Sample-size-restrictive adaptive sampling: an application in estimating localized elements

Arijit Chaudhuri*,[1], Mausumi Bose, Kajal Dihidar

Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 108, India

Abstract

In sample surveys sometimes one encounters a situation where, for many sampling units, one or more variables of interest are valued zero or negligibly low while for some other units they are substantial because of heavy localization of the high-valued units in certain segments. Estimation may then be inaccurate if a chosen sample fails to capture enough of the high-valued units. In such situations, adaptive sampling, as an extension of the initial sample to capture additional high-valued units, may be more serviceable. However, the size of an adaptive sample may often far exceed that of the initial sample. In this paper we present a method to put desirable constraints on the adaptive sample-size to keep the latter in check. To examine the efficacy of this method, we illustrate its application to estimate total numbers of rural earners through specific vocations in a given district in India simultaneously for several vocations.

MSC: 62DO5

Keywords: Adaptive sampling; Finite survey population; Restricted sample-size; Stratified two-stage sampling; Varying probabilities

1. Introduction

Adaptive sampling is a useful technique in effective exploration of rare and localized units as it provides for extending the initially chosen sites to cover neighborhoods of those

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

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

where the materials of interest are sighted. It has been used in the exploration of mineral deposits in unknown regional pockets, in studies to estimate the number of rare animals or plants in an area and other similar surveys. For some examples of the use of this technique, we refer to Thompson and Seber (1996). An inherent problem with this technique is that the size of the adaptive sample may far exceed that of the initial sample. Salehi and Seber (1997, 2002) addressed this issue and gave certain solutions. We provide here a possible alternative and easy remedy by permitting an in-built procedure to keep a check on the "adaptive sample" size.

Adpative sampling was introduced by Thompson (1990) and later further developed by Thompson (1992) and Thompson and Seber (1996). Chaudhuri (2000) clarified that any sampling method admitting an unbiased estimator for a survey population total as well as one for its variance may be extended to an adaptive one yielding the corresponding unbiased estimators. Chaudhuri et al. (2004) showed how this method may be used to work out improved estimators through an adaptive sample, starting from a suitably designed initial one on which the adaptive sample is based. Using live data, they illustrate how this technique could be used to obtain serviceable estimators for the total numbers of earners separately through 10 different industries in an Indian district. Importantly, they have demonstrated how networks, needed for adaptive sampling, may be effectively defined utilizing associations among several variables for the totals of each of which an efficient estimation procedure is in demand at the same time. However, their method had no provision to keep the final sample-size in check and in that work it was evident, as it should be, that if the adaptive sample is to be really effective in estimation for all the industries simultaneously, its size may far exceed that of an initial sample.

In this paper, we show how to keep a check on the realized sample-size and illustrate the method using the same practical problem as considered in Chaudhuri et al. (2004). In the process, we revise the specific definitions of 'neighborhood' and 'network' required for the method applied to render our presented procedure more realistic. The sampling scheme of Chaudhuri et al. (2004) is also revised and the generalized regression technique is additionally used for a possible improvement upon the traditional estimators.

In Section 2, we give an example suited to adaptive sampling as a follow-up of a specific two-stage sample for which alternative estimators are spelled out. Section 3 shows how to obtain estimators from adaptive samples and how to keep the size of the adaptive sample in check. Using live data, Section 4 numerically illustrates how one may achieve improved estimation and restrict the sample-size. Our comments and recommendations are given in the concluding Section 5.

2. Motivation for adaptive sampling in a practical setup and constraining its size

From Chaudhuri et al. (2004) we quote the distribution of numbers of rural earners in 10 different unorganized non-agricultural vocations among all the villages in a particular district in India. The data are from the Economic Census (EC), held in 1990.

In India, nationwide large-scale socio-economic surveys covering diverse aspects are being conducted routinely over more than the last 50 years. But no survey is expected to be specifically designed and executed with any regularity to estimate the number of persons earning through each of the principal non-agricultural vocations in the unorganized sector as illustrated above. However, it is necessary to obtain these estimates regularly in order to effectively ascertain the nation's gross domestic product (GDP). Chaudhuri et al. (2004) illustrated a possible undertaking of a sample survey effectively utilizing available data from the Economic Census (EC), 1990 and the Indian Population Census (IPC), 1991.

We modify here the two-stage stratified sampling with unequal probabilities of selection used in Chaudhuri et al. (2004). As the district concerned has 21 administrative blocks we split them into 3 strata of 7 each, taking account of the aggregated numbers of earners through all the 10 varieties in strata formation. From each stratum, 3 blocks are selected adopting Rao, Hartley and Cochran's (RHC, 1962) celebrated scheme of sampling. The total number of earners in a block through all these 10 industries together is available from the EC and this is taken as the size-measure for the block. From each block chosen, 20% of its villages, rounded upward to integers, are again sampled following the RHC scheme. The village population size as known from the IPC, is taken as the size-measure in village selection.

We briefly recall that in choosing a sample of n units, namely, the blocks bearing normed size-measures $p_i\,(0 < p_i < 1,\, \sum_1^N p_i = 1)$, the RHC-scheme is applied by forming n groups with N_i units in the ith group by simple random sampling without replacement (SRSWOR) out of the N units such that, writing Σ_n as sum over n groups, N_i are positive integers subject to $\Sigma_n N_i = N$. Denoting by Q_i the sum of the normed sizes of the N_i units falling in the ith group, a unit i_k, say, of the ith group is chosen with probability p_{i_k}/Q_i. This is independently repeated across the n groups.

