Granular level estimation using multiple data sources

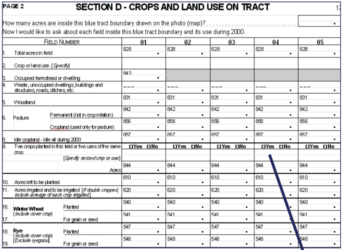
Partha Lahiri

University of Maryland College Park, USA

Based on my joint with Nicola Salvati, University of Pisa, Italy

UN Sustainable Development Goals (SDG)

An Example of Multiple Data Sources for SAE

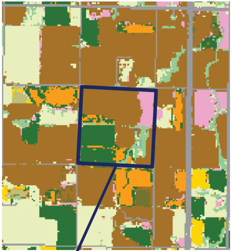


PAGE 2 **SECTION D - CROPS AND LAND USE ON TRACT** 17

How many acres are inside this blue tract boundary drawn on the photo (map)? . . . .

Now I would like to ask about each field inside this blue tract boundary and its use during 2000.

FIELD NUMBER	01	02	03	04	05
1. Total acres in field	828	828	828	828	828
2. Crop or land use. [*Specify*]					
3. Occupied farmstead or dwelling	843				
4. Waste, unoccupied dwellings, buildings and structures, roads, ditches, etc.	– – –	– – –	– – –	– – –	– – –
5. Woodland	831	831	831	831	831
6. Pasture — Permanent (not in crop rotation)	842	842	842	842	842
Pasture — Cropland (used only for pasture)	856	856	856	856	856
8. Idle cropland - Idle all during 2000	857	857	857	857	857
9. Two crops planted in this field or two uses of the same crop.	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No
[*Specify second crop or use*]					
Acres	844	844	844	844	844
10. Acres left to be planted	610	610	610	610	610
11. Acres irrigated and to be irrigated [*If double cropped, include acreage of each crop irrigated.*]	620	620	620	620	620
16. **Winter Wheat** (*include cover crop*) Planted	540	540	540	540	540
17. For grain or seed	541	541	541	541	541
18. **Rye** (*include cover crop*) [*Exclude ryegrass*] Planted	547	547	547	547	547
19. For grain or seed	548	548	548	548	548



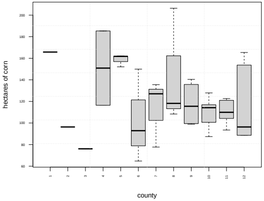
REGRESSION VARIABLES:

	Dependent Y	Independent X
	Enumerated JAS Segments	CDL Classified Acres
Soybeans	227	273
Wheat	337	541

An Example: Estimation of crop acreage at granular levels

Ref: Battese et al. (1988)

- Estimate crop acreage for 12 counties of north central Iowa
- Sampled unit: segment of land
- Combine survey data with satellite data
- 37 observations



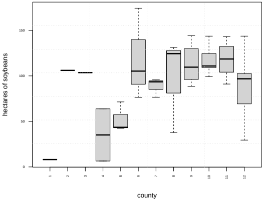
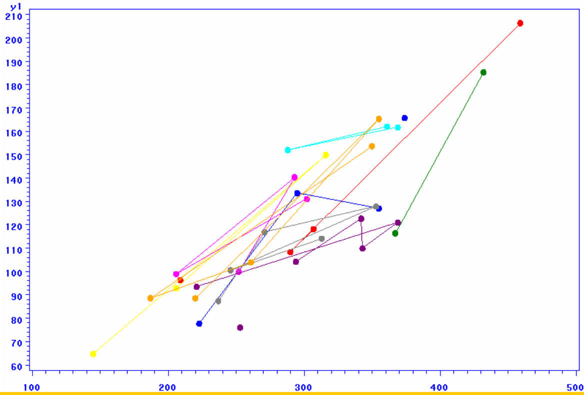


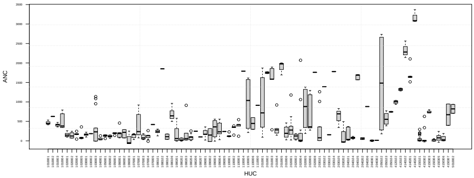
Fig 2: Plot of Corn Hectares versus Corn Pixels by County

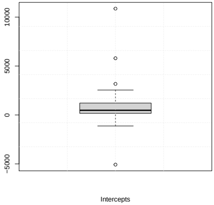


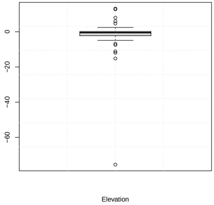
The distribution of estimated intercept and slopes by area

An Example from the EMAP Lake Survey Data

- 334 lakes selected from the population of 21,026 lakes
- 86 Hydrologic Unit Codes (HUCs) are in-sample
- 27 HUCs are out-of-sample
- Estimation of average Acid Neutralising Capacity (ANC) by HUC is of interest.







