Estimating Domain-wise Distribution of Scarce Objects by Adaptive Sampling and Model-based Borrowing of Strength

Arijit Chaudhuri and Sanghamitra Pal[1]
Indian Statistical Institute, Kolkata-700 108
(Received : October, 2002)

SUMMARY

Utilizing the known (1) geographical areas (z) of the districts in India and (2) those of the total wastelands (x) therein we consider estimating (3) the total unknown areas (y), under 'Mining and Industrial Wastelands' for the groups of districts together but separately in the Northern, Southern, Eastern and Western regions in India, restricting to districts each possessing at least 5 percent of its total area as a wasteland. The total number of such districts in these respective 4 regions of separate interest are 48, 48, 42 and 91 giving a total of 229 out of which we consider sampling a total of 73 districts employing Rao – Hartley – Cochran (RHC, 1962) scheme of sampling using the known values of z above as the size-measures. Treating the above 4 regions of districts as the 4 domains of interest we consider utilizing known x above as a regressor in estimating the 'domain total' values of y above to form an idea of the distribution of these district-wise scarce objects in these regions.

For this we employ (a) non-synthetic as well as (b) synthetic versions of generalized regression (greg) estimators motivated respectively by postulated regression lines of y on x through the origin, for simplicity, with (i) domain-specific and alternatively with (ii) domain-invariant 'slope-parameters'.

Next we employ empirical Bayes estimators (EBE) with these greg estimators as the 'initials' with further specifications in the models.

Finally, in order to capture more districts beyond the 'initial sample' accommodating the rare commodities namely the 'mining and industrial wastelands' we employ the technique of Adaptive sampling defining appropriate (1) 'neighbourhoods' and (2) 'networks'. One may refer to Thompson (1992), Thompson and Seber (1996) and Chaudhuri (2000) for a discussion on adaptive sampling technique. The resulting relative performances of the alternative estimators noted above based on 'initial' and 'adaptive' samples are numerically examined through a simulation exercise utilizing known values of all the 3 variables noted above based on a given set of 'Remote sensed' observations.

[1] River Research Institute, Government of West Bengal, Salt-Lake, Kolkata-700 091

The synthetic greg estimates based on adaptive samples turn out to be the most promising ones in terms of the standard twin criteria of (A) actual coverage percentage (ACP) of confidence intervals (CI) based on assumed normality of a standardized 'pivotal' derived from a 'domain-specific' estimator and of (B) average coefficient of variation (ACV) of an estimator-both calculated from 'replicated samples'.

Key words : Adaptive sampling, Domain estimation, Empirical Bayes estimator, Generalized regression predictor, Modelling, Unequal probability sampling.

1. Introduction

From the website 'envfor.nic.in/naeb/naeb.html" entitled "The National Wasteland Identification Project" (NWIP) we gather certain data relating to 48, 48, 42 and 91 districts, each with at least 5 per cent of its total area as 'a Wasteland' area respectively in the northern region of UP, Haryana, Himachal Pradesh, Punjab and Jammu & Kashmir states, the southern region of Karnataka, Andhra Pradesh, Tamil Nadu and Kerala states, the eastern region composed of Arunachal Pradesh, Nagaland, Manipur, Assam, West Bengal, Orissa and Bihar and the western region consisting of the states of Maharashtra, Gujarat, Goa, Rajasthan and Madhya Pradesh.

For each of these 229 districts are separately known the total (1) geographical area (z), (2) the total 'wasteland area' (x) and (3) the total 'mining and industrial wasteland area' (y). Since the value of y for many of the districts is zero while when it is positive magnitude is substantial and 'how far the remote-sensed data on y matches the ground realities' is unknown, we consider it useful to prescribe, through a prior investigation, a fruitful method of (A) sampling of these 229 districts and of (B) estimating the total values of y for all the districts together but separately within the above-noted 4 regions of interest.