Our objective is to estimate the total numbers of earners for each of the 10 industries separately. To keep the notation simple, we use the generic notation y to denote the variable of interest, namely, the number of earners through any one particular industry. Thus, y_i generically denotes the number of earners through any one chosen industry in the ith block and $Y = \sum_1^N y_i$ thus denotes the total number of earners through this industry in the entire district. Thus, the same notation serves for all the 10 industries, simply keeping in mind that when we are estimating the total number of earners for any one chosen industry, we should start with the value of y_i for that industry.

Let p_i be the normed size-measure corresponding to y_i. Then, RHC's unbiased estimator for $Y = \sum_1^N y_i$ is

$$t = \sum_n \frac{Q_i}{p_i} y_i. \tag{2.1}$$

In case y_i is not ascertainable, as in the present case where the ith block is composed of M_i villages, then m_i of the villages may again be selected by the RHC method. Let y_{ij} denote the y-value for the jth village of the ith block. Moreover, Σ_{mi}, M_{ij}, Q_{ij} will denote entities corresponding, respectively, to Σ_n, N_i and Q_i. Then, an unbiased estimator for Y is

$$e_R = \sum_n \frac{Q_i}{p_i} \left(\Sigma_{mi} \frac{Q_{ij}}{p_{ij}} y_{ij} \right). \tag{2.2}$$

For simplicity, we shall write $\hat{y}_i = \Sigma_{mi} \frac{Q_{ij}}{p_{ij}} y_{ij}$ and $C = (\Sigma_n N_i^2 - N)/(N^2 - \Sigma_n N_i^2)$, $C_i = (\Sigma_{mi} M_{ij}^2 - M_i)/(M_i^2 - \Sigma_{mi} M_{ij}^2)$. From RHC (1962) and Chaudhuri et al. (2000) we have an unbiased estimator for the variance of e_R as

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

writing $v_i = C_i \Sigma_{mi} \Sigma_{mi} Q_{ij} Q_{ij'} (\frac{y_{ij}}{p_{ij}} - \frac{y_{ij'}}{p_{ij'}})^2$ and $\Sigma_n \Sigma_n$ as sum over non-overlapping pairs of the n groups of blocks, $\Sigma_{mi} \Sigma_{mi}$ as that over the villages.

Cassel, Särndal and Wretman's (CSW, 1976) generalized regression (greg) technique is here employed by way of a possible improvement upon the above estimator e_R in the following alternative ways. A greg estimator was not however discussed in Chaudhuri et al.'s (2004) work. The motivation of greg is to improve upon the original estimator by using an auxiliary variable which is well correlated (positively) with the variable of interest.

The total number of all non-agricultural workers in a village, as found in the IPC, is treated as the single regressor. Note that this is different from the size variable used for block selection which is the total number of earners through the 10 industries as obtained from EC. We write x_{ij} as the number of non-agricultural workers in the jth village of ith block, $x_i = \Sigma_{M_i} x_{ij}$ and $X = \Sigma_N x_i$. We consider the following alternative models for motivating some greg estimators which may be considered for possible improvements over e_R:

Model M_1 : $y_{ij} = \beta_i x_{ij} +$ error.
Model M_2 : $y_{ij} = \beta x_{ij} +$ error.
Model M_3 : $y_i = \theta x_i +$ error.

M_1 and M_2 are used to first get estimates of y_i using y_{ij} values from the second stage sample and corresponding x_{ij} values from IPC data. These are then used to find estimates of Y. M_3 is used to derive regression estimates of Y by using the estimates of y_i from M_2 and the x_i values from the IPC. M_3 is derived from M_2, but we shall show below that M_3 motivates specific regression estimators for Y.

The estimators motivated by M_1 and M_2, respectively, are

$$g_{11} = \Sigma_n \frac{Q_i}{p_i} \left[\Sigma_{mi} \frac{Q_{ij}}{p_{ij}} (y_{ij} - \hat{b}_i x_{ij}) + \hat{b}_i x_i \right], \tag{2.3}$$

$$g_{12} = \Sigma_n \frac{Q_i}{p_i} \left[\Sigma_{mi} \frac{Q_{ij}}{p_{ij}} (y_{ij} - \hat{b} x_{ij}) + \hat{b} x_i \right], \tag{2.4}$$

while the estimator motivated by M_3 is

$$g_{22} = g_{12} + \hat{B} \left(X - \Sigma_n \frac{Q_i}{p_i} x_i \right). \tag{2.5}$$

In the above,

$$\hat{b}_i = \frac{\Sigma_{mi} y_{ij} x_{ij} R_{ij}}{\Sigma_{mi} x_{ij}^2 R_{ij}}, \quad \hat{b} = \frac{\Sigma_n (Q_i/p_i) \Sigma_{mi} y_{ij} x_{ij} R_{ij}}{\Sigma_n (Q_i/p_i) \Sigma_{mi} x_{ij}^2 R_{ij}}, \quad R_{ij} = \frac{(1-(p_{ij}/Q_{ij}))}{(p_{ij} x_{ij}/Q_{ij})},$$

$$\hat{B} = \frac{\Sigma_n [\Sigma_{mi} (Q_{ij}/p_{ij})(y_{ij} - \hat{b} x_{ij}) + \hat{b} x_i] x_i R_i}{\Sigma_n x_i^2 R_i}, \quad R_i = \left(1 - \frac{p_i}{Q_i}\right) \Big/ \left(\frac{p_i x_i}{Q_i}\right).$$

Here R_{ij}, R_i parallel the choice $\frac{1-\pi_i}{\pi_i x_i}$ in the greg estimator of Cassel et al. (1976).