Notation

- m small areas with N_i units;
- y_{ij} and $\mathbf{x}_{ij}$ denote the values of the study variable and a $p \times 1$ vector of known auxiliary variables for the jth unit of the ith small area, respectively, with $i = 1, \ldots, m$, $j = 1, \ldots, N_i$;
- Parameter of interest: $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij},\ i = 1, \ldots, m$.
- n_i is the sample size for area i and it is not large enough to support the use of a direct estimator: $\bar{y}_i = n_i^{-1} \sum_{j \in s_i} y_{ij}$, where s_i denotes the part of the sample from the ith small area.

Nested error regression model (NER)

- Nested error regression model for the finite population:

$$y_{ij} = \beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i + \epsilon_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, N_i,$$

- β_0 and $\boldsymbol{\beta}$ are unknown fixed intercept and regression coefficients, respectively;
- γ_i is a random effect for area i; ϵ_{ij} is the sampling error for the jth observation in the ith area; γ_i and ϵ_{ij} are all assumed to be independent with $\gamma_i \sim N(0, \sigma_\gamma^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $i = 1, \ldots, m;\ j = 1, \ldots, N_i$;
- the parameters $\boldsymbol{\delta} = (\sigma_\gamma^2, \sigma_\epsilon^2)$ are referred to as the variance components.

An extension of NER

We propose the following extension of the nested error regression model:

$$y_{ij} = \beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \gamma_i + \epsilon_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, N_i,$$

- $\boldsymbol{\beta}_i$ is a $p \times 1$ vector of fixed unknown regression coefficients for area i;
- γ_i and ϵ_{ij} are all independent with $\gamma_i \sim N(0, \sigma^2_\gamma)$ and $\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon i})$.

The Best Predictor (BP)

The best predictor (BP) of $\theta_i = \beta_0 + \bar{\mathbf{X}}_i' \boldsymbol{\beta}_i + \gamma_i$ is given by

$$\begin{aligned}\hat{\theta}_i^{BP} \equiv \hat{\theta}_i(\boldsymbol{\phi}_i) &= \beta_0 + \bar{\mathbf{X}}_i' \boldsymbol{\beta}_i + (1 - B_i)(\bar{y}_i - \beta_0 - \bar{\mathbf{x}}_i' \boldsymbol{\beta}_i) \\ &= (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)' \boldsymbol{\beta}_i + \left\{ B_i(\beta_0 + \bar{\mathbf{x}}_i' \boldsymbol{\beta}_i) + (1 - B_i)\bar{y}_i \right\}\end{aligned}$$

- $\bar{\mathbf{X}}_i$: population mean for area i
- $\bar{\mathbf{x}}_i$: sample mean for area i
- $B_i = \frac{\sigma_{\epsilon i}^2 / n_i}{\sigma_{\epsilon i}^2 / n_i + \sigma_\gamma^2}$;
- $\boldsymbol{\phi}_i = (\beta_0, \boldsymbol{\beta}_i, \sigma_\gamma^2, \sigma_{\epsilon i}^2)'$;
- An empirical best predictor (EBP) of θ_i can be written as $\hat{\theta}_i^{EBP} \equiv \hat{\theta}_i(\hat{\boldsymbol{\phi}}_i)$.

Estimation of $\boldsymbol{\beta}_i$ when variance components are known

Set $\beta_0 = \sum_{i=1}^{m} \alpha_{0i}/m$.
For $t = 1, 2, \ldots,$ define

$$\mathbf{r}_{l;i}^{(t)} = \mathbf{A}_{l;i}(\mathbf{y}_l - \alpha_{0i}^{(t)}\mathbf{1}_{n_l} - \mathbf{X}_l\boldsymbol{\beta}_i^{(t)}),$$

where

- $\mathbf{y}_l$ is a vector of the response variable for area l;
- $\mathbf{X}_l$ denotes a matrix of individual level covariates in area l;
- $\mathbf{A}_{l;i}$ is a suitable known scale matrix.

