%, 10% sampling)
1 (H)	(51.6, 47.4, 44.7),	(52.4, 59.3, 62.0)
2 (B)	(85.9, 84.9, 83.0),	(36.6, 44.1, 48.1)
3 (HU)	(90.0, 88.4, 85.0),	(21.9, 26.7, 28.6)
4 (P)	(86.1, 78.8, 75.4),	(47.2, 58.9, 64.7)
5 (S)	(40.8, 29.6, 25.8),	(79.5, 69.8, 65.5)
6 (SB)	(70.2, 64.3, 61.3),	(68.7, 73.1, 73.2)
7 (BM)	(79.0, 71.7, 71.4),	(46.4, 73.1, 73.2)
8 (IS)	(92.0, 89.5, 87.9),	(24.1, 29.2, 31.7)
9 (C)	(88.6, 85.1, 82.8),	(29.8, 36.2, 39.5)
10 (PC)	(84.8, 78.0, 76.2),	(24.2, 29.3, 31.7)
3, 8 & 9	(88.1, 88.2, 87.9),	(25.4, 29.6, 32.0)

Table 5
Proper combinations of villages to form networks

Industry code	Network by	Industry code	Network by	Industry code	Network by
HU	S, C, B	SB	HU, IS, C	C	HU, IS, H
P	HU, IS, C	BM	HU, IS, C, B	PC	B, HU, IS, C, BM
S	HU, B IS, C	IS	HU, C, B		

4. Numerical illustrations by simulation from economic census 1990–1991

Remark 1. Consistently with their spread across the villages and blocks, Table 4 shows that with the original stratified RHC–SRS sample, only the huskers (3), ironsmiths (8) and carpenters (9) are rather well estimated, with the bamboo-workers (2) slightly lagging behind. All other estimates, particularly the silk-earners (5), handloom-weavers (1), stone-breakers (6) and bidi-makers (7) are too poor to be useful and this is also expected as industries 5, 1, 6 and 7 are quite strongly localized.

Remark 2. We may take the cue from Table 2 to conjecture that for a right application of adaptive sampling, networks should be so formed that to estimate the number of handloom-weavers (1), the villages with huskers (3), ironsmiths (8) and carpenters (9) in the neighbourhood of any sampled village should be treated as the network for that village with a handloom-weaver. Similarly, while estimating the number of bamboo-artisans (2), villages with huskers (3), ironsmiths (8) and carpenters (9) should be taken to form networks. Similarly, one would expect good results as per the combinations in Table 5.

Table 6
ACP, ACV based on 1000 replicates in estimating industry-wise earners by adaptive sampling with industry-specific networks. a,b are the original and adaptive sample village numbers in a single replicate industry-wise; c,d are the same over-all village numbers

Industry Number and Code	Network-industry specification		
	1 (H) (ACP, ACV, a,b),	2 (B) (ACP, ACV, a,b),	3 (HU) (ACP, ACV, a,b)
1 (H)	(70.1, 49.8, 16,27)	(47.6, 49.5, 16,23),	(64.9, 44.6, 16, 52)
2 (B)	(85.3, 35.5, 28, 32)	(89.1, 33.3, 28, 72),	(90.1, 30.2, 28, 88)
3 (HU)	(90.5, 20.0, 60, 68)	(90.9, 20.4, 60, 87),	(92.8, 16.7, 60, 244)
4 (P)	(87.3, 43.8, 13, 15)	(85.1, 42.5, 13, 18),	(88.7, 35.0, 13, 35)
5 (S)	(46.6, 79.6, 0,1)	(44.3, 81.5, 0,0),	(66.1, 77.7, 0, 1)
6 (SB)	(70.3, 68.4, 0, 0)	(70.0, 69.5, 0,1),	(75.6, 62.9, 0, 1)
7 (BM)	(77.8, 43.7, 14, 19)	(81.3, 42.0, 14, 23),	(83.9, 32.8, 14, 44)
8 (IS)	(92.4, 22.5, 31, 39)	(91.0, 22.9 31, 49),	(92.8, 19.8, 31, 108)
9 (C)	(88.2, 28.7, 28, 33)	(90.5, 27.3, 28, 40),	(89.3, 23.1, 28, 87)
10 (PC)	(88.5, 52.0, 11, 13)	(82.0, 52.3, 11, 16),	(86.8, 41.9, 11, 24)
All 10	(93.1, 20.6, -,-)	(83.7, 23.3, -,-),	(90.3, 16.9, -,-)
All but 5, 6, 7	(70.1, 49.8, -,-)	(89.1, 33.3,-,-),	(92.5, 16.7, -,-)
(c, d)	(107, 118)	(107, 151),	(107, 291)

Remark 3. It is remarkable that to estimate the number of earners by a particular industry, the proper network need not be formed by looking for neighbouring villages with earners by that same industry. Instead, a network formed by some other industries ‘associated’ with it, the ‘association’ being empirically revealed by a table similar to Table 2, should fare better. We have no explanation for a particular industry to be associated in a desirable manner with some other industry but we feel that we should be guided by empirical evidences alone as in Table 2.

