Statistics and Applications
Volume 10, Nos. 1&2, 2012 (New Series), pp. 13-25

Estimation of Domain Total for Unknown Domain Size in the Presence of Nonresponse

Kaustav Aditya, U.C. Sud and Hukum Chandra
Indian Agricultural Statistics Research Institute, New Delhi, India

Abstract

This article describes the estimation of domain total in the presence of nonresponse when the domain size unknown and the sampling design is two-stage. Further, the response mechanism is assumed to be deterministic. An estimator based on sub-sampling of non-respondents, collecting data on the sub-sample through specialized efforts, is proposed. Expression for the variance of the estimator is also developed. A suitable cost function is considered for obtaining the optimum sample sizes. Empirical studies are carried out to examine the percentage reduction in the expected cost of proposed estimator.

Keywords: Cost function; Nonresponse; Sub-sample; Two-stage sampling

1 Introduction

For large or medium scale surveys we are often faced with the scenario that the sampling frame of ultimate stage units is not available and the cost of construction of the frame is very high. Sometimes the population elements are scattered over a wide area resulting in a widely scattered sample. Therefore, not only the cost of enumeration of units in such a sample may be very high, the supervision of field work may also be very difficult. For such situations, two-stage or multi-stage sampling designs are very effective.

It is also the case that, in many human surveys, information is not obtained from all the units in surveys. The problem of nonresponse persist even after call backs. The estimates obtained from incomplete data may be biased particularly when the respondents differ from the non-respondents. Hansen and Hurwitz (1946) proposed a technique for adjusting for nonresponse to address the problem of bias. The technique consists of selecting a sub-sample of non-respondents. Through specialized efforts data are collected from the non-respondents so as to obtain an estimate of nonresponding units in the population. Oh and Scheuren (1983) attempted to compensate for nonresponse by weighing adjustment. Kalton and Karsprzyk (1986) tried the imputation technique. Tripathi and Khare (1997) extended the sub-sampling of non-respondents approach to multivariate case. Okafor and Lee (2000) extended the approach to double sampling for ratio and regression estimation. Okafor (2001, 2005) further extended the approach in the context of element sampling and two-phase sampling respectively on two successive occasions.

It may be mentioned that the weighting and imputation procedures aim at elimination of bias caused by nonresponse. However, these procedures are based on certain assumptions on the response mechanism. When these assumptions do not hold good the resulting estimate may be seriously biased. Further, when the nonresponse is confounded i.e. the response probability is dependent on the survey character; it becomes difficult to eliminate the bias entirely. Rancourt, Lee and Sarndal (1994) provided a partial correction for the situation. Hansen and Hurwitz's sub-sampling approach although costly, is free from any assumptions. When the bias caused by nonresponse is serious this technique is very effective i.e. one does not have to go for 100 percent response, which can be very expensive.

Many a times, besides the overall estimates, the estimates for different subgroups/domains of population are also required (Sarndal *et al.*, 1992). In the context of estimation of the domain parameters, Agrawal and Midha (2007) proposed a two phase sampling design when the size of the domain was not known. Sud *et al.* (2010) considered the problem of estimation of finite population mean of a domain in the presence of nonresponse under a deterministic response mechanism. Chhikara and Sud (2009) used the sub-sampling of non-respondents approach for estimation of population and domain totals in the context of item nonresponse. However, the results in both the above cases were limited to uni-stage sampling design. Again, Sud *et al.* (2012) considered the problem of estimation of finite population mean in the presence of nonresponse under two stage sampling design when the response mechanism was assumed to be deterministic.

In what follows, an estimator of domain total using two-stage two phase sampling designs are developed in Section 2 based on the technique of sub-sampling of the non-respondents when the domain size is unknown. In this case, the response mechanism is assumed to be deterministic. Also given in this section is expression for variance of the estimator. Besides, optimum values of sample sizes are obtained by minimizing the expected cost for a fixed variance. The results are empirically illustrated in Section 3.