Obtain $(\alpha_{0i}^{(t)}, \boldsymbol{\beta}_i^{(t)})$ by solving the following system of estimating equations for $(\alpha_{0i}, \boldsymbol{\beta}_i)$:

$$\sum_{l=1}^{m} \mathbf{W}_{l;i}\, \psi_i(\mathbf{r}_{l;i}^{(t)}) = \mathbf{0}, \quad i = 1, \ldots, m.$$

where $\mathbf{W}_{l;i}$ is a suitable known weight matrix.

Choices of $\psi_i(\mathbf{r}_{l;i}^{(t)})$

- $\psi_i(\mathbf{r}_{l;i}^{(t)})$ is a $n_l \times 1$ vector obtained from the vector of residuals $\mathbf{r}_{l;i}^{(t)}$ with its jth component, say $r_{lj;i}^{(t)}$, replaced by $\psi_i(r_{lj;i}^{(t)})$, a chosen known function of $r_{lj;i}^{(t)}$;
- $\psi_i(r) = 2\psi(r)\left[\tau_i I(r > 0) + (1 - \tau_i) I(r \leq 0)\right], \; -\infty < r < \infty$, where $\psi(r)$ is a known monotone non-decreasing function with $\psi(-\infty) < \psi(0) < \psi(\infty)$, $\tau_i \in \Omega = (0, 1)$ known.
- Examples of $\psi(r)$: $\psi(r) = r$ and Huber influence function.

A parametric bootstrap estimator of $f\left(E[d(\hat{\theta}_i, \theta_i)]\right)$

Step 1 Given $\boldsymbol{\phi}_i$, generate R parametric bootstrap replicates $\{y_{ij}^{(r)}, i = 1, \ldots, m; j = 1, \ldots, n_i, r = 1, \ldots, R\}$ using the following model:

$$y_{ij}^{(r)} = \hat{\beta}_0 + \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_i + \gamma_i^{(r)} + \epsilon_{ij}^{(r)},$$

where $\gamma_i^{(r)} \sim N(0, \hat{\sigma}_\gamma^2)$ and $\epsilon_{ij}^{(r)} \sim N(0, \hat{\sigma}_{\epsilon i}^2)$ are all independently distributed, $i = 1, \ldots, m;\ j = 1, \ldots, n_i$.

Step 2 For each replication r, compute the simulated parameter of interest: $\theta_i^{(r)} = \hat{\beta}_0 + \bar{\mathbf{X}}'_i\hat{\boldsymbol{\beta}}_i + \gamma_i^{(r)},\ r = 1, \ldots, R.$

Step 3 For each replication r, compute $\hat{\boldsymbol{\phi}}_i^{(r)}$ using the estimation algorithm and compute $\hat{\theta}_i^{(r)}$, which may depend on $\hat{\boldsymbol{\phi}}_i^{(r)}\ \ r = 1, \ldots, R.$

Step 4 A parametric bootstrap estimator of $f\left(E[d(\hat{\theta}_i, \theta_i)]\right)$ is:

$$f\left(\mathsf{E}_*[d(\hat{\theta}_i^*, \theta_i^*)]\right) \approx f\left(\frac{1}{R}\sum_{r=1}^{R} d(\hat{\theta}_i^{(r)}, \theta_i^{(r)})\right),$$

where E_* is the expectation with respect to the parametric bootstrap distribution.

EMAP Lake Survey Data Analysis

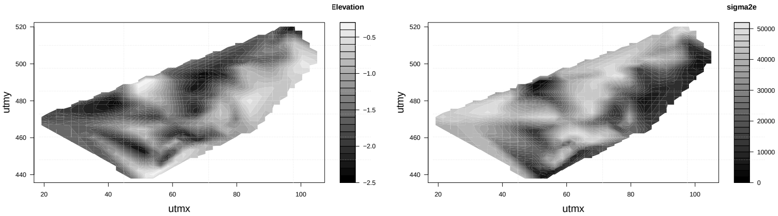
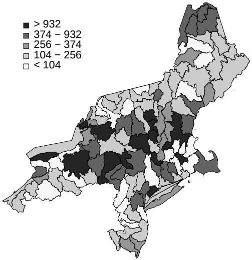


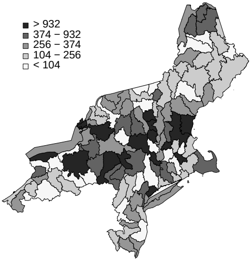
Figure: Maps showing the spatial variation in the HUC-specific area elevation slope coefficient (left) and sampling variance (right) estimates that are generated when the proposed nested error regression model with high dimensional parameter is fitted to the EMAP data.