For simplicity, we may write g_{11}, g_{12} and g_{22} in (2.3), (2.4) and (2.5) as

$$g_{11} = \Sigma_n \frac{Q_i}{p_i} \hat{y}_i(1), \quad g_{12} = \Sigma_n \frac{Q_i}{p_i} \hat{y}_i(2) \text{ and } g_{22} = \Sigma_n \frac{Q_i}{p_i} \hat{y}_i(2) h_i,$$

say. Next, let $e_{ij}(1) = y_{ij} - \hat{b}_i x_{ij}$, $e_{ij}(2) = y_{ij} - \hat{b} x_{ij}$ and $e_i = \hat{y}_i(2) - \hat{B} x_i$.

Then, the estimators of MSE's of g_{11}, g_{12} and g_{22} are, respectively,

$$v_1 = C \Sigma_n \Sigma_n Q_i Q_{i'} \left(\frac{\hat{y}_i(1)}{p_i} - \frac{\hat{y}_{i'}(1)}{p_{i'}}\right)^2 + \Sigma_n \frac{Q_i}{p_i} v_i(1),$$

$$v_2 = C \Sigma_n \Sigma_n Q_i Q_{i'} \left(\frac{\hat{y}_i(2)}{p_i} - \frac{\hat{y}_{i'}(2)}{p_{i'}}\right)^2 + \Sigma_n \frac{Q_i}{p_i} v_i(2)$$

and

$$v_3 = C \Sigma_n \Sigma_n Q_i Q_{i'} \left(\frac{e_i h_i}{p_i} - \frac{e_{i'} h_{i'}}{p_{i'}}\right)^2 + \Sigma_n \frac{Q_i}{p_i} h_i v_i(2),$$

where $v_i(1) = C_i \Sigma_{mi} \Sigma_{mi} Q_{ij} Q_{ij'} \left(\frac{e_{ij}(1)}{p_{ij}} - \frac{e_{ij'}(1)}{p_{ij'}}\right)^2$ and $v_i(2) = C_i \Sigma_{mi} \Sigma_{mi} Q_{ij} Q_{ij'} \left(\frac{e_{ij}(2)}{p_{ij}} - \frac{e_{ij'}(2)}{p_{ij'}}\right)^2$.

If enough non-zero valued y_{ij}'s are not covered in a selected sample then improvements upon e_R may not be effected by any of g_{11}, g_{12} or g_{22}. So, it is considered important, rather imperative, to enhance the "information-content" in a realized sample by extending from the initial sample to an adaptive sample and accordingly revise each of e_R, g_{11}, g_{12}, g_{22}, basing them each on a finally extended adaptive sample. Here, for simplicity, we revise only y_{ij}'s for the adaptive sample but not the x_{ij}'s.

In the present empirical situation we are motivated to implement an adaptive sample by studying the mutual relations of association among the 10 industries across the 1286 villages in our illustrated district by a reference to Table 2, partially reproduced from Chaudhuri et al. (2004).

The figures in the parentheses in table 2 indicate percentages of the values in the rows for the respective columns in terms of the diagonals to which the respective columns correspond.

From Table 1 it seems that if an initial sample is chosen with selection probabilities that make no use of the distribution of the earners among the respective villages, for

Table 1
Showing a distribution of rural industry-specific earners

	Handloom (H) 1	Bamboo (B) 2	Husking (HU) 3	Pottery (P) 4	Silk (S) 5	Stone-breaking (SB) 6
No. of earners	4582	3715	2352	2012	1543	3886
No. of villages	199	314	648	146	19	36
	Tobacco (T) 7	Iron smithy (IS) 8	Carpentry (C) 9	Paddy-crushing (PC) 10	Total	
No. of earners	1539	1523	1381	1139	23672	
No. of villages	154	474	372	75	1286	

Table 2
Presenting the respective numbers of villages with earners industry-wise showing a specimen of association of the industries in the district

	1(H)	3(HU)	4(P)	5(S)	8(IS)	10(PC)
1(H)	199 (100)	121 (18.67)	30 (20.55)	8 (42.11)	90 (18.99)	13 (17.33)
3(HU)	121 (60.80)	648 (100)	76 (52.05)	12 (63.16)	27 (5.70)	33 (44.00)
4(P)	30 (15.08)	76 (11.73)	146 (100)	3 (15.79)	63 (13.29)	13 (17.33)
5(S)	8 (4.02)	12 (1.85)	3 (2.05)	19 (100)	9 (1.90)	4 (5.33)
8(IS)	90 (45.23)	272 (41.98)	63 (43.15)	9 (47.37)	474 (100)	26 (34.67)
10(PC)	13 (6.53)	33 (5.09)	13 (8.90)	4 (21.05)	26 (5.49)	75 (100.00)

example if there is no or scant representation in the sample of the 19 villages with silk related or of the 75 villages accommodating the paddy-crushing earners, then appropriate estimation of the numbers of earners through these industries will be of dubious levels of accuracy.

Table 2 indicates for example, that of the 19 villages in which silk-earners live, 63.16% have earners by husking and 47.37% have iron-smiths. So, if the sample contains some of the

12 villages with earners by husking, or some of the 9 with iron-smiths, then, through these sampled villages some of the silk-earners may be reached. Such associations among the industries through village-wise co-inhabitance may be utilized to get estimates of the number of silk-related earners. Similarly, this can be done for the other industries as well. In this way, Table 2 is exclusively used to effectively construct networks of the villages exploiting this association to enhance the information content in a suitably extended adaptive sample. This however may entail excessive sample coverage beyond one's means. So, it is necessary to specify a way to keep the resulting size of an adaptive sample in check.