Using the known values of z as size-measures it seems plausible to adopt a suitable 'unequal probability sampling' scheme to start with and since x-values are known, a generalized regression estimator seems worthy of application. Further, since even with as high as a 25% sample of districts we may not find enough 'region-wise' sample-sizes, it may be useful to apply the 'principle of borrowing strength' as in small area estimation by appropriate modelling. Finally, since y is positive only for a very few districts region-wise, in order to capture more districts with positive y's we may contemplate employing adaptive sampling to extend the original sample to hope for improved estimation.

In section 2 we describe the procedures of sample selection and the estimation methods along with the motivating models. In section 3 we present a numerical evaluation of the competing procedures by a simulation exercise. We give our recommendations in Section 4 with which we conclude.

2. Sampling and Estimation Methods

For a simple presentation we need the following notations. Let $U = (1, \ldots, i, \ldots, N)$ denote a population of units labelled $i = 1, \ldots, N$ and let this be a union of D non-overlapping sets of units U_d, called 'domains', with known sizes N_d, $d = 1, \ldots, D$. Let y_i, x_i, z_i, $i \in U$ be the values of the variables respectively y, x, z with (1) totals Y, X, Z and (2) domain totals Y_d, X_d, Z_d, $d = 1, \ldots, D$. By $p_i = \frac{z_i}{Z}$, we shall denote the 'normed size-measures' of the units.

From U let a sample of n units be chosen employing the Rao-Hartley-Cochran (RHC, 1962) scheme. For this, U is randomly divided into n groups of $M_1, \ldots, M_i, \ldots, M_n$ units with M_i's as integers closest to N/n with their sum $\Sigma_n M_i$ over the n groups equal to N. From the ith group so formed one unit, say, ij is chosen with a probability $\frac{p_{ij}}{r_i}$, writing $r_i = p_{i1} + \ldots + p_{iMi}$; this is repeated independently over all these n groups.

Let $I_{di} = 1$ if $i \in U_d$; 0 else and (p_i, y_i) be the normed size-measure and the y-value for the unit chosen from the ith group. Let Σ denote summing over i in U. Then

$Y_d = \Sigma y_i I_{di}$ and RHC's unbiased-estimator for Y_d is

$$\hat{Y}_d = \Sigma_n \frac{r_i}{p_i} y_i I_{di}$$

Writing $B = \frac{\Sigma_n M_i^2 - N}{N^2 - \Sigma_n M_i^2}$, RHC's unbiased estimator of $V(\hat{Y}_d)$, the variance of $\hat{Y}_d$ is $v(\hat{Y}_d) = B \Sigma_n \Sigma_n r_i r_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 I_{di} I_{dj}$, writing $\Sigma_n \Sigma_n$ as sum over pairs of distinct groups with no overlaps. Let us postulate a model so that we may write

$$y_i = \beta_d x_i + \epsilon_i, i \in U_d, d = 1, \ldots, D$$

with β_d as constants and ϵ_i's as random variables. Following Chaudhuri *et al.* (1995) we may employ the following version of a possible improvement upon $\hat{Y}_d$, namely

$$t_{gd} = \Sigma_n \frac{r_i}{p_i} y_i I_{di} + b_{Qd}\left(X_d - \Sigma_n \frac{r_i}{p_i} x_i I_{di} \right) = \Sigma_n \frac{r_i}{p_i} g_{di} y_i I_{di}$$

Here $b_{Qd} = \frac{\Sigma_n y_i x_i Q_i I_{di}}{\Sigma_n x_i^2 Q_i I_{di}}$ with Q_i as a suitably assignable positive constant, for example, as $\frac{1}{x_i}$, $\frac{1}{x_i^2}$, $\frac{r_i}{p_i x_i}$, $\frac{1 - \frac{p_i}{r_i}}{\frac{p_i}{r_i} x_i}$ etc. and

$$g_{di} = 1 + \left(X_d - \Sigma_n \frac{r_i}{p_i} x_i I_{di} \right) \frac{x_i Q_i \frac{p_i}{r_i}}{\Sigma_n x_i^2 Q_i I_{di}}$$