Maps of estimated average ANC for HUCs using direct and EBP under NERHDP

Direct Estimates



EBP



Boxplot of CVs ratios

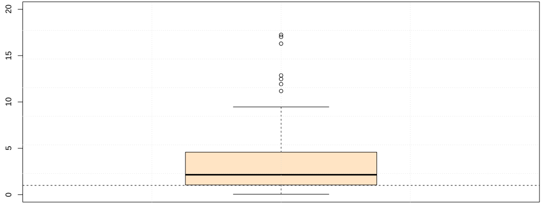


Figure: Boxplot showing the ratio between the CVs of the direct estimates and the CVs of the estimates obtained by the nested error regression model with high dimensional parameter. Values greater than 1 indicates that the CVs of the direct estimates are higher than the other ones.

Concluding Remarks

- Flexible modeling
- Area specific estimating equation
- Design consistency
- Starightforward parametric bootstrap for measuring uncertainty
- Method is extendable to estimate nonlinear finite population parameters.

References I

Arora, V., Lahiri, P. and Mukherjee, K. (1997) Empirical bayes estimation of finite population means from complex survey. *Journal of the American Statistical Association*, **92**, 1555–1562.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Breckling, J. and Chambers, R. (1988) M-quantiles. *Biometrika*, **75 (4)**, 761–771.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014) Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, **76 (1)**, 47–69.

Chambers, R. and Tzavidis, N. (2006) M-quantile models for small area estimation. *Biometrika*, **93 (2)**, 255–268.

Datta, G. S. and Lahiri, P. (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, **10**, 613–627.

References II

Ghosh, M. and Meeden, G. (1997) *Bayesian Methods for Finite Population Sampling*. London: Chapman & Hall.

Hall, P. and Maiti, T. (2006) On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, **68 (2)**, 221–238.

Jiang, J., Lahiri, P. and Nguyen, T. (2018) A unified monte-carlo jackknife for small area estimation after model selection. *Annals of Mathematical Sciences and Applications*, **3 (2)**, 405–438.

Jiang, J., Lahiri, P. and Wan, S.-M. (2002) A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics*, **30**, 1782–1810.

Jiang, J. and Nguyen, T. (2012) Small area estimation via heteroscedastic nested-error regression. *Canadian Journal of Statistics*, **40**, 588–603.

Jiang, J., Nguyen, T. and Rao, J. (2011) Best predictive small area estimation. *Journal of the American Statistical Association*, **106**, 732–745.

References III

Kubokawa, T., Sugasawa, S., Ghosh, M. and Chaudhuri, S. (2016) Prediction in heteroscedastic nested error regression models with random dispersions. *Statistica Sinica*, **26**, 465–492.

Lahiri, P. and Salvati, N. (2023). A nested error regression model with high-dimensional parameter for small area estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **85**, 212-213. https://doi.org/10.1093/jrsssb/qkac010

Opsomer, J., Claeskens, G., Ranalli, M., Kauermann, G. and Breidt, F. (2008) Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B*, **70**, 265–283.

Rao, J. N. K. and Molina, I. (2015) *Small Area Estimation*. New York: Wiley, 2nd edition edn.

Sugasawa, S. and Kubokawa, T. (2017) Heteroscedastic nested error regression models with variance functions. *Statistica Sinica*, **27**, 1101–1123.

SAE 2024: Small Area Estimation, Surveys, and Data Science, Lima, Peru, June 3-7, 2024

Contact Information

Partha Lahiri
Professor and Director, Joint Program in Survey Methodology
& Professor, Department of Mathematics
1218 Lefrak Hall
University of Maryland
College Park, MD 20742
Email: plahiri@umd.edu

Thank You!