2 Theoretical Developments

For estimation of domain parameters, we assume that the domain size is known. However, this may not be the case always. In this section we develop the necessary theory of estimation of domain total when domain size is unknown. Let the finite population U under consideration consists of N known primary stage units (psus) labeled 1 through N. Let the i-th psu comprise M second stage units (ssus). Let us consider a population $U = (1, \ldots, k, \ldots, N)$ of size N partitioned into D sub-sets $U_1, \ldots, U_d, \ldots U_D$ (hereafter we refer them as domains) and let N_d (which is assumed large) be the size of $U_d\,(d = 1, \ldots, D)$ such that $U = \bigcup_{d=1}^{D} U_d$ and $N = \sum_{d=1}^{D} N_d$. Here, N_d is assumed to be unknown. We consider two phase sampling for estimation of domain total. The domain size i.e. N_d is estimated at the first phase and at the second phase we select the sample for estimation of domain total. In this case let, n' psus are selected from N psus by simple random sampling without replacement (srswor) sampling design at the first phase and later at the second phase a sample of size n is selected from n' by srswor, n'_d and n_d out of n' and n psus fall in the d-th domain respectively.

When the domain sizes are small, n_d may turn out to be very small or it may be equal to '0' in some cases. In such cases small area estimation techniques are needed for reliable estimation at the domain level. However, we do not consider this case here. Let M_d be the size of the units in each psu belonging to the d-th domain and from each selected psu m_d ssus are selected by srswor and letters/mails containing questionnaires are sent to each unit in the sample. With the random sample of observations, the statistician's task is to make the best possible estimate for the domain. Let y_{dkj} be the value of study character pertaining to j-th ssu in the k-th psu in d-th domain, k=1,2,…,N_d, j=1,2,…, M_d, d=1, 2,..., D. Our objective here is to estimate the domain total $Y_d = \sum_{k=1}^{N_d} \sum_{j=1}^{M_d} y_{dkj}$.

For the estimator of the domain total when domain size is assumed unknown, let, n' psus are selected from N psus by srswor design at the first phase and later at the second phase a sample of size n is selected from n' by srswor, n'_d and n_d out of n' and n psus fall in the d-th domain respectively. At the second phase, within each selected psu, m_d ssus are also selected from M_d ssus by srswor.

Let out of a sample m_d ssus selected from M_d ssus, m_{dk_1} ssus respond while m_{dk_2} ssus do not respond, $m_{dk_1}+m_{dk_2}=m_d$. From the m_{dk_2} nonresponding ssus a sub-sample of h_{dk_2} ssus is selected by srswor, $m_{dk_2}=h_{dk_2}f_{dk_2}$, k=1, 2,..., n_d. Let $\overline{y}_{m_{dk_1}}$ denote the mean of the sample from the response class in the d-th domain while $\overline{y}_{h_{dk_2}}$ denote the mean of the sub-sample for the nonresponse class, where $\overline{y}_{m_{dk_1}}=\frac{1}{m_{dk_1}}\sum_{j=1}^{m_{dk_1}} y_{dkj}$ and $\overline{y}_{h_{dk_2}}=\frac{1}{h_{dk_2}}\sum_{k=1}^{h_{dk_2}} y_{dkj}$.

An unbiased estimator of the domain total is given as,

$$\hat{Y}'_{1d}=\frac{NM_d n'_d}{n'n_d}\sum_{k=1}^{n_d}\frac{1}{m_d}(m_{dk_1}\overline{y}_{m_{dk_1}}+m_{dk_2}\overline{y}_{h_{dk_2}}) \tag{1}$$

with variance,

$$V(\hat{Y}'_{1d})=\frac{N(N-n')M_d^2}{n'(N-1)}N_dQ_d\overline{Y}_d^{\,2}+\frac{N(N-n')(N_d-1)}{n'(N-1)}M_d^2S_{bd}^2+\frac{NN_dM_d^2}{n'}\left(\frac{(n'-n)}{n}\right)S_{bd}^2$$
$$+\frac{N}{n}\sum_{k=1}^{N_d}(\frac{1}{m_d}-\frac{1}{M_d})M_d^2S_{dk}^2+\frac{N}{n}\sum_{k=1}^{N_d}\frac{M_{dk_2}M_d}{m_d}(f_{dk_2}-1)S_{M_{dk_2}}^2 \tag{2}$$