We now illustrate some numerical evidence.

Remark 4. The number of villages to be covered in adaptive sampling may far exceed the size of the original sample depending on the choice of industry in defining the networks. But this need not be prohibitive because (1) the neighbouring villages are geographically so and (2) in rural areas, ascertaining the presence/absence of earners by specific industries locally is not laborious or expensive. In a single replicate, a village with 7 or more industry-type earners is rare and so the blanks appear in the tables. With networks as in Table 6 the estimation about handloom (1), silk (5), stone-breaking (6) and to some extent bidi-making (7) too, continue to be poor.

Remark 5. When the industries pottery (4), silk (5) and stone-breaking (6) are used for forming the networks, Table 7 shows that the quality of estimation rather declines and adaptive sampling hardly covers any additional villages over the original sample. So, these 3 industries should not be used in defining networks. This is consistent with

Table 7
ACP, ACV based on 1000 replicates in estimating industry-wise earners by adaptive sampling with industry-specific networks. a,b are the original and adaptive sample village numbers in a single replicate industry-wise; c,d are the same over-all village numbers

Industry Number and Code	Network-industry specification		
(1)	4 (P) (ACP, ACV, a,b), (2)	5 (S) (ACP, ACV, a,b), (3)	6 (SB) (ACP, ACV, a,b) (4)
1 (H)	(51.2, 52.1, 16, 16)	(51.3, 52.3, 16, 16),	(51.6, 52.4, 16, 16)
2 (B)	(86.7, 36.2, 28, 29)	(85.8, 36.6, 28, 28),	(85.9, 36.6, 28, 28)
3 (HU)	(90.0, 21.7, 60, 60)	(90.3, 21.2, 60, 60),	(89.5, 21.9, 60, 60)
4 (P)	(88.0, 44, 2, 13, 15)	(86.2, 47.2, 13, 13),	(86.4, 47.1, 13, 13)
5 (S)	(40.8, 79.5, 0, 0)	(40.7, 79.2, 0, 0),	(40.8, 79.5, 0, 0)
6 (SB)	(69.7, 68.1, 0, 0)	(70.2, 68.7, 0, 0),	(78.0, 63.7, 0, 0)
7 (BM)	(81.1, 44.7, 14, 14)	(78.5, 46.0, 14, 14),	(80.8, 46.2, 14, 14)
8 (IS)	(91.6, 23.5, 31, 31)	(91.7, 24.0, 31, 31),	(92.1, 23.9, 31, 31)
9 (C)	(88.1, 29.0, 28, 28)	(88.5, 29.7, 28, 28),	(89.1, 29.5, 28, 28)
10 (PC)	(80.7, 32.2, 11, 11)	(80.2, 57.2, 11, 11),	(78.5, 58.8, 11, 11)
(c, d)	(107, 109)	(107, 107)	(107, 107)

Table 5 which does not recognize these three industries to suit any of the industries in network building.

Remark 6. Particularly for the industries silk (5), handloom (1) and stone-breaking (6), a phenomenal improvement is achieved when networks as in Table 8 are used. But the resultant sample village size may increase by a large amount and this may sometimes be impractical.

Remark 7. Except for the silk (5) earners, the three networks as in Table 9 achieve commendable improvements over the original sampling and the enhanced sample-sizes too remain sensibly low.

Remark 8. From our arguments and illustrations it is evident that adaptive sampling may be made fruitful by dint of appropriate network formations based on empirical verifications of ‘associations’ among the various combinations of industries in a given region, namely Birbhum district in our illustration. Adaptive sampling leads to increased sample-size in course of additional capture of relevant units throwing additional data for the improvement in estimation. However, with mere enhancement of sampling rates, without exploitation of the ‘associations’ among the industries of interest, one cannot expect or claim proportionate betterment in estimation, i.e., ‘the unitary method’ does not apply. So, in practice, tables similar to Tables 2 and 5 should be utilized in setting up guidelines in network formation. Proper networking also depends on the specific industry and industry groups which are considered important for accurate estimation. In the present case if silk (5), stone-breaking (6) and handloom (1) totals are the most important ones to estimate, relatively more costly and penetrative networking will be needed as illustrated.