3. Adaptive sampling and constraining it in size

Let $U=(1,\ldots,i,\ldots,N)$ denote a survey population, y_i the value of y on its ith unit and s be a sample chosen from U with a probability $p(s)$. Let $N(i)$ denote a uniquely defined neighborhood of units corresponding to i. If y_i fails to satisfy a condition, say, C^*, then i itself is a 'Singleton Network' for i. If y_i satisfies C^*, then let $C(i)$ be a cluster of i containing i itself and all units in its neighborhood. Again, C^* is checked for the units in $C(i)$ and if it is satisfied for any unit, units in its neighborhood are added to $C(i)$. This process of scanning and adding units continues, stopping only on reaching units with C^* unsatisfied. The units in the clusters with C^* unsatisfied are the edge-units of the cluster which of course are each a singleton network. The cluster $C(i)$ with all its edge-units dropped is the 'Network of i' denoted by $A(i)$. Regarding each "Singleton" network also as a network, it follows that all the networks are mutually non-overlapping and they together exhaust the entire population. Consequently, if we denote by m_i, the cardinality of $A(i)$ and define

$$t_i=\frac{1}{m_i}\sum_{j\varepsilon A(i)} y_j,$$

it follows that $T=\sum_{i=1}^{N} t_i$ equals $Y=\sum_{i=1}^{N} y_i$. So, estimating Y using the survey data $(s, y_i|i\varepsilon s)$ is equivalent to estimating T using the observations $(s, t_i|i\varepsilon s)$. Writing $\underline{Y}=(y_1,\ldots,y_i,\ldots,y_N)$ and $\underline{T}=(t_1,\ldots,t_i,\ldots,t_N)$, if one employs an estimator $t=t(s,\underline{Y})$ which is unbiased for Y then $t=t(s,\underline{T})$, is also unbiased for T and hence for Y as well. If we follow this procedure then we must observe that instead of covering only the original sample s, we have to effectively cover the units in $A(s)$, the union of $A(i)$ over i in s, which is an extension of s. This $A(s)$ is an Adaptive sample corresponding to s and this process of extending the sample from s to $A(s)$ is called Adaptive sampling.

An inherent hazard in Adaptive sampling is that compared to the size n of s, that of $A(s)$, say, v may be exorbitantly larger. So, to at least partially cut down the additional cost of surveys, our proposal is to modify Adaptive Sampling into a "Size-constrained Adaptive Sampling". To implement the latter we propose that after ascertaining the sets $A(i)$, for $i\varepsilon s$, one confines the determination of the values of y_j's only for $B(i)$, $i\varepsilon s$, where $B(i)$'s are suitable subsets of $A(i)$ to be drawn by simple random sampling without replacement. Writing l_i as the cardinality of $B(i)$, we suggest that $B(i)$'s are to be chosen so that $\sum_{i\varepsilon s} l_i$ may not exceed a predetermined limit, say, L, which may be fixed as a certain fraction of

$\sum_{i \varepsilon s} m_i$. Size-constrained Adaptive sample corresponding to s is then $B(s)$, which is the union of $B(i)$ over i in s. Corresponding to t_i we now define

$$e_i = \frac{1}{l_i} \sum_{j \varepsilon B(i)} y_j$$

and employ, instead of $t = t(s, t_i | i \varepsilon s)$, an estimator $t = t(s, e_i | i \varepsilon s)$, which is also unbiased for Y.

Let E_R, V_R denote expectation, variance operators with respect to SRSWOR of $B(i)$ from $A(i)$'s independently across i in s, E_p, V_p the same over the initial sampling using the design p and $E = E_p E_R$, $V = E_p V_R + V_p E_R$ the overall expectation, variance operators. Then, since $E_R(e_i) = t_i$, $v_R(e_i) = (\frac{1}{l_i} - \frac{1}{m_i}) \frac{1}{(l_i - 1)} \sum_{j \varepsilon B(i)} (y_j - e_i)^2$ satisfies

$$E_R(v_R(e_i)) = V_R(e_i) = \left(\frac{1}{l_i} - \frac{1}{m_i}\right) \left(\frac{1}{m_i - 1}\right) \sum_{j \varepsilon A(i)} (y_j - t_i)^2$$

and it is easy to work out an unbiased estimator for $V(t(s, e_i | i \varepsilon s))$ if $t(s, y_i | i \varepsilon s)$ is an unbiased estimator for Y admitting an unbiased estimator for the variance of this estimator. For example, if

$$t(s, y_i | i \varepsilon s) = \sum_{i \varepsilon s} \frac{y_i}{\pi_i}, \quad \pi_i = \sum_{s \ni i} p(s) > 0$$

with s containing a fixed number of distinct units for every s with $p(s) > 0$, then, provided $\pi_{ij} = \sum_{s \ni i, j} p(s) > 0$ for every i, j then an unbiased estimator for $V(t(s, e_i | i \varepsilon s))$ can easily be seen to be

$$v = \sum_{i \varepsilon s} \frac{v_R(e_i)}{\pi_i} + \sum_{i < j \varepsilon s} \sum \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2,$$

following the arguments in Raj (1968).