where,

$$S_{bd}^2=\frac{1}{\left(N_d-1\right)}\sum_{k=1}^{N_d}(\overline{Y}_{dk}-\overline{Y}_d)^2,\ \overline{Y}_{dk}=\frac{1}{M_d}\sum_{j=1}^{M_d}Y_{dk_j}\ \text{and}\ \overline{Y}_d=\frac{1}{N_d}\sum_{k=1}^{N_d}\overline{Y}_{dk}.$$
$$S_{dk}^2=\frac{1}{(M_d-1)}\sum_{j=1}^{M_d}(Y_{dk_j}-\overline{Y}_{dk})^2,P_d=\frac{N_d}{N},Q_d=1-P_d.$$
$$S_{M_{dk_2}}^2=\frac{1}{(M_{dk_2}-1)}\sum_{j=1}^{M_{dk_2}}(Y_{dk_j}-\overline{Y}_{M_{dk_2}})^2,\overline{Y}_{M_{dk_2}}=\frac{1}{M_{dk_2}}\sum_{j=1}^{M_{dk_2}}Y_{dk_j}.$$

The proof is given as below,

$$E\left(\hat{Y}'_{1d}\right) = E_1E_2E_3E_4E_5\left[E_6\left\{\frac{NM_d n'_d}{n'n_d}\sum_{k=1}^{n_d}\frac{1}{m_d}(m_{dk_1}\bar{y}_{m_{dk_1}} + m_{dk_2}\bar{y}_{h_{dk_2}})\right\}\right]$$

$$= E_1E_2E_3E_4\left[E_5\left\{\frac{NM_d n'_d}{n'n_d}\sum_{k=1}^{n_d}\frac{1}{m_d}(m_{dk_1}\bar{y}_{m_{dk_1}} + m_{dk_2}\bar{y}_{m_{dk_2}})\right\}\right]$$

$$= E_1E_2E_3\left[E_4\left\{\frac{M_d N n'_d}{n'n_d}\sum_{k=1}^{n_d}\bar{Y}_{dk}\right\}\right]$$

$$= E_1E_2E_3\left[E_4\left\{\frac{M_d N n'_d}{n'}\bar{y}_{n_d}\right\}\right]$$

$$= E_1E_2\left[E_3\left\{\frac{M_d N n'_d}{n'}\bar{y}_{n'_d}\right\}\right]$$

$$= E_1\left[E_2\left\{\frac{M_d N n'_d}{n'}\bar{y}_{n'_d}\right\}\right]$$

$$= E_1\left\{\frac{M_d N n'_d}{n'}\bar{Y}_d\right\}$$

$$= N_d M_d \bar{Y}_d = Y_d .$$

Here $\bar{Y}_{dk} = \dfrac{1}{M_d}\sum_{j=1}^{M_d} y_{dkj}$ and $\bar{Y}_d = \dfrac{1}{N_d}\sum_{k=1}^{N_d}\dfrac{1}{M_d}\sum_{j=1}^{M_d} y_{dkj}$. This indicates that $\hat{Y}'_{1d}$ is a unbiased estimator of the domain total. Further, E_6 represents conditional expectations of all possible samples of size h_{dk_2} drawn from m_{dk_2} , E_5 is the conditional expectation of all possible samples of size m_d drawn from M_d, E_4 is the conditional expectation of all possible samples of size n_d drawn from n'_d keeping n_d fixed, E_3 is the conditional expectation arising out of randomness of n_d, E_2 is the conditional expectation of all possible samples of size n'_d drawn from N_d keeping n'_d fixed while E_1 refers to expectation arising out of randomness of n'_d. The variance of the above estimator is given as,