In this article we shall mostly take Q_i as $\left(\frac{1 - \frac{p_i}{r_i}}{\frac{p_i}{r_i} x_i} \right) = \frac{r_i - p_i}{p_i x_i}$. Letting $e_{di} = y_i - b_{Qd} x_i$, following Sarndal (1982) the mean square error (MSE) of t_{gd} about Y_d may be estimated by

$m_{kgd} = B\Sigma_n \Sigma_n r_i r_j \left(\frac{e_{di} a_{kdi}}{p_i} - \frac{e_{dj} a_{kdj}}{p_j} \right)^2 I_{di} I_{dj}$, $k = 1, 2$ on writing $a_{1di} = 1$ and $a_{2di} = g_{di}$. In order to improve upon t_{gd} by 'borrowing strength' from outside the 'intersection of the sample with U_d' but within the initially chosen sample s, let us postulate an alternative model for which β_d above is replaced by β for every d but keeping everything else intact. This revised model motivates the 'synthetic' greg predictor for Y_d as t_{gSd} which is t_{gd} with b_{Qd} replaced by $b_Q = \frac{\Sigma_n y_i x_i Q_i}{\Sigma_n x_i^2 Q_i}$. Then we may write

$$t_{gSd} = \Sigma_n \frac{r_i}{p_j} y_i \left[I_{di} + \left(X_d - \Sigma_n \frac{r_i}{p_i} x_i I_{di} \right) \frac{x_i Q_i \frac{p_i}{r_i}}{\Sigma_n x_i^2 Q_i} \right]$$

$$= \Sigma_n \frac{r_i}{p_j} y_i g_{sdi}, \text{ say}$$

with g_{sdi} as 'within the square brackets'. Then, following Sarndal (1982), MSE-estimators for t_{gSd} are

$$m_{kgSd} = B\Sigma_n\Sigma_n r_i r_j\left(\frac{e_i b_{kdi}}{p_i} - \frac{e_j b_{kdj}}{p_j}\right)^2 I_{di} I_{dj},\ k = 1, 2$$

on writing

$$e_i = y_i - b_Q x_i, b_{1di} = I_{di}, b_{2di} = g_{sdi}$$

In contrast with t_{gSd}, the t_{gd} is a 'non-synthetic' greg predictor.

Writing t_d for an initial estimator/ predictor for Y_d let us now postulate the more sophisticated model permitting us to write

(i) $t_d \mid Y_d \overset{ind}{\bigcap} N(Y_d, m_d)$, with m_d as a known MSE-estimator for t_d

(ii) $Y_d \overset{ind}{\bigcap} N(\theta X_d, A), \theta,$ A as unknown constants

(iii) $\epsilon_d = (t_d - Y_d)$ "independent" of $\eta_d = Y_d - \theta X_d$ for $d = 1, \ldots, D$

Then, from Fay and Herriot (1979) we have

$$t_{Bd} = \left(\frac{A}{A + m_d}\right)t_d + \left(\frac{m_d}{A + m_d}\right)(\theta X_d)$$

as the Bayes estimator of Y_d, $d = 1, \ldots, D$

$$\text{Writing } \hat{\theta} = \frac{\sum_{d=1}^{D} t_d X_d \Big/ (A + m_d)}{\sum_{d=1}^{D} X_d^2 \Big/ (A + m_d)}$$

and solving by iteration for θ and A starting with a 'zero value for A', the equation

$$\sum_{d=1}^{D} (t_d - \tilde{\theta} X_d)^2 \Big/ (A + m_d) = D - 1$$

we may derive moment estimates $\hat{\theta}$, $\hat{A}$ respectively for θ, A. Then

$$t_{EBd} = \left(\frac{\hat{A}}{\hat{A} + m_d}\right)t_d + \left(\frac{m_d}{\hat{A} + m_d}\right)(\hat{\theta} X_d)$$

gives the EBE for Y_d.

