Small area estimation of proportions of Arsenic affected wells in Bangladesh

By

Sanghamitra Pal

West Bengal State University, India

(Joint work with Prof. Partha Lahiri)

Agenda

- Problem Statement
- Proposed Solution
- Simulation Results
- Conclusion
- References

Problem Statement

Arsenic – a Health Hazard

- **Arsenic (As): toxic metal** --- widespread in groundwater in many countries

- India(especially in Bengal), Bangladesh, Nepal, Thailand, China, Mongolia and Tibet, Viet Nam, Laos, Cambodia, Myanmar, various South American countries and areas in North America and Western Australia-----------As affected


- Negative health impacts are related to:
  - its concentration in food or water

As Level Limits

- WHO guidelines for maximum level of As in drinking water:

  10 μg/L for safe water

- Different countries have adopted different standards for As

- Bangladesh: 50 μg/L

Data Map

- In 1997 British Geological Survey had taken out a project "Survey on Arsenic affected wells in Bangladesh"

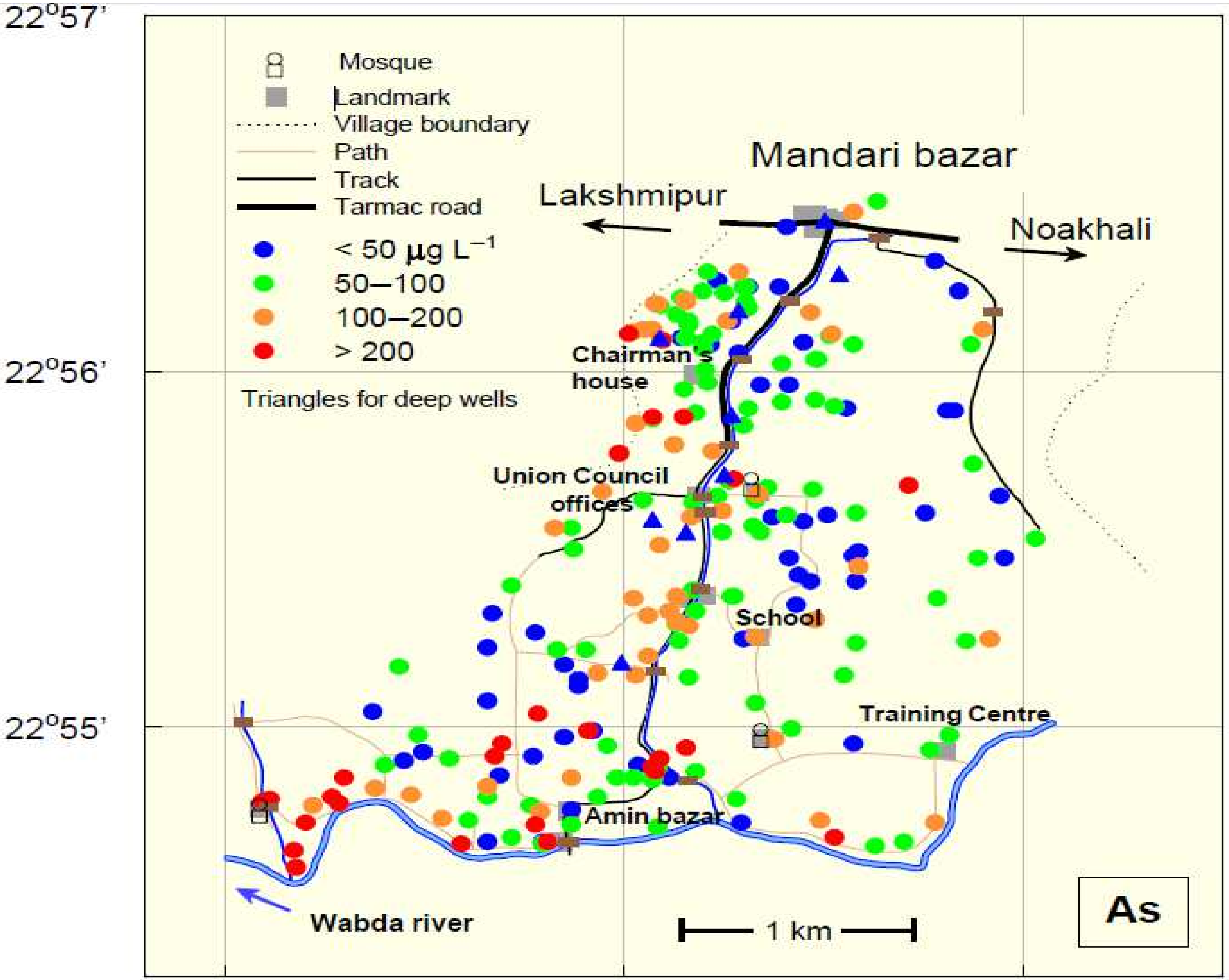
- A sample of 3540 wells were surveyed to measure Arsenic affected wells