$$V\left(\hat{Y}'_{1d}\right)=V_1E_2E_3E_4E_5E_6\left(\hat{Y}'_{1d}\right)+E_1V_2E_3E_4E_5E_6\left(\hat{Y}'_{1d}\right)+E_1E_2V_3E_4E_5E_6\left(\hat{Y}'_{1d}\right)+E_1E_2E_3V_4E_5E_6\left(\hat{Y}'_{1d}\right)$$
$$+E_1E_2E_3E_4V_5E_6\left(\hat{Y}'_{1d}\right)+E_1E_2E_3E_4E_5V_6\left(\hat{Y}'_{1d}\right)$$

$$V_1E_2E_3E_4E_5E_6\left(\hat{Y}'_{1d}\right)=\frac{N(N-n')M_d^2}{n'(N-1)}N_dQ_d\bar{Y}_d^{\,2},$$
$$E_1V_2E_3E_4E_5E_6\left(\hat{Y}'_{1d}\right)=\frac{N(N-n')(N_d-1)}{n'(N-1)}M_d^2S_{bd}^2,$$
$$E_1E_2V_3E_4E_5E_6\left(\hat{Y}'_{1d}\right)=0,$$
$$E_1E_2E_3V_4E_5E_6\left(\hat{Y}'_{1d}\right)=\frac{NN_dM_d^2}{n'}\left(\frac{(n'-n)}{n}\right)S_{bd}^2,$$
$$E_1E_2E_3E_4V_5E_6\left(\hat{Y}'_{1d}\right)=\frac{N}{n}\sum_{k=1}^{N_d}\left(\frac{1}{m_d}-\frac{1}{M_d}\right)M_d^2S_{dk}^2,$$
$$E_1E_2E_3E_4E_5V_6\left(\hat{Y}'_{1d}\right)=\frac{N}{n}\sum_{k=1}^{N_d}\frac{M_{dk_2}M_d}{m_d}(f_{dk_2}-1)S^2_{M_{dk_2}}.$$

Here, V_1,V_2,V_3,V_4,V_5,V_6 are defined similarly as E_1,E_2,E_3,E_4,E_5,E_6. Now adding all the above variance terms we get the required expression in equation (2). We determine the optimum values of n', n, m_d and f_{dk_2} by minimizing the expected cost for a fixed variance. To achieve this consider the following cost function

$$C=C_{1d}n'_d+C_{2d}n_d+C_{3d}\sum_{k=1}^{n_d}m_{dk_1}+C_{4d}\sum_{k=1}^{n_d}h_{dk_2},$$

where,
C: Total cost
C_{1d}: per unit travel and miscellaneous cost at the first phase in the d-th domain.
C_{2d}: Per unit travel and miscellaneous cost at the second phase in the d-th domain.
C_{3d}: Cost per unit for collecting the information on the study character in the first attempt in the d-th domain.
C_{4d}: Cost per unit for collecting the information by expensive method after the first attempt failed in the d-th domain.

The cost function considered above is suitable for situations prevailing in mail surveys. In these surveys the first attempt to collect information from the respondents is made through e-mail/postal mail. Many of the respondents may not send the required information through mails. To collect information, a sub-sample of non-respondents may be collected for data collection by specialized effort, say, personal interview.

The expected cost in this case is,

$$E(C)=\frac{N_d}{N}[C_{1d}n'+C_{2d}n+C_{3d}\sum_{k=1}^{N_d}\frac{M_{dk_1}m_d}{M_d}n+C_{4d}\sum_{k=1}^{N_d}\frac{M_{dk_2}m_d}{M_d f_{dk_2}}n]$$

Consider the function $\phi = E(C) + \lambda\{V(\hat{Y}'_{1d}) - V_0\}$. Here, λ is the Lagrangian multiplier.