From Prasad and Rao (1990) we get the MSE-estimator for t_{EBd} as

$$m_{EBd} = g_{1d}(\hat{A}) + g_{2d}(\hat{A}) + 2g_{3d}(\hat{A})$$

where $g_{1d}(\hat{A}) = \gamma_d m_d$

$$\gamma_d = \frac{\hat{A}}{\hat{A} + m_d}$$

$$g_{2d}(\hat{A}) = (1 - \gamma_d^2)\frac{X_d^2}{\sum_{d=1}^{D} \frac{X_d^2}{(\hat{A} + m_d)}}$$

$$g_{3d}(\hat{A}) = \frac{m_d^2}{(\hat{A} + m_d)^3} V(\overline{A})$$

where $$V(\overline{A}) = \frac{2}{D^2}\sum_{d=1}^{D}(\hat{A} + m_d)^2$$

For the validity of m_{EBd}, D is required to be large. But in the present case we employ this even though D is only 4 hoping that this may still work.

Suspecting that the initial sample s drawn as above may not yield enough units with positive values of y 'respective domain-wise', we may apply in the following way the technique of adaptive sampling to enhance the capture of more sampled units with positive and possibly high positive y-values.

For every unit, namely district in the present investigation, let a 'neighbourhood' be defined as the collection of districts including this unit itself and those with a common boundary with it as is determined from the map of the 229 districts we are considering.

Any unit, rather district with a zero value for y is called an 'edge' unit or a singleton network. For any unit with a positive y-value one should check for the positive/ zero-value of y for each of its neighbouring units and proceed with this checking until every neighbouring unit has a zero value. The 'set of units thus checked starting with the positive y-valued unit' constitutes a 'cluster' for the unit including itself. Those units with positive y-values in the cluster constitute a 'network' for the initial unit. Writing A(i) for the 'network' to which the unit i belongs and m_i for its cardinality, let

$$t_i = \frac{1}{m_i}\sum_{k \in A(i)} y_k, \; l_i = \frac{1}{m_i}\sum_{k \in A(i)} x_k$$

Then, as is recorded by Chaudhuri (2000), one may check that

$T = \sum_{i \in U} t_i$ equals Y and $L = \sum_{i \in U} l_i$ equals X

Similarly, letting

$$t_{id} = \frac{\sum_{jI_{dj} \in A(i)} y_i I_{dj}}{\left(\sum_{jI_{dj} \in A(i)} I \right)} \text{ and } \frac{\sum_{jI_{dj} \in A(i)} x_j I_{dj}}{\left(\sum_{jI_{dj} \in A(i)} I \right)}$$

it follows that

$$Y_d = \sum_{i \in U} t_{id} = T_d, \text{ say}$$

$$X_d = \sum_{i \in U} l_{id} = L_d, \text{ say}$$

The collection of the units in the original sample s together with those in their respective clusters constitutes an adaptive sample.

Corresponding to $\hat{Y}_d$ the RHC-estimator for Y_d based on the adaptive sample is

$$\hat{Y}_d(A) = \Sigma_n \frac{r_i}{p_i} t_i I_{di}$$

Similarly corresponding to t_{gd} the non-synthetic greg predictor for Y_d based on the adaptive sample is

$$t_{gd}(A) = \left(\Sigma_n \frac{r_i}{p_i} t_{id} I_{di} \right) + b_{Qd}(A) \left(X_d - \Sigma_n \frac{r_i}{p_i} l_{id} I_{di} \right) \text{ writing}$$

$$b_{Qd}(A) = \frac{\Sigma_n t_{id} l_{id} Q_i I_{di}}{\Sigma_n l_{id}^2 Q_i I_{di}}$$

Since l_{id} is often zero, we shall take Q_i as $\left(1 - \frac{p_i}{r_i}\right) \Big/ \frac{p_i}{r_i}$ omitting l_{id} in the denominator which we might use as equivalent to x_i.