From Chaudhuri (2000) it is apparent that, for the stratified two-stage sampling with RHC scheme employed in both stages, it is simple to work out the versions of e_R, g_{11}, g_{12}, g_{22} and the estimators of their variance or MSE based on the adaptive sample corresponding to the initial one. In order to note that the corresponding revisions on size-constraining as indicated above may also be done in an obvious way, let us illustrate this with reference to an unstratified single-stage sampling with the constraining of the size of an adaptive sample. The corresponding expressions for stratified two-stage sampling by RHC method in both the stages is evident from this and hence is omitted here to save space. Using the notations already introduced in the previous sections, corresponding to

$$t_R = \Sigma_n y_i \frac{Q_i}{p_i},$$

let $f_R = \Sigma_n t_i \frac{Q_i}{p_i}$, $g_R = \Sigma_n e_i \frac{Q_i}{p_i}$. Then,

$$V_p(f_R) = A\left(\Sigma \frac{t_i}{p_i} - T\right)^2, \quad A = \frac{\Sigma_n N_i^2 - N}{N(N-1)},$$

$$v_p(f_R) = C\Sigma_n Q_i\left(\frac{t_i}{p_i} - f_R\right)^2 = C\left(\Sigma_n t_i^2 \frac{Q_i}{p_i^2} - f_R^2\right)$$

$$v_R(g_R) = \Sigma_n\left(\frac{Q_i}{p_i}\right)^2 v_R(e_i),$$

and $f_i = e_i^2 - v_R(e_i)$ satisfies $E_R(f_i) = t_i^2$.

So, $\hat{v}(g_R) = v_R(g_R) + C[\Sigma_n \frac{Q_i}{p_i^2} f_i - (g_R^2 - \Sigma_n\left(\frac{Q_i}{p_i}\right)^2 v_R(e_i))]$ is an unbiased estimator for $V(g_R)$. $\hat{v}(g_R)$ may more elegantly be written as

$$v^*(g_R) = (1+C)\Sigma_n\left(\frac{Q_i}{p_i}\right)^2 v_R(e_i) + C\left[\Sigma_n \frac{Q_i}{p_i^2} f_i - g_R^2\right].$$

In the next section we present certain numerical figures to illustrate our live-data-based findings of the application of the above procedures.

4. Simulation-based illustrations of findings

Using the EC data indicated in Section 1 we present here the following details. We first choose any one industry out of the 10 industries and consider the estimation of the total number of earners in the district through this chosen industry. The whole exercise is then repeated for each of the 10 industries.

First, the estimators e_R, g_{11}, g_{12}, g_{22} as obtained in Section 2 are computed. These are based on two-stage RHC–RHC sampling.

Then, we obtain estimators based on adaptive sampling where the initial sample is a two-stage RHC–RHC sample. For each village in the Indian district illustrated here, we define its ‘neighborhood’ to consist of itself and all other villages with a common geographical boundary with it. We take advantage of the figures in Table 2 to form “networks” appropriately. To construct a network for a village, we take specific sets of industries and then check the condition C^* which we take as the existence of at least one earner through any one or more industries of this set. It is interesting to note that the set of industries considered for defining the network, may or may not include the particular industry for which we are estimating the total. Then, we compute estimates based on adaptive sampling and also estimates based on ‘size-constrained’ adaptive sampling, starting from the same initial two-stage RHC–RHC sample.

We can evaluate the method of ‘size-constrained’ adaptive sampling for our example, since we have all necessary census data at hand. For this, we replicate the sampling, both the original and the adaptive, a total of 1000 times. We consider the standardized pivotal, namely, $\tau = \frac{e-\theta}{\sqrt{v}}$, as a standard normal deviate where e is the estimate, θ the parameter and

v the estimate of the variance or the MSE of e. Then, we take $(e - 1.96\sqrt{v}, e + 1.96\sqrt{v})$ as a 95% Confidence Interval (CI) for θ.

We use two measures based on this CI to compare the performance of the alternative estimators. One is the 'ACP', or the average coverage percentage, which is the percent of the replicated samples for which θ is covered by CI. The second measure is 'ARL', or the average relative length, which is the relative length of the CI, averaged over the 1000 replicates.

Table 3 compares some alternative estimators for the total numbers of earners through different industries. The values for ACP and ARL are shown for the traditional estimator, greg estimators and also their adaptive sample versions. Moreover, for a specific replicate, the number (c) of villages chosen in the original sample and the number (d) in the adaptive sample were also computed. But since it was found that d far exceeds c, we had recourse to 'size-restricted' adaptive sampling. We present the values of ACP and ARL, based on SRSWOR's from the networks in various percentages, namely, 8, 10, 15 and 25, rounded upwards to the nearest integer. In addition, we also compute for a specific replicate, the number (a) of villages containing a particular industry in the original sample and the corresponding number (b) in the adaptive sample.

Note that for a specific replicate, one single original sample and its corresponding adaptive sample gives all the estimates for the different industries and so the values of c and d remain constant for all estimates. The value of a depends on the industry for which the estimation is done and b depends on the choice of industry and also the network used. Finally, for estimation for the group of all 10 industries together, $a = c$ and $b = d$.

For ease of understanding, we explain the notations of Table 3 below:

- e_R, g_{11}, g_{12}, g_{22} are as given by equations (2.2), (2.3), (2.4) and (2.5). These are based on two-stage RHC–RHC sampling.
- e_R^*, g_{11}^*, g_{12}^*, g_{22}^* denote modified estimators based on adaptive sampling. The corresponding network used is shown in parentheses. Networks 1, 2, 3 and 4 denote the network formed according to industry sets [1(H), 2(B) & 4(P)], [9(C)], [8(IS) & 9(C)] and [3(HU) & 8(IS)], respectively.
- $\hat{e}_R$, $\hat{g}_{11}$, $\hat{g}_{12}$, $\hat{g}_{22}$ denote estimators based on adaptive sampling by network 4, on which size restriction is applied. The respective percentage(8, 10, 15 or 25) of the adaptive sample-sizes allowed is shown in parentheses.
- a, b, c, d as explained above.