- Here we are going to estimate District wise proportion of wells less than the threshold value

Data: BGS Survey on As of Bangladesh

Sample_ID	Latitude	Longitude	Yr_ Const	Well type	Well Depth (m)	owner	division	district	As (Ug/L)
S-98-00	22.87	90.78	1992	Shallow	10.7	--	Chittagong	Lakshmipur	13
S-98-01	23,02	90.87	1971	HP	7.2	--	Dhaka	Faridpur	256

Map showing the distribution of As in Mandari



Problem & proposed solution

- District-wise proportion of arsenic affected wells
- Problem of Small area estimation
- Districts : small areas (Number of districts =61)
- Normal/Normal model
- Beta-Binomial Model
- Benchmarking (Number of Divisions=7)

Problem

- y_{ij}=arsenic level for well j in ith district ; t: threshold value $I(y_{ij} \le t) = 1,\ i = 1,..,m$

  $m = \text{No of districts}$
- **Population proportion** $\pi_i = \frac{(\#\ wells\ \text{ in POPU.})\ < t}{N_i}$
-
- **Sample proportion** $p_i = \frac{\#\ wells \text{ in Sample} < t}{n_i}$ x_i
-
- N_i = Population size for ith district
- And n_i = Sample size for ith district

Covariate:

x_i =coverage (person per water source) in district i.

The Fay-Herriot Model (FH Model)

Sampling Model :

$$p_i / \pi_i \overset{ind}{\sim} N(\pi_i, D_i)$$

Linking Model :

$$\pi_i \overset{ind}{\sim} N(x_i' \beta, A)$$

x_i

Linear Mixed Model :

$$p_i = \pi_i + e_i = x_i' \beta + V_i + e_i$$

Where $e_i \sim N(0, D_i)$

$V_i \sim$ N(0, A)

Sampling variance : D_i (Known)

Model variance : A (Unknown)

(Fay - Herriot, 1979)

Small area estimation

Fay-Herriot (FH) Model (1979)

An empirical Bayes estimator of π_i is given by

$$\hat{\pi}_i^{EB} = (1-\hat{B}_i)\,p_i + \hat{B}_i\hat{\mu}_i$$

$$\hat{B}_i = \frac{D_i}{\hat{A}+D_i},\; D_i = \frac{\overline{p}\,\overline{q}}{n_i}\ \text{(Morris, 1983)},\quad \overline{p} = \frac{\sum_1^m N_j p_j}{\sum_1^m N_j}$$

$$\hat{\mu}_i = x_i^T\hat{\beta},\quad \hat{\beta}^T = (\beta_0, \beta_1)$$

$$\hat{\beta} = (X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}p$$

$$p = (p_1,\dots\dots\dots p_m) \qquad V = diag(A+D_1,\dots\dots\dots, A+D_m)$$

$\hat{A}, \hat{\beta}_0, \hat{\beta}_1$ are obtained from REML

Fay-Herriot Model (Contd…)

MSE estimation:

1. **Datta-Lahiri (2000) , Prasad-Rao (1990)**

$$mse(\hat{\pi}_i^{EB}) = g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2g_{3i}(\hat{A})$$

$$\text{where } g_{1i}(A) = (1 - B_i)D_i$$

$$g_{2i}(A) = B_i^2 Var(x_i^T \hat{\beta}) = B_i^2 x_i^T \left(\sum_1^m \frac{1}{A + D_j} x_j x_j^T\right)^{-1} x$$

$$g_{3i}(A) = \frac{D_i^2}{(A + D_i)^3} \cdot \frac{2}{\sum_1^m (A + D_j)^{-2}}$$

Arc-Sine Transformation

Apply above following FH model

- Back-Transformation to get CI for the Population proportion

$$y_i = \sqrt{n_i} Sin^{-1}(2p_i - 1)$$

$$\theta_i = \sqrt{n_i} Sin^{-1}(2\pi_i - 1)$$

Benchmarking

Benchmarking

- Seven divisions (large areas) in Bangladesh
- Use that data for benchmarking

Benchmarking with Divisions

With FH Model

– Define

$$l_j = \overline{p}_j - 1.96 se(\overline{p}_j)$$

$$u_j = \overline{p}_j + 1.96 se(\overline{p}_j)$$

$$\overline{p}_j = \sum_{k=1}^{di} W_{kj} p_k \qquad j = 1,2,.....,7$$

$$W_{kj} = \frac{N_k}{\sum_{i=1}^{d_i} N_i}, \quad se(\bar{p}_j) = \sqrt{\sum_{k=1}^{di} W_{kj}^{\ 2} \frac{p_k q_k}{n_k}}$$

d_j = No of district in division j

Benchmarked Confidence Intervals

$$\hat{\pi}_{i,lower} \frac{l_j}{\sum_{k=1}^{dj} W_{kj} \hat{\pi}_{k,lower}}, \hat{\pi}_{i,upper} \frac{u_j}{\sum_{k=1}^{dj} W_{kj} \hat{\pi}_{k,upper}}$$

$$\hat{\pi}_{i,lower} = \hat{\pi}_i^{EB} - 1.96 se(\hat{\pi}_i^{EB})$$

$$\hat{\pi}_{i,upper} = \hat{\pi}_i^{EB} + 1.96 se(\hat{\pi}_i^{EB})$$

Approximate Bayesian method :Beta-Binomial Model

Beta-Binomial:

$$u_i / \pi_i \sim Bin(n_i, \pi_i)$$
$$\pi_i \sim Beta[\mu_i, \gamma\mu_i(1-\mu_i)]$$
$$\mu_i = \frac{\exp(b_o + b_1 x_i)}{1 + \exp(b_o + b_1 x_i)}$$
(Lohr - Rao, 2009)

Approximate Bayesian method

$$\pi_i / data \sim Beta(mean = (1-\hat{B}_i) p_i + \hat{B}_i \mu_i = \hat{\pi}_i^{EB}$$
$$\text{var}\, iance = v_i)$$

Approximate Bayesian (Contd.)

$$\text{variance} = v_i = \hat{C}_i \hat{\pi}_i^{EB}(1-\hat{\pi}_i^{EB}) -$$

$$\frac{m-1}{m}\sum_{1}^{m}\{\hat{C}_i(-j)\hat{\pi}_i^{EB}(-j)[1-\hat{\pi}_i^{EB}(-j)] - \hat{C}_i\hat{\pi}_i^{EB}(1-\hat{\pi}_i^{EB})\}$$

$$+\frac{m-1}{m}\sum_{1}^{m}[\hat{\pi}_i^{EB}(-j)-\hat{\pi}_i^{EB}]^2$$

(Rao, 2003)

$\hat{\pi}_i^{EB}(-j)$ are calculated with Bayesian Jackknife Formula (Delete-one)

Approximate Bayesian (Contd.)