Also, V_0 can be determined by fixing the coefficient of variation, say equal to 5%. To get closed form expression of the optimum values we assume that $m_{dk_2} = h_{dk_2} f_{2d}$, k=1, 2,..., n_d in place of $m_{dk_2} = h_{dk_2} f_{dk_2}$, k=1, 2,..., n_d.

Differentiation with respect to n, m_d, n', λ and f_{2d}, equating the resultant derivatives equal to '0' we get, the optimum values as,

$$n_{opt}=\frac{K_{19}}{K_{20}}\ , m_{dopt}=\frac{-b_{11}\pm\sqrt{b_{11}{}^2+4a_{11}c_{11}}}{2a_{11}},\quad f_{2dopt}=\pm\sqrt{\frac{b_3}{b_2}}\ \text{and}$$

$$n'=\pm\sqrt{\frac{C_{4d}\sum_{k=1}^{N_d}\frac{M_{dk_2}}{f^2{}_{2d}}n^2m_d^2\left(\frac{NN_dQ_d\bar{Y}_d{}^2}{(N-1)}+\left\{\frac{N(N_d-1)}{(N-1)}-N_d\right\}S_{bd}^2\right)}{C_{1d}\sum_{k=1}^{N_d}M_{dk_2}S_{M_{dk_2}}^2}}$$

We consider only positive values, hence,

$$m_{dopt}=\frac{-b_{11}+\sqrt{b_{11}{}^2+4a_{11}c_{11}}}{2a_{11}},\quad f_{2dopt}=\sqrt{\frac{b_3}{b_2}}\ \text{and}$$

$$n'=\sqrt{\frac{C_{4d}\sum_{k=1}^{N_d}\frac{M_{dk_2}}{f^2{}_{2d}}n^2m_d^2\left(\frac{NN_dQ_d\bar{Y}_d{}^2}{(N-1)}+\left\{\frac{N(N_d-1)}{(N-1)}-N_d\right\}S_{bd}^2\right)}{C_{1d}\sum_{k=1}^{N_d}M_{dk_2}S_{M_{dk_2}}^2}}$$

where,

$$b_2 = C_{3d}\sum_{k=1}^{N_d}\frac{M_{dk_1}}{m_d}\sum_{k=1}^{N_d}M_{dk_2}M_dS^2_{M_{dk_2}}, b_3 = C_{4d}\sum_{k=1}^{N_d}\frac{M_{dk_2}}{M_d}\left(\sum_{k=1}^{N_d}M_d^2S_{dk}^2 - \sum_{k=1}^{N_d}M_{dk_2}M_dS^2_{M_{dk_2}}\right)$$

$$a_{11} = C_{4d}\sum_{k=1}^{N_d}\frac{M_{dk_2}}{M_d f^2{}_{2d}}\left[M_d^2N_dS_{bd}^2 - \sum_{k=1}^{N_d}M_dS_{dk}^2\right]$$

$$b_{11} = [C_{4d}\sum_{k=1}^{N_d}\frac{M_{dk_2}}{M_d f^2{}_{2d}}\left(\sum_{k=1}^{N_d}M_d^2S_{dk}^2 - \sum_{k=1}^{N_d}M_{dk_2}M_dS^2_{M_{dk_2}}\right) - C_{3d}\sum_{k=1}^{N_d}\frac{M_{dk_1}}{M_d}\sum_{k=1}^{N_d}M_{dk_2}M_dS^2_{M_{dk_2}}]$$

$$c_{11} = \left(C_{2d}\sum_{k=1}^{N_d}M_{dk_2}M_dS^2_{M_{dk_2}}\right),$$

$$K_{19} = NM_d^2N_dS_{bd}^2 + NM_d^2\sum_{k=1}^{N_d}(\frac{1}{m_d} - \frac{1}{M_d})M_d^2S_{dk}^2 + N\sum_{k=1}^{N_d}\frac{M_{dk_2}M_d}{m_d}(f_{dk_2} - 1)S^2_{M_{dk_2}},$$

$$K_{20} = V_0 + \frac{NM_d^2N_d}{n'}S_{bd}^2 - \frac{N(N-n')M_d^2}{n'(N-1)}N_dQ_d\bar{Y}_d{}^2 - \frac{N(N-n')(N_d-1)}{n'(N-1)}M_d^2S_{bd}^2$$

and $V_0 = 0.0025 \times Y_d^2$.