5. Comments and recommendations

For a good estimator, the value of ACP should be high and close to 95.0 and that of ARL should be as small as possible. Also, 'b' should be large compared to 'a' but 'd' should not exceed 'c' by too much. Applying these collective criteria, the 3 greg estimators do not seem to show appreciably better results than the original RHC estimator. This is possibly because the original estimator itself does reasonably well as it is based on an initial sample which has been chosen well enough through suitable stratification and use of appropriate size measures at both stages. The single regressor used here in greg is incapable of yielding

Table 3
Numerical performances of alternative procedures: a few illustrative cases

Industry type	Estimator	ACP	ARL	Estimator	ACP	ARL	Estimator	ACP	ARL
1(H)	e_R	77.9	1.71	$e_R^*(1)$	82.5	1.36	$e_R^*(4)$	85.0	1.28
	g_{11}	77.8	1.70	$g_{11}^*(1)$	83.0	1.37	g_{11}^* (4)	85.8	1.28
	g_{12}	77.9	1.71	g_{12}^* (1)	83.0	1.37	g_{12}^* (4)	85.6	1.28
	g_{22}	78.8	1.90	g_{22}^* (1)	85.4	1.60	g_{22}^* (4)	87.2	1.53
			$a=24$			$b=74$			$b=64$
	$\hat{e}_R(10)$	77.5	1.98	$\hat{e}_R(15)$	74.7	1.94	$\hat{e}_R(25)$	76.5	1.67
	$\hat{g}_{11}(10)$	77.4	1.99	$\hat{g}_{11}(15)$	74.5	1.94	$\hat{g}_{11}(25)$	75.7	1.67
	$\hat{g}_{12}(10)$	77.5	1.98	$\hat{g}_{12}(15)$	74.7	1.94	$\hat{g}_{12}(25)$	76.1	1.67
	$\hat{g}_{22}(10)$	77.6	2.10	$\hat{g}_{22}(15)$	76.2	2.08	$\hat{g}_{22}(25)$	77.1	1.80
			$b=35$			$b=39$			$b=47$
2(B)	e_R	85.5	1.54	$e_R^*(1)$	88.7	1.29	$e_R^*(4)$	87.4	1.24
	g_{11}	85.9	1.54	$g_{11}^*(1)$	89.2	1.29	$g_{11}^*(4)$	87.5	1.24
	g_{12}	85.8	1.55	$g_{12}^*(1)$	88.9	1.29	$g_{12}^*(4)$	87.7	1.24
	g_{22}	86.4	1.46	$g_{22}^*(1)$	88.6	1.20	$g_{22}^*(4)$	88.5	1.16
			$a=34$			$b=115$			$b=95$
	$\hat{e}_R(10)$	89.9	1.40	$\hat{e}_R(15)$	86.5	1.31	$\hat{e}_R(25)$	88.6	1.27
	$\hat{g}_{11}(10)$	89.9	1.40	$\hat{g}_{11}(15)$	87.0	1.31	$\hat{g}_{11}(25)$	88.7	1.27
	$\hat{g}_{12}(10)$	90.0	1.41	$\hat{g}_{12}(15)$	86.9	1.31	$\hat{g}_{12}(25)$	88.6	1.27
	$\hat{g}_{22}(10)$	89.9	1.33	$\hat{g}_{22}(15)$	86.7	1.25	$\hat{g}_{22}(25)$	89.3	1.19
			$b=52$			$b=60$			$b=72$
3(HU)	e_R	88.6	0.81	$e_R^*(3)$	89.2	0.82	$e_R^*(4)$	88.7	0.73
	g_{11}	89.3	0.82	$g_{11}^*(3)$	89.7	0.83	$g_{11}^*(4)$	89.4	0.73
	g_{12}	89.2	0.82	$g_{12}^*(3)$	89.8	0.83	$g_{12}^*(4)$	89.3	0.73
	g_{22}	88.6	0.66	$g_{22}^*(3)$	88.2	0.67	$g_{22}^*(4)$	92.4	0.54
			$a=73$			$b=115$			$b=260$
	$\hat{e}_R(8)$	89.6	0.88	$\hat{e}_R(15)$	90.8	0.81	$\hat{e}_R(25)$	94.3	0.80
	$\hat{g}_{11}(8)$	89.1	0.87	$\hat{g}_{11}(15)$	90.8	0.81	$\hat{g}_{11}(25)$	94.8	0.80
	$\hat{g}_{12}(8)$	89.4	0.88	$\hat{g}_{12}(15)$	90.8	0.81	$\hat{g}_{12}(25)$	95.0	0.80
	$\hat{g}_{22}(8)$	91.0	0.70	$\hat{g}_{22}(15)$	91.1	0.67	$\hat{g}_{22}(25)$	94.1	0.63
			$b=114$			$b=154$			$b=190$
4(P)	e_R	83.8	1.83	$\hat{e}_R(10)$	87.6	1.81	$\hat{e}_R(25)$	88.0	1.53
	g_{11}	83.8	1.83	$\hat{g}_{11}(10)$	87.5	1.81	$\hat{g}_{11}(25)$	88.4	1.53
	g_{12}	83.9	1.83	$\hat{g}_{12}(10)$	87.4	1.81	$\hat{g}_{12}(25)$	87.8	1.53
	g_{22}	84.3	1.73	$\hat{g}_{22}(10)$	87.7	1.70	$\hat{g}_{22}(25)$	89.6	1.49
			$a=16$			$b=22$			$b=30$
5(S)	e_R	49.3	3.13	$e_R^*(2)$	69.2	3.14	$e_R^*(4)$	69.8	2.61
	g_{11}	49.8	3.14	$g_{11}^*(2)$	68.8	3.18	$g_{11}^*(4)$	69.7	2.62
	g_{12}	49.6	3.14	$g_{12}^*(2)$	69.0	3.17	$g_{12}^*(4)$	69.8	2.61
	g_{22}	47.9	3.02	$g_{22}^*(2)$	68.5	3.07	$g_{22}^*(4)$	71.8	2.48
			$a=4$			$b=5$			$b=9$
	$\hat{e}_R(8)$	58.1	3.41	$\hat{e}_R(15)$	60.5	3.34	$\hat{e}_R(25)$	63.3	3.21
	$\hat{g}_{11}(8)$	57.7	3.44	$\hat{g}_{11}(15)$	60.8	3.42	$\hat{g}_{11}(25)$	63.9	3.21
	$\hat{g}_{12}(8)$	57.8	3.42	$\hat{g}_{12}(15)$	60.5	3.39	$\hat{g}_{12}(25)$	63.3	3.21
	$\hat{g}_{22}(8)$	59.4	3.35	$\hat{g}_{22}(15)$	60.7	3.30	$\hat{g}_{22}(25)$	63.3	3.10
			$b=5$			$b=6$			$b=7$