Confidence Interval with Beta-Binomial

- *Calculate shape parameters—find out CI*
- *Calculate Benchmarked Estimates proceeding as above*

Simulation Results

Simulation

Data source: BGS Survey in Bangladesh, 1997

We adopt Design based approach to see the performances of the estimators

Pseudo Population:

Generate $4n_i$ for the domain i to get a Population

For simplicity we adopt SRSWR to draw sample for simplicity only

Population ($N_i = 4n_i$)

$\Downarrow$ SRSWR

Sample (n_i)

Simulation – Comparison Criterion

ACP - Actual Coverage Percentage *(the closer to 95, the better)*

AL - Average Length of CI *(the Lesser the better)*

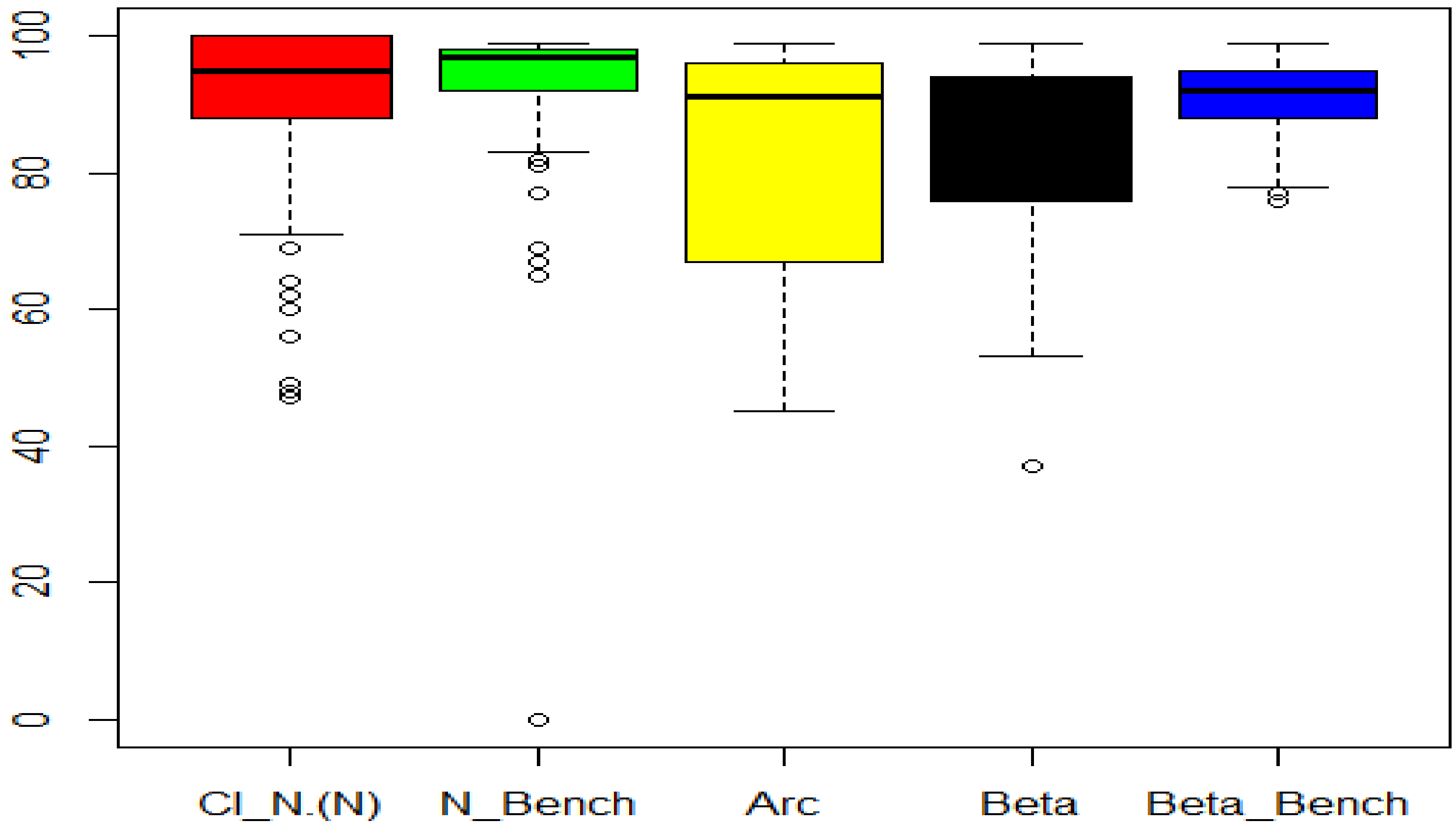
ACP, ACV and AL : all are calculated from replicated samples (1000 samples)