We consider a control situation. Here we assume that N_d is unknown. We make specialized efforts to collect data so that there is no nonresponse. An unbiased estimator of domain total is given as,

$$\hat{Y}'_{2d} = \frac{NM_dn'_d}{n_dn'}\left\{\sum_{k=1}^{n_d}\bar{y}_{dk}\right\}, \tag{3}$$

where, n' psus are selected from N psus by srswor design at the first phase and later at the second phase a sample of size n is selected from n' by srswor, n'_d and n_d out of n' and n psus fall in the d-th domain respectively. Within each selected psu, m_d ssus are also selected from M_d ssus by srswor. Data are collected through specialized efforts i.e. there is no nonresponse. The variance of $\hat{Y}'_{2d}$ is given by,

$$V(\hat{Y}'_{2d}) = \frac{N(N-n')M_d^2}{n'(N-1)}N_dQ_d\bar{Y}_d{}^2 + \frac{N(N-n')(N_d-1)}{n'(N-1)}M_d^2S_{bd}^2 + \frac{NN_dM_d^2}{n'}\left(\frac{(n'-n)}{n}\right)S_{bd}^2 + \frac{N}{n}\left[\sum_{k=1}^{N_d}(\frac{1}{m_d} - \frac{1}{M_d})M_d^2S_{dk}^2\right], \tag{4}$$

and various terms in the above expressions are defined earlier. Prof is given as below,

$$E\left(\hat{Y}'_{2d}\right) = E_1E_2E_3E_4\left[E_5\left\{\frac{NM_d n'_d}{n'n_d}\sum_{k=1}^{n_d}\overline{y}_{dk}\right\}\right]$$
$$= E_1E_2E_3\left[E_4\left\{\frac{M_d Nn'_d}{n'n_d}\sum_{k=1}^{n_d}\overline{Y}_k\right\}\right]$$
$$= E_1E_2E_3\left[E_4\left\{\frac{M_d Nn'_d}{n'}\overline{y}_{n_d}\right\}\right]$$
$$= E_1E_2\left[E_3\left\{\frac{M_d Nn'_d}{n'}\overline{y}_{n'_d}\right\}\right]$$
$$= E_1\left[E_2\left\{\frac{M_d Nn'_d}{n'}\overline{y}_{n'_d}\right\}\right]$$
$$= E_1\left\{\frac{M_d Nn'_d}{n'}\overline{Y}_d\right\}$$
$$= N_d M_d \overline{Y}_d$$
$$= Y_d.$$

Hence, it can be seen that $\hat{Y}'_{2d}$ is an unbiased estimator of domain total and E_5 is the conditional expectation that of all possible samples of size m_d drawn from M_d, E_4 is the conditional expectation of all possible samples of size n_d is drawn from n'_d keeping n_d fixed, E_3 is the conditional expectation arising out of randomness of n_d, E_2 is the conditional expectation of all possible samples of size n'_d drawn from N_d keeping n'_d fixed while E_1 refers to expectation arising out of randomness of n'_d.

The variance of the above estimator is given as,

$$V\left(\hat{Y}'_{2d}\right) = V_1E_2E_3E_4E_5\left(\hat{Y}'_{2d}\right) + E_1V_2E_3E_4E_5\left(\hat{Y}'_{2d}\right) + E_1E_2V_3E_4E_5\left(\hat{Y}'_{2d}\right) + E_1E_2E_3V_4E_5\left(\hat{Y}'_{2d}\right)$$
$$+ E_1E_2E_3E_4V_5\left(\hat{Y}'_{2d}\right)$$

$$V_1E_2E_3E_4E_5\left(\hat{Y}'_{2d}\right) = \frac{N(N-n')M_d^2}{n'(N-1)}N_d Q_d \overline{Y}_d^{\,2},$$