Table 8
ACP, ACV based on 1000 replicates in estimating industry-wise earners by adaptive sampling with industry-specific networks. a,b are the original and adaptive sample village numbers in a single replicate industry-wise; c,d are the same over-all village numbers

Industry Number and Code	Network-industry specification		
	8 (IS) (ACP, ACV, a,b),	3 (HU) and 8 (IS) (ACP, ACV, a,b),	3 (HU) and 9 (C) (ACP, ACV, a,b)
(1)	(2)	(3)	(4)
1 (H)	(74.3, 47.4, 16,24),	(81.7, 33.2, 16, 61),	(78.7, 37.3, 16, 70)
2 (B)	(86.7, 32.6, 28, 43)	(90.3, 26.5, 28, 105),	(91.0, 28.1, 28, 114)
3 (HU)	(92.8, 18.7, 60, 82)	(93.6, 15.2, 60, 285),	(92.0, 15.4, 60, 292)
4 (P)	(85.4, 41.2, 13, 20)	(84.6, 30.6, 13, 46),	(89.1, 33.0, 13, 52)
5 (S)	(57.7, 79.0, 0, 1),	(82.8, 56.4, 0, 2),	(73.6, 72.0, 0, 2)
6 (SB)	(79.0, 60.4, 0, 1),	(80.7, 49.9, 0, 5),	(74.8, 57.1, 0, 3)
7 (BM)	(82.5, 38.6, 14, 23),	(84.0, 27.3, 14, 57),	(82.5, 29.2, 14, 62)
8 (IS)	(92.9, 22.0, 31, 71),	(93.4, 18.3, 31, 184),	(93.1, 18.4, 31, 154)
9 (C)	(90.4, 24.9, 28, 46),	(90.5, 20.7, 28, 121),	(92.5, 19.8, 28, 169)
10 (PC)	(86.9, 50.3, 11, 16),	(87.7, 34.8, 11, 26),	(88.2, 36.4, 11, 31)
All 10	(92.6, 17.9, -,-),	(91.5, 11.9, -,-),	(92.1, 13.8, -,-)
(c, d)	(107, 147),	(107, 485),	(107, 480)

Table 9
ACP, ACV based on 1000 replicates in estimating industry-wise earners by adaptive sampling with industry-specific networks. a,b are the original and adaptive sample village numbers in a single replicate industry-wise; c,d are the same over-all village numbers

Industry Number and Code	Network-by		
	1 (H), 2 (B), 4 (P) (ACP, ACV, a,b),	9 (C) (ACP, ACV, a,b),	8 (IS) and 9 (C) (ACP, ACV, a,b)
1 (H)	(83.0, 41.5, 16, 47),	(78.6, 47.1, 16, 27),	(82.2, 39.6, 16, 43)
2 (B)	(88.5, 28.0, 28, 90),	(87.9, 34.3, 28, 47),	(88.5, 30.0, 28, 77)
3 (HU)	(90.5, 18.0, 60, 113),	(90.7, 20.3, 60, 89),	(92.8, 17.5, 60, 148)
4 (P)	(91.6, 30.0, 13, 29),	(86.9, 42.8, 13, 21),	(84.6, 38.3, 13, 37)
5 (S)	(56.8, 76.4, 0, 2),	(48.2, 81.6, 0, 0),	(62.5, 77.8, 0, 3)
6 (SB)	(70.0, 69.0, 0, 1),	(68.0, 66.4, 0, 1),	(83.2, 51.5, 0, 6)
7 (BM)	(80.6, 37.7, 14, 29),	(76.4, 40.6, 14, 31),	(23.0, 32.6, 14, 46)
8 (IS)	(91.0, 21.4, 31, 65),	(93.3, 21.6, 31, 52),	(93.2, 19.2, 31, 139)
9 (C)	(89.4, 24.4, 28, 55),	(90.4, 25.4, 28, 78),	(90.1, 21.0, 23, 132)
10 (PC)	(89.5, 41.0, 11, 19),	(83.3, 49.9, 11, 14),	(87.7, 41.1, 11, 21)
All 10	(93.4, 16.4, -,-),	(89.4, 20.2, -,-),	(93.2, 14.6, -,-)
(c, d)	(107, 216)	(107, 157),	(107, 319)

Next, for a particular single sample, the original and the corresponding adaptive one obtained by various industry-specific network formations, the estimated total industry-wise numbers along with estimated coefficients of variation are shown in Table 10.