(1) CI_Normal: CI where "MSE estimation is by Dutta-Lahiri (REML) method"
(2) CI_Normal_Bench: Benchmarking on CI_Normal
(3) Arc-Sine transformation
(4) Beta: with Beta-Binomial model
(5) Bench_Beta Benchmarking on Beta

Results – Summary of ACP values

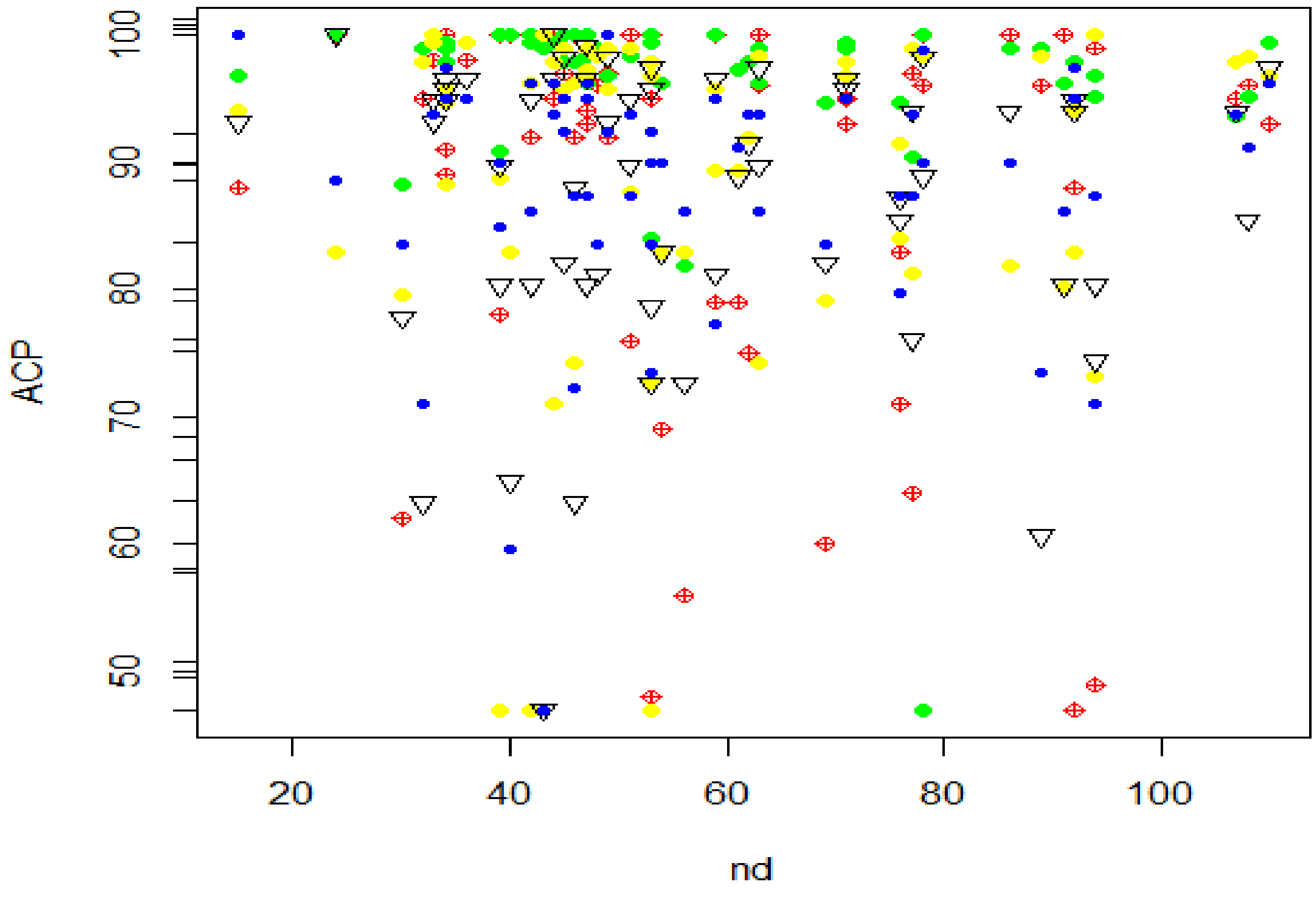
Summary	n_i	**CI_Normal**	**CI_Normal_Bench**	**Arc_Sine**	**Beta**	**Beta_Bench**
Min	15	47	59	45	37	76
1st Qu.	43	88	92	67	76	88
Median	53	95	97	91	89	92
Mean	57	89	91	81	84	90
3rd Qu.	76	100	98	96	94	95
Max	110	100	99	99	99	99