$$E_1V_2E_3E_4E_5\left(\hat{Y}'_{2d}\right) = \frac{N(N-n')(N_d-1)}{n'(N-1)}M_d^2 S_{bd}^2,$$

$$E_1E_2V_3E_4E_5E_6\left(\hat{Y}'_{2d}\right)=0,$$

$$E_1E_2E_3V_4E_5\left(\hat{Y}'_{2d}\right)=\frac{NN_dM_d^2}{n'}\left(\frac{(n'-n)}{n}\right)S_{bd}^2,$$

$$E_1E_2E_3E_4V_5\left(\hat{Y}'_{2d}\right)=\frac{N}{n}\sum_{k=1}^{N_d}(\frac{1}{m_d}-\frac{1}{M_d})M_d^2S_{dk}^2$$

Where, V_1,V_2,V_3,V_4,V_5 are defined similarly as E_1,E_2,E_3,E_4,E_5.Now adding all the terms above we get the variance expression.

We determine the optimum values of n, m_d, n' by minimizing the expected cost for a fixed variance. To achieve this consider the following cost function

$$C = C_{1d}n'_d + C_{2d}n_d + C_{4d}n_dm_d$$

where the various costs appearing in the cost function are same as defined earlier.
The expected cost in this case is,

$$E(C)=\frac{N_d}{N}[C_{1d}n'+C_{2d}n+C_{4d}m_dn]$$

The optimum values are,

$$n_{opt}=\frac{NN_dS_{bd}^2+N\sum_{k=1}^{N_d}(\frac{1}{m_d}-\frac{1}{M_d})S_{dk}^2}{V_0+\frac{NM_d^2N_d}{n'}S_{bd}^2-\frac{N(N-n')M_d^2}{n'(N-1)}N_dQ_d\bar{Y}_d^2-\frac{N(N-n')(N_d-1)}{n'(N-1)}M_d^2S_{bd}^2},$$

$$m_{dopt}=\sqrt{\frac{C_{2d}\sum_{k=1}^{N_d}S_{dk}^2}{C_{4d}(N_dS_{bd}^2-\sum_{k=1}^{N_d}\frac{S_{dk}^2}{M_d})}},\quad \text{and}$$

$$n'=\sqrt{\frac{C_{4d}n^2m_d^2\left(\frac{NN_dQ_d\bar{Y}_d^2}{(N-1)}+\left\{\frac{N(N_d-1)}{(N-1)}-N_d\right\}S_{bd}^2\right)}{C_{1d}\sum_{k=1}^{N_d}M_{dk_2}S_{M_{dk_2}}^2}}$$

3 Empirical Studies

For empirical illustration data pertaining to MU 284 population given in Sarndal *et al.* (1992) was used. The variable of interest here was P85 which was human Population (in thousands) of 284 municipalities of Sweden in 1985. Using this data, a population U of size N = 27 psus was generated by combining the adjacent 10 units and allocating them to the respective psus. From the N psus a sample of n'=24 (i.e., first phase sample) psus each of size M_d=10 was drawn using srswor and then from the first phase sample of n' psus a second phase sample of size n=21 psus was selected using srswor. Here the population U was divided into three domains $U_d (d = 1,...,3)$ each of equal size $N_d = 9$ and we considered M_{dk_1}=5 and M_{dk_2}=5. Various combinations of C_{1d}, C_{2d}, C_{3d} and C_{4d} were considered. The percentage reduction in expected cost of $\hat{Y}'_{1d}$ along with optimum values of sample sizes and C_{1d}, C_{2d}, C_{3d} and C_{4d} are given in Table 1 for domain 1, in Table 2 for domain 2 and in Table 3 for domain 3. The %RIEC of the estimator is given as $\frac{(C-C')}{C}\times 100$. Empirical study was implemented using SAS 9.3 package.

