

Integrating survey data with alternative databases for small area estimation

Partha Lahiri

University of Maryland College Park, USA

The IV Congress of Polish Statistics, Warsaw

3 July, 2024

Joint Program in Survey Methodology (JPSM)



(a) JPSM Onsite programs



(b) JPSM Online programs



(c) JPSM Short Courses

UN Sustainable Development Goals (SDG)



UN-SDG WEB Banner



The 17 Sustainable Development Goals (SDG)

- GOAL 1: No Poverty
- GOAL 2: Zero Hunger
- GOAL 3: Good Health and Well-being
- GOAL 4: Quality Education
- GOAL 5: Gender Equality
- GOAL 6: Clean Water and Sanitation
- GOAL 7: Affordable and Clean Energy
- GOAL 8: Decent Work and Economic Growth
- GOAL 9: Industry, Innovation and Infrastructure
- GOAL 10: Reduced Inequality
- GOAL 11: Sustainable Cities and Communities
- GOAL 12: Responsible Consumption and Production
- GOAL 13: Climate Action
- GOAL 14: Life Below Water
- GOAL 15: Life on Land
- GOAL 16: Peace and Justice Strong Institutions
- GOAL 17: Partnerships to achieve the Goal

Sample surveys

- Probability sample surveys

- Nonprobability sample surveys



Merfeld, J., Chen, H., Lahiri, P., and Newhouse, D. (2024). Small Area Estimation with Geospatial Data: A Primer (draft). Inter-Secretariat Working Group on Household Surveys, Background document to the 55th session of the United Nations Statistical Commission.



Uses of Geo-spatial Datasets by WFP

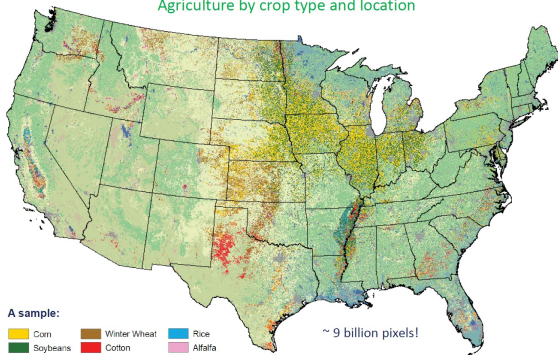
- Building footprint data from Google Open Buildings dataset
- Road density data from Open Street Map
- Nighttime lights
- Vegetation Index

Ref: [WFP Global Data Strategy 2024 – 2026](#)

Satellite Data

Cropland Data Layer

Agriculture by crop type and location



Zakzeski, A., National Agricultural Statistics Service



Retail Scanner Data (Nielsen)

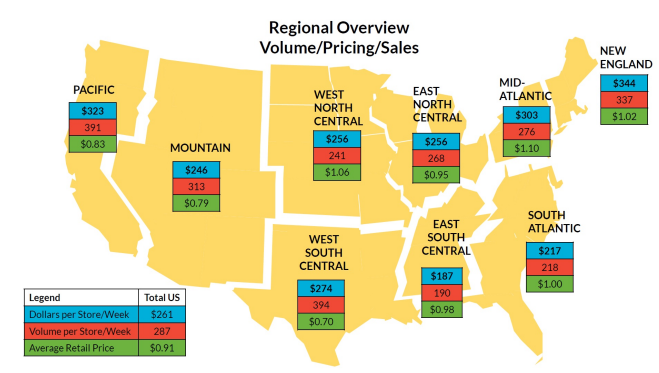


Figure: Scanner data of mango sales in grocery stores over different geographical regions