Results BOX Plots of ACP values under different methods



Results

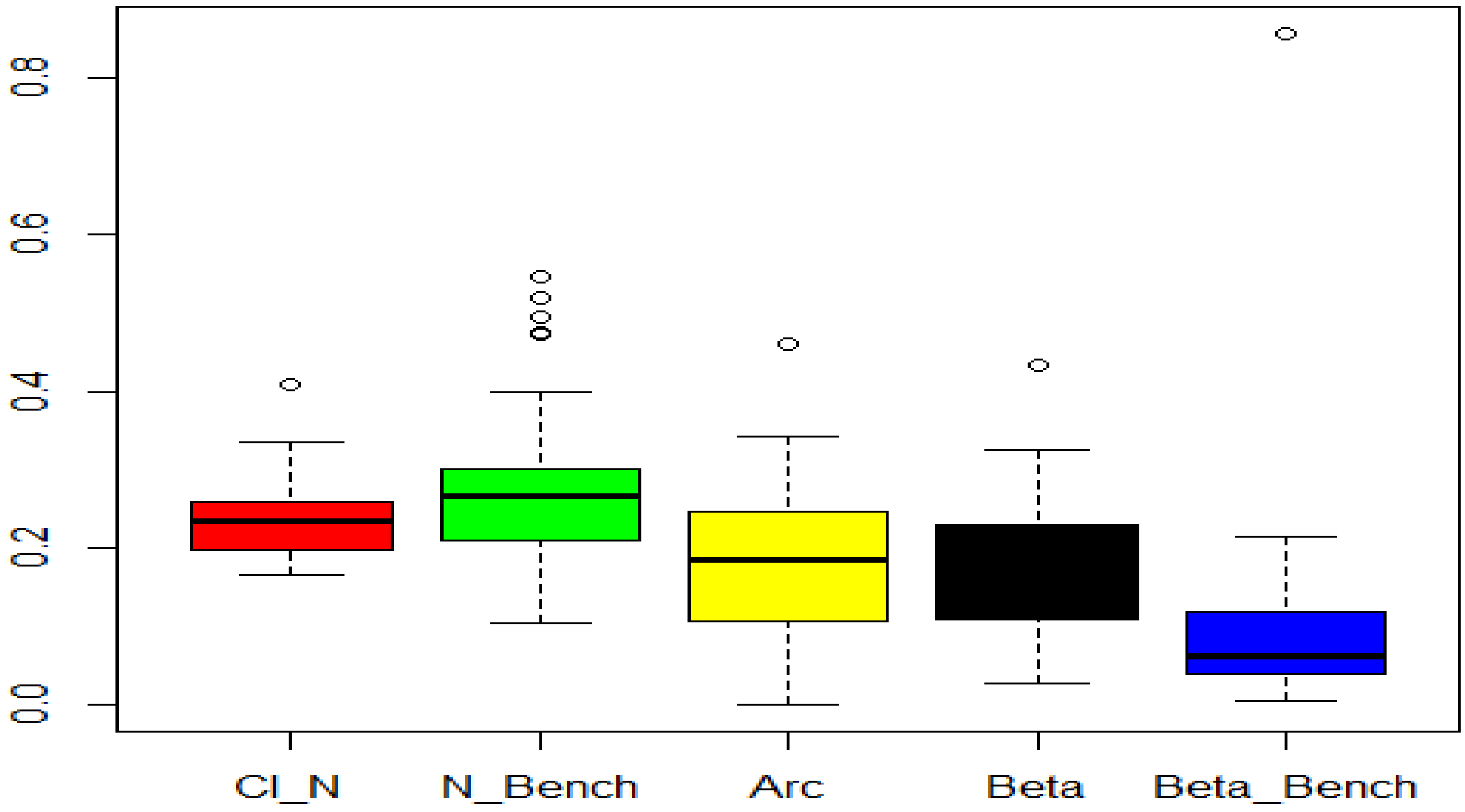
Red="CI_Normal"; Green=" CI_Normal _Bench"; Yellow="Arc-Sine";
Black="Beta"; Blue="Beta_Bench"