Table 3 (*continued*)

Industry type	Estimator	ACP	ARL	Estimator	ACP	ARL	Estimator	ACP	ARL
$6(SB)$	e_R	65.9	2.62	$e^*_R(4)$	83.7	1.96	$\hat{e}_R(8)$	86.8	2.37
	g_{11}	65.9	2.60	$g^*_{11}(4)$	83.5	1.97	$\hat{g}_{11}(8)$	86.9	2.38
	g_{12}	65.8	2.61	$g^*_{12}(4)$	83.5	1.97	$\hat{g}_{12}(8)$	86.7	2.38
	g_{22}	71.3	2.90	$g^*_{22}(4)$	87.2	2.38	$\hat{g}_{22}(8)$	89.6	2.71
			$a=6$			$b=16$			$b=9$
$7(T)$	e_R	87.0	1.53	$e^*_R(4)$	89.5	1.02	$\hat{e}_R(25)$	89.3	1.30
	g_{11}	86.9	1.53	$g^*_{11}(4)$	89.6	1.02	$\hat{g}_{11}(25)$	89.4	1.31
	g_{12}	86.8	1.53	$g^*_{12}(4)$	89.7	1.03	$\hat{g}_{12}(25)$	89.2	1.31
	g_{22}	86.8	1.68	$g^*_{22}(4)$	91.7	1.22	$\hat{g}_{22}(25)$	90.0	1.54
			$a=21$			$b=54$			$b=44$
$8(IS)$	e_R	92.1	0.86	$e^*_R(2)$	92.7	0.84	$e^*_R(4)$	92.8	0.77
	g_{11}	92.5	0.86	$g^*_{11}(2)$	93.5	0.84	$g^*_{11}(4)$	93.1	0.77
	g_{12}	92.3	0.87	$g^*_{12}(2)$	93.2	0.85	$g^*_{12}(4)$	93.0	0.77
	g_{22}	93.6	0.75	$g^*_{22}(2)$	92.9	0.73	$g^*_{22}(4)$	93.2	0.64
			$a=53$			$b=77$			$b=187$
	$\hat{e}_R(10)$	93.5	0.92	$\hat{e}_R(15)$	92.7	0.89	$\hat{e}_R(25)$	91.3	0.89
	$\hat{g}_{11}(10)$	93.2	0.92	$\hat{g}_{11}(15)$	93.1	0.90	$\hat{g}_{11}(25)$	91.0	0.89
	$\hat{g}_{12}(10)$	93.3	0.92	$\hat{g}_{12}(15)$	93.2	0.90	$\hat{g}_{12}(25)$	91.2	0.90
	$\hat{g}_{22}(10)$	93.8	0.77	$\hat{g}_{22}(15)$	93.0	0.80	$\hat{g}_{22}(25)$	92.4	0.79
			$b=94$			$b=108$			$b=133$
$9(C)$	e_R	91.7	1.04	$e^*_R(4)$	91.8	0.86	$\hat{e}_R(25)$	90.0	1.04
	g_{11}	91.6	1.04	$g^*_{11}(4)$	92.0	0.86	$\hat{g}_{11}(25)$	89.8	1.04
	g_{12}	91.9	1.05	$g^*_{12}(4)$	92.1	0.86	$\hat{g}_{12}(25)$	90.1	1.04
	g_{22}	91.4	0.93	$g^*_{22}(4)$	90.3	0.72	$\hat{g}_{22}(25)$	91.0	0.91
			$a=41$			$b=113$			$b=82$
$10(PC)$	e_R	85.9	1.85	$e^*_R(1)$	88.6	1.60	$e^*_R(4)$	87.0	1.53
	g_{11}	86.0	1.85	$g^*_{11}(1)$	88.4	1.61	$g^*_{11}(4)$	87.8	1.54
	g_{12}	85.9	1.85	$g^*_{12}(1)$	88.1	1.61	$g^*_{12}(4)$	87.7	1.54
	g_{22}	86.3	1.83	$g^*_{22}(1)$	87.8	1.61	$g^*_{22}(4)$	87.9	1.52
			$a=9$			$b=20$			$b=21$
	$\hat{e}_R(10)$	90.2	1.97	$\hat{e}_R(15)$	83.8	1.85	$\hat{e}_R(25)$	87.3	1.87
	$\hat{g}_{11}(10)$	90.5	1.99	$\hat{g}_{11}(15)$	84.3	1.86	$\hat{g}_{11}(25)$	87.8	1.87
	$\hat{g}_{12}(10)$	90.3	1.98	$\hat{g}_{12}(15)$	84.2	1.86	$\hat{g}_{12}(25)$	87.3	1.87
	$\hat{g}_{22}(10)$	90.2	1.99	$\hat{g}_{22}(15)$	84.2	1.84	$\hat{g}_{22}(25)$	87.8	1.84
			$b=13$			$b=15$			$b=17$
All10	e_R	91.1	0.65	$e^*_R(1)$	89.5	0.65	$e^*_R(2)$	91.2	0.68
	g_{11}	91.1	0.65	$g^*_{11}(1)$	90.3	0.66	$g^*_{11}(2)$	91.8	0.68
	g_{12}	91.2	0.66	$g^*_{12}(1)$	90.6	0.66	$g^*_{12}(2)$	91.7	0.69
	g_{22}	93.5	0.78	$g^*_{22}(1)$	91.8	0.78	$g^*_{22}(2)$	91.6	0.81
			$a=112$ $c=112$			$b=241$ $d=241$			$b=164$ $d=164$
g^*_{11}	$e^*_R(3)$	90.2	0.58	$e^*_R(4)$	91.6	0.46	$\hat{e}_R(8)$	91.4	0.85
	$g^*_{11}(3)$	91.5	0.58	$g^*_{11}(4)$	91.8	0.46	$\hat{g}_{11}(8)$	91.1	0.84
	$g^*_{12}(3)$	91.3	0.59	$g^*_{12}(4)$	91.7	0.46	$\hat{g}_{12}(8)$	91.1	0.84
	$g^*_{22}(3)$	92.4	0.73	$g^*_{22}(4)$	93.8	0.64	$\hat{g}_{22}(8)$	92.4	0.96
			$b=273$ $d=273$			$b=357$ $d=357$			$b=167$ $d=167$