The variance estimator for $\hat{Y}_d(A)$ is given by

$$v(\hat{Y}_d(A)) = B \Sigma_n \Sigma_n r_i r_j \left(\frac{t_i}{p_i} - \frac{t_j}{p_j} \right)^2 I_{di} I_{dj}, \; k = 1, 2$$

The MSE estimators for $t_{gd}(A)$ are $m_{kgd}(A)$ obtained from m_{kgd} replacing therein y_i by t_{id}, x_i by l_{id}, b_{Qd} by $b_{Qd}(A)$.

Instead of t_{gSd} we shall employ $t_{gSd}(A)$ for the adaptive sample obtained on replacing y_i, x_i by t_{id}, l_{id} in the former. The MSE estimator for $t_{gSd}(A)$ will be taken as $m_{gSd}(A)$ obtained from m_{1gd} on replacing y_i, x_i in the latter by t_{id} and l_{id} respectively in the terms involving I_{di} and by t_i, l_i for the terms free of I_{di}. Because of the form of $m_{gSd}(A)$ it is not possible to use a second MSE-estimator corresponding to m_{gSd} because

$$t_{gSd}(A) = \Sigma_n \frac{r_i}{p_i} t_{id} I_{di} + \left(L_d - \Sigma_n \frac{r_i}{p_i} l_{id} I_{di} \right) \left(\frac{\Sigma_n t_i l_i Q_i}{\Sigma_n l_i^2 Q_i} \right)$$

cannot be expressed as a weighted sum of $(t_{id} I_{di})$-values. Corresponding to (t_{EBd}, m_{kEBd}), (t_{EBSd}, m_{kEBSd}) we obviously have $(t_{EBd}(A), m_{kEBd}(A))$, $k = 1, 2$ and $(t_{EBSd}(A), m_{EBSd}(A))$ with obvious notations for the EB estimators based on adaptive sampling and the MSE-estimators corresponding to m_{kEBd}, $k = 1, 2$ and m_{1EBSd}.

3. Simulation-based Numerical Evaluation of Relative Efficacies

Given an estimator/predictor f_d for Y_d with an MSE-estimator v_d we shall treat $s_d = (f_d - Y_d)/\sqrt{v_d}$ as a standard normal deviate and take $(f_d - 1.96\sqrt{v_d}, f_d + 1.96\sqrt{v_d})$ as the 95% confidence interval (CI) for Y_d. To compare alternative choices of (f_d, v_d) we shall calculate, based on $R = 1000$ replicates of the samples, the criteria measures (I) ACP the actual coverage percentage which is the percent of the replicated samples with CI's covering Y_d - the closer it is to 95 the better and (II) ACV, the average coefficient of variation namely the average over the R replicates of the values of $100 \frac{\sqrt{v_d}}{f_d}$ – the less it is the less the width of CI and the more accurate is the point estimator f_d for Y_d.

For the NWIP data mentioned earlier our numerical observations are as in the table below.

It may be noted that the number of districts to be covered by adaptive sampling varies between 117 and 146 with an average of 134 while the initial sample size is only 73. Adaptive sampling always involves additional costs.

The question is whether and how much it pays in terms of gain in accuracy in estimation.