Table 10
Estimated earner-numbers t by industry along with coefficients of variation $cv(t)$ single sample-wise

Industry Number & Code	True number (Y)	Estimated number (t) by single-sample, cv (%)			
		Original sample	Adaptive samples		
			Network by 1 (H)	Network by 2 (B)	Network by 3 (HU)
1 (H)	4582	6971, 84	7563, 54	6256, 96	7187, 56
2 (B)	3715	4823, 43	4930, 40	4482, 28	5255, 26
3 (HU)	2352	3254, 14	3118, 14	3818, 13	2982, 17
4 (P)	2012	1958, 38	1848, 39	2191, 33	2983, 28
5 (S)	1543	0,0	11,107	0, 0	3, 94
6 (SB)	3886	0,0	0,0	31, 126	27, 140
7 (BM)	1539	639, 35	722, 33	1027, 24	1001, 20
8 (IS)	1523	1254, 16	1131, 14	1611, 16	1671, 13
9 (C)	1381	1673, 35	1616, 34	1959, 19	1465, 21
10 (PC)	1139	2581, 48	2798, 45	1886, 43	2264, 36
All 10	23672	23155, 27	23737, 20	23259, 23	24836, 16
	Estimated earner-number (t) and cv (%) using network by industry				
	8 (IS)	3 (HU) & 8 (IS)	8 (IS) & 9 (C)	3 (HU) & 9 (C)	2 (B), 8 (IS) & 9 (C)
1 (H)	6741, 55	7488, 47	4934, 64	6448, 53	3895, 41
2 (B)	5651, 34	5661, 24	5160, 28	4519, 24	5169, 17
3 (HU)	3295, 14	3045, 14	3177, 13	2758, 16	2968, 14
4 (P)	2249, 34	3298, 26	2988, 23	3234, 25	3251, 23
5 (S)	9, 107	15, 58	5, 107	5, 88	8, 86
6 (SB)	202, 88	338, 80	323, 76	29, 43	310, 74
7 (BM)	794, 30	1241, 16	757, 26	1007, 18	932, 16
8 (IS)	1525, 14	1889, 12	1840, 14	1754, 12	1784, 11
9 (C)	1831, 20	1696, 14	1738, 16	1934, 17	2095, 16
10 (PC)	2552, 43	2300, 34	2289, 46	2101, 37	1893, 38
All 10	24848, 16	26970, 11	23214, 14	23789, 15	22305, 8

Remark 9. If one's interest is chiefly to estimate the number of earners by silk industry (5), which actually in Birbhum is known to be 1543 in 1990–1991, then a sample originally designed by us may not capture any one of them and so result in a 'zero' estimate. Even by adaptive technique, this estimate may be raised to at most 15 and that is also borne by a heavy bulk of a 58% cv. So, a somewhat drastically different sampling scheme needs to be tried. For the case of stone-breakers (6), the situation is roughly similar. For the huskers (3) and the ironsmiths (8), our original sampling scheme is good enough.

However, it is remarkable that for all the 10 types of earners combined, we can achieve an appreciable reduction in cv by using adaptive scheme with the right networks, as compared to the original one. For other categories too, similar observations can be made from our exercise above.

So the message is that to cover most of the industries, adaptive sampling should be useful. For industries which are evenly spread out over the region, the original sample can be aptly designed to give good estimates; but for the very highly localized industries more special efforts are needed and for this further research is a must. It can be easily seen following Thompson and Seber (1996) that in the case where the initial sample is a simple random sample, the efficiency of adaptive sampling based on it will be large if the variation within the networks is high. The idea of network construction on the basis of associations, as proposed in this paper, is likely to lead to networks which will have high within-network variability. Moreover, adaptive sampling is operational if (1) the 'added units' are easy to survey and (2) the common association ensures geographical contiguity thereby keeping down the additional costs.

To improve upon these estimates we could consider an RHC–RHC scheme where the villages are also selected using the RHC technique with a suitable size measure. This may lead to a better original sample than the RHS-SRSWOR sample illustrated here.

Finally, we are inclined to draw a reader's attention to Table 4 and column 3 of Table 8 to notice how, for 1 (H), 5 (S) and 6 (SB) the respective ACP's have dramatically moved, through the network formations by 3 (HU) and 8 (IS), from mere 44.7, 25.8 and 61.3 based on the original sample to respectively 81.7, 82.8 and 80.7 along with reductions also in the ACV's. More importantly, (a) for the same network better results are also derived for the other variables and (b) for other networks these advantages are not quite apparent.

Acknowledgements

We are thankful to two referees and an editorial board member for their recommendations leading to improvement upon an earlier draft of the paper. We also thank Mr. Rajiv Parekh for his computational assistance at all stages of this work.

References

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Roy. Statist. Soc. Ser. B 36, 192–236.

Chaudhuri, A., 2000. Network and adaptive sampling with unequal probabilities. Cal. Statist. Assoc. Bull. 50, 237–253.

Chaudhuri, A., Adhikary, A.K., Dihidar, S., 2000. Mean square error estimation in multistage sampling. Metrika 52, 115–131.

Rao, J.N.K., Hartley, H.O., Cochran, W.G., 1962. On a simple procedure of unequal probability sampling without replacement. J. Roy. Statist. Soc. B 24, 482–491.

Thompson, S.K., 1992. Sampling. Wiley, New York.

Thompson, S.K., Seber, G.A.F., 1996. Adaptive Sampling. Wiley, New York.