Summary of AL Values

Summary	n_i	CI_Normal	CI_Normal_ Bench	Arc_Sine	Beta	Beta_Bench
Min	15	.1652	.1041	.0011	.0285	.0051
1st Qu.	43	.1973	.2092	.1069	.1076	.0389
Median	53	.2341	.2665	.1855	.1831	.0626
Mean	57	.2359	.2367	.1696	.1751	.0907
3rd Qu.	76	.2580	.2134	.2467	.2287	.1160
Max	110	.4091	.4458	.4611	.4339	.3768

Results (Box-Plot of Average lengths of CI)



Conclusion

- We can not use Direct estimators as the se is zero for some domains.

- We have adopted SAE problem as the domain sizes are small

- <u>Benchmarked Empirical Bayes estimators perform better than others</u>

- We proceed with <u>Beta-Binomial Model with Benchmarking</u>

References

✓BGS and DPHE, 2001. Arsenic contamination of Groundwater in Bangladesh., *British*

✓*Geological Survey and Department of Public Health Engineering, Govt. of Bangladesh. Final report;* Vol-2, 267p

✓Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear

✓unbiased predictors in small area estimation problems. *Statist. Sinica* **10** 613–627.

✓Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations.

✓*J. Amer. Statist. Assoc.* **70** 311–319.

✓Fay, R. E., and Herriot, R. (1979). Estimates of income for small places: An application ofJames-Stein procedures to census data, J. Am. Statist. Ass., 74, 269-277.

✓Ghosh, M. and Rao, J.N.K. 1994). Small area estimation: an appraisal.Statistical Sc. 81, 1058-1062

✓Kinniburgh, D.G and Kosmus, W. Arsenic contamination in groundwater: some analytical

✓Considerations, Talanta 58 (2002) 165–180

References

✓Lohr, S. L. and Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. Biometrika 96 457–468.

✓Michael Berg, Hong Con Tran, Thi Chuyen Nguyen, Hung Viet Pham, Roland Schertenlieb, Walter Giger, .Arsenic contamination of groundwater and drinking water in Viet Nam: a human health threat., *Environmental Science and*

✓*Technology*, vol. 35, no. 13, 2001, pp. 2621.6.

✓**Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion).** ***J. Amer.Statist. Assoc.,*** **78, 47-65.**

✓Prasad, N. G. N., and Rao, J. N. K. (1990). The estimation of mean squared errors of small area estimators.

✓Journal of the American Statistical Association 85, pp. 163-171.

✓Rao, J. N. K. (2003). Small Area Estimation. John Wiley and Sons, Hoboken, New Jersey.

Thank You

Email: mitra_pal@yahoo.com