Table. Relative Efficacies of alternative procedures

		ACP/ACV Values			
S. No.	Domain Specifications numbered (d)	North	South	East	West
		(1)	(2)	(3)	(4)
	Domain size				
	N_d	48	48	42	91
	f_d/v_e				
1.	$\hat{Y}_d / v(\hat{Y}_d)$	64.3/11.1	68.5/14.6	62.5/13.2	67.1/12.6
2.	$\hat{Y}_d(A) / v(\hat{Y}_d(A))$	74.1/10.6	70.5/11.3	65.9/12.9	71.6/9.8
3.	t_{gd}/m_{1gd}	80.2/20.0	83.4/24.1	80.6/23.9	87.1/14.1
4.	t_{gd}/m_{2gd}	82.1/21.3	84.4/27.2	83.5/22.2	88.0/17.2
5.	$t_{gd}(A)/m_{1gd}(A)$	87.1/15.6	89.4/19.3	86.3/21.4	93.0/9.8
6.	$t_{gd}(A)/m_{2gd}(A)$	89.4/19.1	92.7/26.3	88.1/25.9	93.6/10.3
7.	t_{EBd}/m_{1EBd}	90.3/23.5	91.1/32.0	84.1/39.1	92.1/22.2
8.	t_{EBd}/m_{2EBd}	92.5/31.4	93.7/34.1	83.1/32.3	97.1/31.6
9.	$t_{EBd}(A)/m_{1EBd}(A)$	94.6/41.3	90.1/29.2	86.3/41.4	96.3/29.5
10.	$t_{EBd}(A)/m_{2EBd}(A)$	94.4/37.4	92.1/32.2	88.1/43.5	95.1/24.7
11.	t_{gSd}/m_{1gSd}	88.8/27.1	82.1/29.1	85.6/25.1	89.6/18.2
12.	t_{gSd}/m_{2gSd}	91.3/28.5	84.9/29.3	83.6/24.9	90.0/19.1
13.	$t_{gSd}(A)/m_{gSd}(A)$	93.9/18.5	92.4/20.1	89.9/23.8	96.1/10.5
14.	t_{EBSd}/m_{1EBSd}	95.4/24.3	90.3/34.9	89.1/40.2	95.1/24.1
15.	t_{EBSd}/m_{2EBSd}	95.8/32.9	90.6/29.6	94.8/37.1	96.1/29.5
16.	$t_{EBSd}(A)/m_{EBSd}(A)$	95.3/42.1	94.9/30.1	91.2/43.2	95.1/26.3

4. *Concluding Remarks and Recommendations*

If guided by the criterion of ACP, one may be convinced that empirical Bayes estimators for adaptive samples as well as the original samples fare better than t_{gd} and t_{gSd} and more so if coupled with m_{2gd}, m_{2gSd} respectively rather than with m_{1gd}, m_{1gSd}.

Moreover, adaptive sampling coupled with non-synthetic, synthetic greg estimators and the empirical Bayes estimators based thereupon seems to have an edge over the original one.

In terms of the ACV criterion empirical Bayes methods perform poorer than the initial ones on which they are based. But adaptive sampling achieves improvements when combined with t_{gd} with both m_{1gd}, m_{2gd} and also with t_{gSd} but the ACV increases when it is used with empirical Bayes versions of their greg estimators. Taking both the criteria together, the synthetic greg estimator

$t_{gSd}(A)$ based on adaptive sampling seems to be the most promising one. So, if the resources permit, our recommendation is in favour of adaptive sampling even at an additional cost. Compared to (t_{gSd}, m_{kgSd}), k = 1, 2 the pair $(t_{gSd}(A), m_{gSd}(A))$ is a better choice - this vindicates the efficacy of adaptive sampling. Keeping in mind simultaneously the width of the confidence interval and the accuracy in point estimation, empirical Bayes procedure does not seem to be a right option in the present exercise. But adaptive sampling coupled with synthetic greg estimator is a promising choice.

A possible reason for a partial failure of the empirical Bayes estimation approach in the present exercise may be the inadequacy of m_{EBd} as an MSE-estimator in view of the number of domains here being too small-only four.

REFERENCES

Chaudhuri, A. (2000). Network and adaptive sampling with unequal probabilities. *Cal. Stat. Assoc.* Bull., **50,** 237-253.

Chaudhuri, A. and Maiti, T. (1995). On the regression adjustments to Rao-Hartley-Cochran estimator. *Jour. Stat. Res.,* **29,** 71-78.

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

Prasad, N.G.N. and Rao, J.N.K (1990). The estimation of the mean square error of small area estimates. *Jour. Stat. Assoc.,* **85,** 163-171.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. Roy. Statist. Soc.,* **B24,** 482-491.

Sarndal, C.E. (1982). Implications of survey design for generalized regression estimation of linear Junctions. *J. Statist. Plann. Inf.*, **7,** 155-170.

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

Thompson, S.K. and Seber, G.A.F. (1996). *Adaptive Sampling.* John Wiley & Sons, New York.