GPS Probe Data Collection

The following figure (from FHWA, 1998) summarizes the collection of probe data

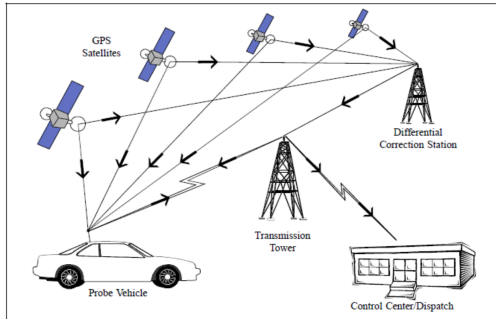
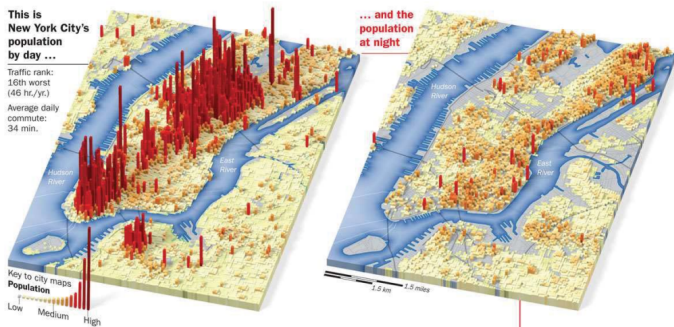


Figure: GPS Data Collection (FHWA, 1998; Source Kartika, C.S.D., 2015)

Cell Phone Data

Location data from mobile phones



Source: Pfeffermann (2017)

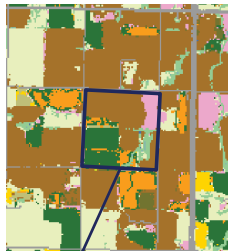


PAGE 2 SECTION D - CROPS AND LAND USE ON TRACT

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

How 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
Total acreage total					
2. Crop or land use (Specify)					
3. Occupied (rented or owned)					
4. Value, unoccupied (sheds, buildings and structures, roads, ditches, etc.)					
5. Woodland					
6. Pasture Permanent (not in crop rotation) Crotched (used only for pasture)					
7. Not cropland - No. of days 2000					
8. The crop planted in the field or two uses of the same crop. (Specify acreage or use)	Crop Code	Crop Code	Crop Code	Crop Code	Crop Code
9. Acre left to be planted					
10. Acre irrigated and to be irrigated (if double-crop or double-cropland or double-crop irrigated)					
Winter Wheat (include cover crop)	Planted For grain or seed				
17. Rice (include cover crop) (include prepared)	Planted For grain or seed				



REGRESSION
VARIABLES:

Dependent
Y

Independent
X



	Enumerated JAS Segments	CDL Classified Acres
Soybeans	227	273
Wheat	337	541



28

Zakzeski, A., National Agricultural Statistics Service

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 j th unit of the i th small area, respectively, with $i = 1, \dots, m, j = 1, \dots, N_i$;
- Parameter of interest: $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}, i = 1, \dots, m.$
- 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 i th small area.

Nested error regression model (NER)

- Nested error regression model for the finite population:

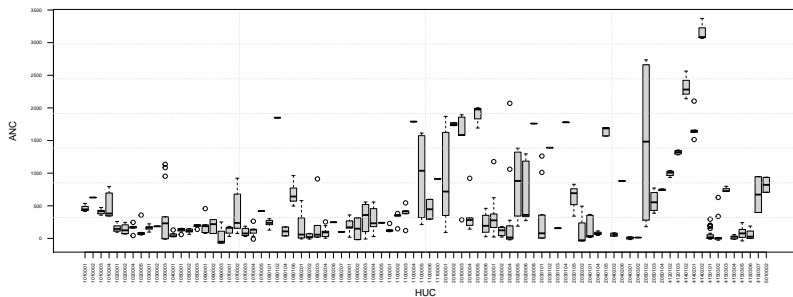
$$y_{ij} = \beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i + \epsilon_{ij}, \quad i = 1, \dots, m; \quad j = 1, \dots, N_i,$$

- 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 j th observation in the i th area; γ_i and ϵ_{ij} are all assumed to be independent with $\gamma_i \sim N(0, \sigma_\gamma^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $i = 1, \dots, m; j = 1, \dots, N_i$;
- The model parameters σ_γ^2 and σ_ϵ^2 are referred to as the variance components.

