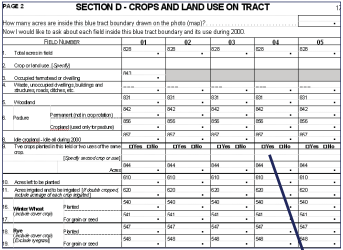
Can Geospatial Auxiliary Variables be Beneficial in Small Area Estimation? If so, How Can They be Effectively Utilized?

Partha Lahiri

University of Maryland College Park, USA

Based on my joint research with my collaborators Priyanka Anjoy, Yuting Chen, Santanu Pramanik, Nicola Salvati, Jiraphan Suntornchost



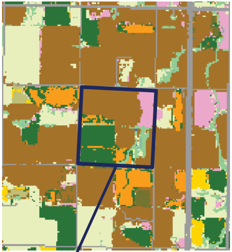
PAGE 2

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
	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

Zakzeski, A., National Agricultural Statistics Service

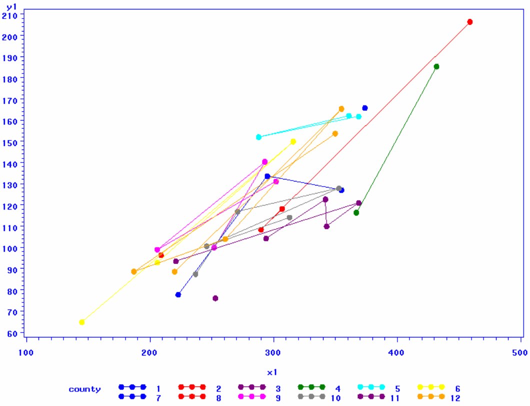
An Example

Ref: Battese, Harter and Fuller (1988 JASA)

Table 1. Survey and Satellite Data for Corn and Soybeans in 12 Iowa Counties

	No. of segments		Reported hectares		No. of pixels in sample segments		Mean number of pixels per segment*	
County	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	189.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humboldt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	29.46	340	87		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

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



This plot also reflects the strong relationship between the

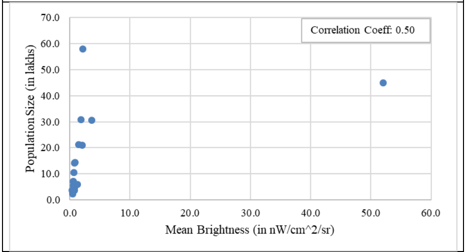
Anjoy, Lahiri, Pramanik (2025)

Correlation between Population Size vs. Nighttime Light Data for 2012

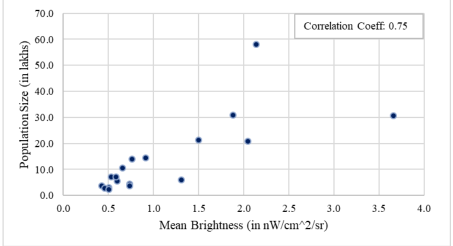
States	Rural	Urban	Aggregate
Uttar Pradesh	-0.49	0.82	0.27
Maharashtra	-0.28	0.48	0.26
Madhya Pradesh	-0.27	0.90	0.59
Gujarat	-0.32	0.56	0.50
Tamil Nadu	-0.51	0.79	0.58
West Bengal	-0.14	0.75	0.56
Odisha	-0.05	0.32	0.37
Assam	-0.35	0.71	0.20

Correlation between Estimated Poverty Rates vs. Nighttime Light Data for 2022

States	Rural	Urban	Aggregate
Uttar Pradesh	0.003	-0.40	0.03
Maharashtra	-0.29	-0.38	-0.30
Madhya Pradesh	-0.14	-0.61	-0.35
Gujarat	-0.30	-0.38	-0.47
Tamil Nadu	-0.06	-0.03	-0.14
West Bengal	-0.28	-0.33	-0.25
Odisha	-0.48	-0.86	-0.64



