A Nested Error Regression Model with High Dimensional Parameter for Small Area Estimation

Partha Lahiri

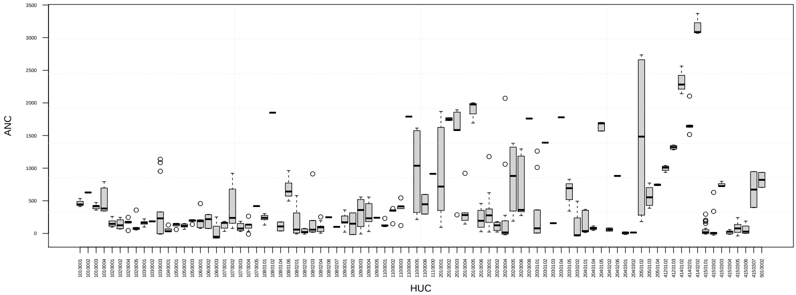
University of Maryland College Park, USA

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

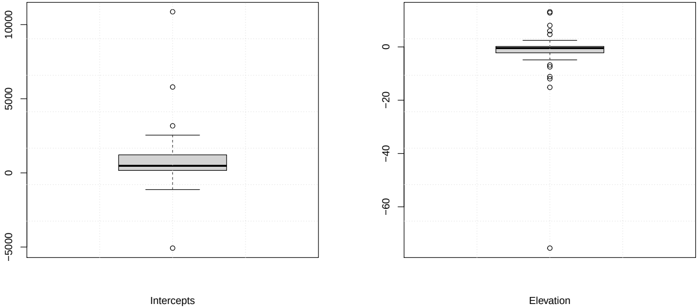
An Example from the EMAP Lake Survey Data

- 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)



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:

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

- 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

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

where

- $\bar{\mathbf{X}}_i$: population mean for area i
- $\bar{\mathbf{x}}_i$: sample mean for area i
- $B_i = \frac{\sigma^2_{\epsilon i}/n_i}{\sigma^2_{\epsilon i}/n_i + \sigma^2_\gamma}$;
- $\boldsymbol{\phi}_i = (\beta_0, \boldsymbol{\beta}_i, \sigma^2_\gamma, \sigma^2_{\epsilon i})'$;
- 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

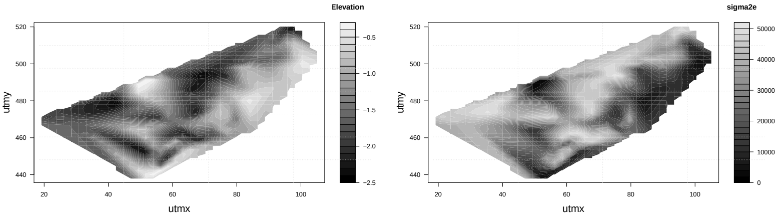
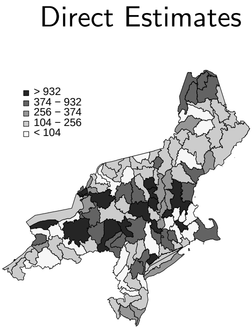
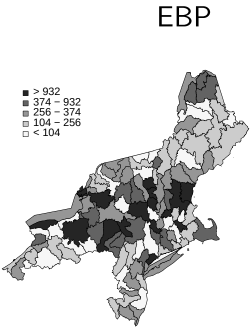


Figure: Maps showing the spatial variation in the HUC-specific area elevation slope coefficient (left) and sampling variance (right) estimates that are generated when the proposed nested error regression model with high dimensional parameter is fitted to the EMAP data.

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





Boxplot of CVs ratios

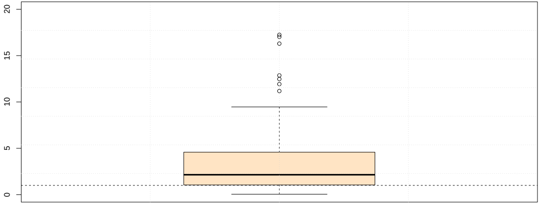


Figure: Boxplot showing the ratio between the CVs of the direct estimates and the CVs of the estimates obtained by the nested error regression model with high dimensional parameter. Values greater than 1 indicates that the CVs of the direct estimates are higher than the other ones.

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
- 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.

Thank You!