4 Discussion and Conclusion

A close perusal of Table 1 shows that for domain 1, the percentage reduction in expected cost decreases with increase in per unit travel and miscellaneous cost at the first phase (C_{1d}) for the proposed estimator. The percentage reduction in expected cost increases with increase in per unit travel and miscellaneous cost at the second phase (C_{2d}). The percentage reduction in expected cost increases with the increase in cost per unit of collecting the information on the study character in the first attempt (C_{3d}) and it decreases with increase in cost per unit of collecting the information by expensive method after the first attempt failed (C_{4d}).

The results in Table 2 show that for domain 2, the percentage reduction in expected cost decreases with increase in per unit travel and miscellaneous cost at the first phase (C_{1d}) for the proposed estimator. The percentage reduction in expected cost increases with increase in per unit travel and miscellaneous cost at the second phase (C_{2d}). The percentage reduction in expected cost increases with the increase in cost per unit of collecting the information on the study character in the first attempt (C_{3d}) and it decreases with increase in cost per unit of collecting the information by expensive method after the first attempt failed (C_{4d}). For domain 3, results in Table 3 reveal that the percentage reduction in expected cost decreases with increase in per unit travel and miscellaneous cost at the first phase (C_{1d}) for the proposed estimator. The percentage reduction in expected cost increases with increase in per unit travel and miscellaneous cost at the

second phase (C_{2d}). The percentage reduction in expected cost increases with the increase in cost per unit of collecting the information on the study character in the first attempt (C_{3d}) and it decreases with increase in cost per unit of collecting the information by expensive method after the first attempt failed (C_{4d}).

A close look of all the tables reveals that, the %RIEC is highest in case of domain 1 and it is almost the same in the other two domains. Hence, from the point of view of the %RIEC, the estimator of the domain total in the presence of nonresponse when domain size is assumed unknown was found to be better than the estimator with no nonresponse.

References

Agrawal, M. C., and Mrida, C. K. (2007). Some efficient estimators of the domain parameters. *Statistics and Probability Letters*, **77**, 704-709.

Chhikara, Raj S., and Sud, U. C. (2009). Estimation of population and domain totals under two-phase sampling in the presence of non-response. *Journal of the Indian Society of Agricultural Statistics,* **63**(3), 297-304.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley & Sons, Inc.

Hansen, M. H., and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, **41**, 517-529.

Okafor, F. C. (2001). Treatment of non-response in successive sampling. *Statistica,* **61(2)**, 195-204.

Okafor, F. C. (2005). Sub-sampling the non-respondents in two-stage sampling over successive occasions. *Journal of Indian Statistical Association*, **43**, 33-49.

Okafor, F. C., and Lee, H. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. *Survey Methodology*, **26 (2),** 183-188.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, **12**, 1-16.

Oh, H. L., and Scheuren, F. J. (1983). Weighting adjustment for unit non-response. In: W.G. Madow, I. Olkin, and B. Rubin (eds.), *Incomplete data in sample surveys*, Vol.2. New York: Academic press, 143-184.

Rancourt, E., Lee, H., and Särndal, C. E. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded non-response. *Survey Methodology*, **20**, 137-147.

Singh, R., and Mangat, N. P. S. (1996). *Elements of Survey Sampling*. Kluwer Academic publishers.

Sarndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Sud, U. C., Chandra H., and Chikkara, R. S. (2010). Domain estimation in the presence of non-response. *Journal of the Indian Society of Agricultural Statistics*, **66(3)**, 447-457.

Sud, U. C., Aditya, K., Chandra, H., and Parsad R. (2012). Two Stage Sampling For Estimation of Population Mean With Sub-Sampling of Non-respondents. *Journal of the Indian Society of Agricultural Statistics*, **64**, 343-347.

Tripathi, T. P., and Khare, B. B. (1997). Estimation of mean vector in presence of non-response. *Communications in Statistics - Theory and Methods*, **26(9)**, 2255 - 2269.

Author for correspondence

Hukum Chandra
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA
New Delhi
e-mail: *hchandra@iasri.res.in*