Including Kolkata



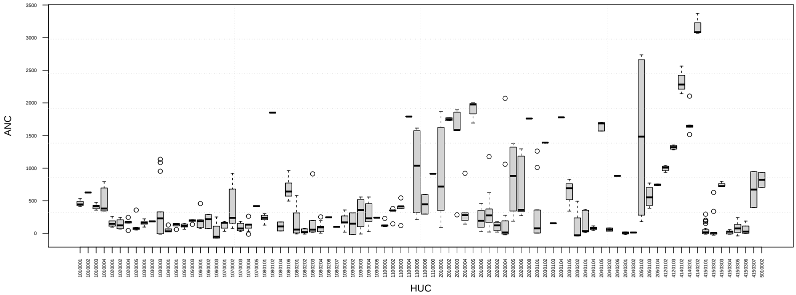
Excluding Kolkata

An Example from the EMAP Lake Survey Data

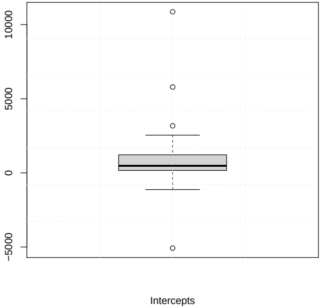
Ref: Lahiri and Salvati (JRSS B, 2024)

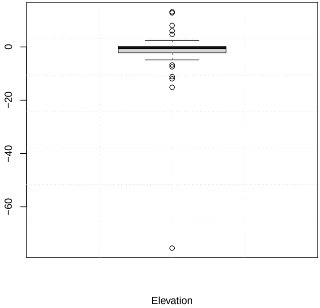
- 334 lakes selected from the population of 21,026 lakes
- 86 Hydrologic Unit Codes (HUCs) are in-sample
- 27 HUCs are out-of-sample
- Estimation of average Acid Neutralising Capacity (ANC) by HUC is of interest.

An Example from the EMAP Lake Survey Data (Cont'd)



An Example from the EMAP Lake Survey Data (Cont'd)





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 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 model parameters σ_γ^2 and σ_ϵ^2 are referred to as the variance components.

An extension of NER model

We propose the following extension of NER model:

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

- β_0 is a common intercept term;
- $\boldsymbol{\beta}_i$ is a $p \times 1$ vector of fixed 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 $\bar{Y}_i \approx \theta_i = \beta_0 + \bar{\mathbf{X}}_i' \boldsymbol{\beta}_i + \gamma_i$ is given by

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

where

- $\bar{\mathbf{X}}_i$: 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 $\hat{\theta}_i^{EBP} \equiv \hat{\theta}_i(\hat{\boldsymbol{\phi}}_i)$.

EMAP Lake Survey Data Analysis

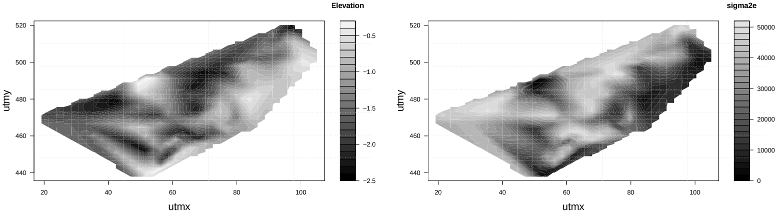
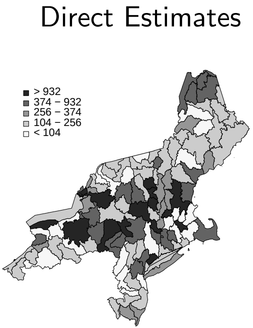


Figure: Maps showing the spatial variation in the HUC-specific area elevation slope coefficient (left) and sampling variance (right) estimates that are generated when the proposed nested error regression model with high dimensional parameter is fitted to the EMAP data.

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



Boxplot of CVs ratios

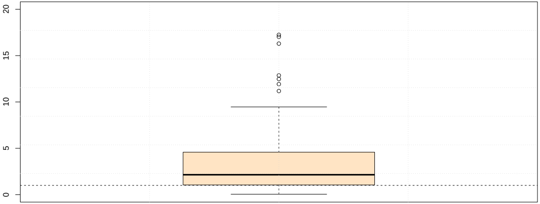


Figure: Boxplot showing the ratio between the CVs of the direct estimates and the CVs of the estimates obtained by the nested error regression model with high dimensional parameter. Values greater than 1 indicates that the CVs of the direct estimates are higher than the other ones.

R package: NERHD

The R package is at:

`https://github.com/nicolasalvati73/NERHD/blob/main/NERHD_0.1.1.tar.gz`

Figure:

Thank You!