Table 3 (*continued*)

Industry type	Estimator	ACP	ARL	Estimator	ACP	ARL	Estimator	ACP	ARL
	$\hat{e}_R(10)$	94.9	0.87	$\hat{e}_R(15)$	91.8	0.74	$\hat{e}_R(25)$	92.1	0.73
	$\hat{g}_{11}(10)$	94.9	0.88	$\hat{g}_{11}(15)$	92.7	0.74	$\hat{g}_{11}(25)$	92.3	0.73
	$\hat{g}_{12}(10)$	95.1	0.88	$\hat{g}_{12}(15)$	92.5	0.75	$\hat{g}_{12}(25)$	92.7	0.73
	$\hat{g}_{22}(10)$	95.9	1.01	$\hat{g}_{22}(15)$	92.9	0.88	$\hat{g}_{22}(25)$	93.7	0.88
			$b = 185$			$b = 220$			$b = 263$
			$d = 185$			$d = 220$			$d = 263$

further accuracy in estimation by the regression technique. More improvement could have been achieved if the available auxiliary variable was better associated with the variable of interest. However, in large scale surveys, a search for such a highly associated variable is not always practicable and one has to make do with what is readily available.

For the highly localized industries, namely, Handloom(1), Silk(5) and Stone-breaking(6), only 'adaptive' sampling achieves significant improvement. For the rest, original sampling is good enough. Using husking and iron-smithy (3 and 8) as the condition C^* for networking seems to be the most suitable in enhancing the precision in estimation. But this entails increasing the overall initial sample-size, say from 112 to 357 in the adaptive sample in one replicate. So, constraining the sample-size is important.

Our suggested procedures happen to bring down the sample-size from 357 to 167, 185, 220 and 263, respectively, with 8%, 10%, 15% and 25% sub-sampling, with upward rounding to integers. It is sometimes found that an estimator based on a size-restricted adaptive sample has a lower ACP for a larger sample-size than for a smaller sample-size. This is because the former estimator gives a smaller estimate of MSE and consequently a narrower CI which fails to cover the true value. It may thus be noted that the former estimator may sometimes have smaller ACP but it will perform better in terms of ARL having a smaller ARL.

So, our final recommendation is that first a good initial sampling scheme has to be employed utilizing available auxiliary data. This may achieve desirable levels of efficiency for estimation of many of the characteristics. If it fails with respect to a few variables, as ascertained on computing the estimators for the coefficients of variation, then greg estimators may be tried as alternatives. If their estimated coefficients of variation also turn out to be large, then adaptive sampling may be tried. Since, fieldworks for the adaptive sampling have to be implemented prior to the data analysis, one should undertake it for those variables for which one anticipates possible drops in efficiency level prior to the survey. However, if the resulting adaptive sample-size goes on spiraling up, a decision for sub-sampling has to be implemented at the fieldwork stage. Whenever one considers going for adaptive sampling with or without size-constraining, prior data as in Tables 1 and 2 must be exploited for a proper guidance.

Acknowledgements

The authors would like to thank two referees for their detailed and helpful comments.

References

Cassel, C.M., Särndal, C.E., Wretman, J.H., 1976. Some results on generalized difference estimation and generalized regression estimation for finite population. Biometrika 63, 615–620.

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

Chaudhuri, A., Adkikary, A.K., Dihidar, S., 2000. Mean square error estimation in multi-stage sampling. Metrika 52, 115–131.

Chaudhuri, A., Bose, M., Ghosh, J.K., 2004. An application of adaptive sampling to estimate highly localized population segments. J. Statist. Plann. Inference 121, 175–189.

Raj, D., 1968. Sampling Theory, Mc-Graw Hill, New York.

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. Ser. B 24, 482–491.

Salehi, M.M., Seber, G.A.F., 1997. Adaptive cluster sampling with networks selected without replacement. Biometrika 84 (1), 209–219.

Salehi, M.M., Seber, G.A.F., 2002. Unbiased estimators for restricted adaptive cluster sampling. Austrl. & New Zealand J. Statist. 44 (1), 63–74.

Thompson, S.K., 1990. Adaptive cluster sampling. J. Amer. Statist. Assoc. 85, 1050–1059.

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

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