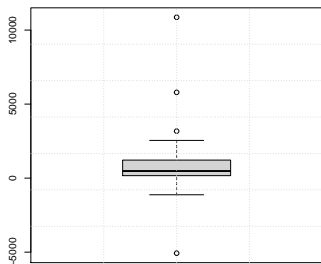
An Example from the EMAP Lake Survey Data

- 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.

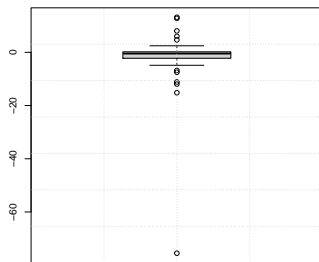
An Example from the EMAP Lake Survey Data (Cont'd)



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



Intercepts



Elevation

An extension of NER model

We propose the following extension of NER model:

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

- 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 $\gamma_i \sim N(0, \sigma_\gamma^2)$ and $\epsilon_{ij} \sim N(0, \sigma_{\epsilon i}^2)$.

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

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

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 $\hat{\theta}_i^{EBP} \equiv \hat{\theta}_i(\hat{\boldsymbol{\phi}}_i)$.

Data-driven method for model parameter estimation

- For estimating the model parameters ϕ_i , generalized estimating equations (GEE) with area specific tuning parameters are used to improve prediction accuracy.
- Method allows to borrow strength across areas when estimating each area specific vector of parameters.
- For known area specific tuning parameters, our estimating equation method yields consistent estimators of the model parameters.

Measures of uncertainty of EBP

- Parametric bootstrap
- First-order unbiased when tuning parameters are known.
- We deviate from the standard second-order unbiasedness property of mean squared error (MSE) estimators.
- Method can estimate various uncertainty measures (e.g., MSE, RRMSE, CV, etc.)

EMAP Lake Survey Data Analysis

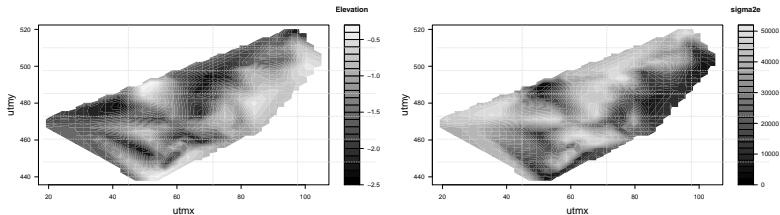
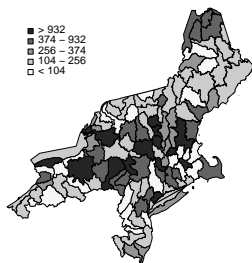


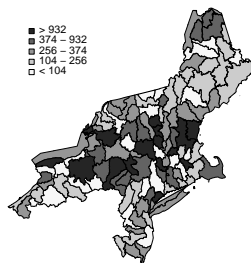
Figure: Maps showing the spatial variation in the HUC-specific area elevation slope coefficient (left) and sampling variance (right) estimates that are generated when the proposed nested error regression model with high dimensional parameter is fitted to the EMAP data.

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

Direct Estimates



EBP



Boxplot of CVs ratios

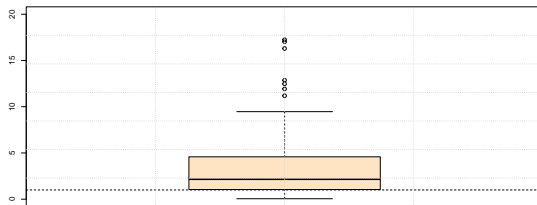


Figure: Boxplot showing the ratio between the CVs of the direct estimates and the CVs of the estimates obtained by the nested error regression model with high dimensional parameter. Values greater than 1 indicates that the CVs of the direct estimates are higher than the other ones.

R package: NERHD

The R package is at:

```
https://github.com/nicolasalvati73/NERHD/blob/main/  
NERHD\_0.1.1.tar.gz
```









Figure







Concluding Remarks

- Flexible modeling
- Area specific estimating equation
- Design consistency
- Straightforward 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/jrsssrb/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.

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!