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

Nicola Salvati

University of Pisa

Joint work with
Partha Lahiri (University of Maryland)

IASS Webinar 45

Hukum Chandra Memorial Prize

30.10.2024
A memory of Hukum Chandra

- I met Hukum for the first time at SAE2009 in Elche. We started to collaborate even earlier, thanks to Ray Chambers when he was Hukum's PhD supervisor.

- Hukum collaborated with Pisa SAE Team from 2009 and he visited Pisa for three months in 2014.

- Made a significant contribution in survey sampling, small area estimation, official statistics, spatial analysis applied to official and survey statistics and agricultural statistics.
Outline

1. Background
   - Notation
   - A motivating example: EMAP estimates of ANC at HUC level

2. Nested Error Regression Model with High Dimensional Parameter
   - An extension of the nested error regression model
   - The estimation algorithm
   - Uncertainty measures

3. EBP estimates of the ecological condition of lakes in the northeastern USA

4. Model-based simulation experiments

5. Conclusions and future developments

A Nested Error Regression Model with High Dimensional Parameter for Small Area Estimation
Battese et al. (1988) proposed an empirical best linear unbiased prediction (EBLUP) method, using a nested error regression (NER) model, in order to estimate acreage under corn and soybeans for 12 counties in north-central Iowa, USA.

Such a model is likely to poorly fit the data when the number of small areas of interest is large; Jiang and Lahiri (2006); Morris (2006).

The assumption of the same regression coefficients and/or variance components in the NER model may not be tenable for all the small areas (Arora et al. 1997; Salvati et al., 2012).
Notation

- $m$ small areas with $N_i$ units;
- $y_{ij}$ and $x_{ij}$ 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 $i = 1, \ldots, m$, $j = 1, \ldots, N_i$;
- prediction of small area mean:
  $$\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}, \quad i = 1, \ldots, m$$
  using a simple random sample $s$ of size $n$ drawn from the finite population covering all $m$ areas and $\bar{X}_i = N_i^{-1} \sum_{j=1}^{N_i} x_{ij}$, the vector $p \times 1$ of finite population means of the auxiliary variables for area $i$;
- $n_i$ 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. Non-informative sampling.
Nested error regression model (Battese et al., 1988)

- Nested error regression model for the finite population:
  \[ y_{ij} = \beta_0 + x_{ij}'\beta + \gamma_i + \epsilon_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, N_i, \]  
  (1)

- \( \beta_0 \) and \( \beta \) are unknown fixed intercept and regression coefficients, respectively;

- \( \gamma_i \) is a random effect for area \( i \); \( \epsilon_{ij} \) is the sampling error for the \( j \)th observation in the \( i \)th area; \( \gamma_i \) and \( \epsilon_{ij} \) are all assumed to be independent with \( \gamma_i \sim N(0, \sigma^{2}_{\gamma}) \) and \( \epsilon_{ij} \sim N(0, \sigma^{2}_{\epsilon}) \), \( i = 1, \ldots, m; \quad j = 1, \ldots, N_i; \)

- the parameters \( \delta = (\sigma^{2}_{\gamma}, \sigma^{2}_{\epsilon}) \) are referred to as the variance components of model (1).
Under model (1), the finite population mean $\bar{Y}_i$ can be well approximated by $\theta_i = \beta_0 + \bar{X}_i'\hat{\beta} + \gamma_i$, for large $N_i$. Battese et al. (1988) proposed an Empirical Best Linear Unbiased Predictor (EBLUP) for estimating $\theta_i$ given by

$$\hat{\theta}_i^{BHF} = \hat{\beta}_0 + \bar{X}_i'\hat{\beta} + \hat{\gamma}_i,$$  (2)

- $\hat{\beta}_0$ and $\hat{\beta}$ are weighted least square estimators of $\beta_0$ and $\beta$;
- $\hat{\gamma}_i = (1 - \hat{B}_i)(\bar{y}_i - \hat{\beta}_0 - \bar{x}_i'\hat{\beta})$ is the EBLUP of $\gamma_i$ with $\bar{x}_i = n_i^{-1}\sum_{j \in s_i} x_{ij}$;
- $\hat{B}_i = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_e^2/n_i + \hat{\sigma}_\gamma^2}$ and $\hat{\delta} = (\hat{\sigma}_\gamma^2, \hat{\sigma}_e^2)$ is a consistent estimator of $\delta = (\sigma_\gamma^2, \sigma_e^2)$ under model (1).
A motivating example: EMAP estimates of ANC at HUC level

A motivating example: the distribution of individual residuals

- 334 lakes selected from the population of 21,026 lakes
- 86 HUCs are in-sample
- 27 HUCs are out-of-sample

N. Salvati
A Nested Error Regression Model with High Dimensional Parameter for Small Area Estimation
A motivating example: EMAP estimates of ANC at HUC level

A motivating example: the distribution of the estimated values of the intercept and the slope by area
The synthetic assumption of identical regression coefficients $\beta$ and sampling variance $\sigma^2_\epsilon$ across all small areas to be combined may be unrealistic when the number of small areas $m$ is large;

The synthetic assumption on $\beta$ can be relaxed if we replace $\beta$ in (1) by random area specific regression $\beta_i, i = 1, \ldots, m$, generated from a common model. This additional assumption for random effects is necessary to reduce the number of unknown parameters;

Rao and Molina (2015) relaxed the homogeneity assumption of the individual errors by replacing $\sigma^2_\epsilon$ by $k_{ij}\sigma^2_\epsilon$, where $k_{ij} > 0$ is a known auxiliary variable. However, identifying $k_{ij}$ in a real-life data analysis may be hard and between area variability may not be fully explained by $k_{ij}$.
A motivating example: EMAP estimates of ANC at HUC level

**Alternative models**

- Random area specific regression coefficients models (Prasad and Rao, 1990; Hobza and Morales, 2013; Rao and Molina, 2015).

- Heteroscedastic nested error regression model by allowing different fixed sampling variances (Jiang and Nguyen, 2012; Maiti et al. 2014; Kubokawa et al., 2016; Sugasawa et al. 2017; Naves et al. 2020, Das and Chambers, 2024).

**BUT**

- Random area specific regression coefficients models involve specifications of distributions of a large number of random effects.

- Fixed effects assumptions on the area specific regression coefficients and sampling variances generally lead to unstable estimates of these fixed effects due to small area specific sample sizes.
A motivating example: EMAP estimates of ANC at HUC level

The proposed approach

The Nested Error Regression with High Dimensional Parameter (NERHDM):

- Fixed effects for both regression coefficients and sampling variances.
- Area specific estimating equations applied to data from all areas for estimating these area specific regression coefficients.
- Appropriately constructed residuals for estimation of variance components.

N. Salvati
A Nested Error Regression Model with High Dimensional Parameter for Small Area Estimation
Outline

1 Background

2 Nested Error Regression Model with High Dimensional Parameter
   - An extension of the nested error regression model
   - The estimation algorithm
   - Uncertainty measures

3 EBP estimates of the ecological condition of lakes in the northeastern USA

4 Model-based simulation experiments

N. Salvati
A Nested Error Regression Model with High Dimensional Parameter for Small Area Estimation
The model

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

\[ y_{ij} = \beta_0 + x'_{ij}\beta_i + \gamma_i + \epsilon_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, N_i, \quad (3) \]

- \( \beta_i \) is a \( p \times 1 \) vector of fixed unknown regression coefficients for area \( i \);
- \( \gamma_i \) and \( \epsilon_{ij} \) are all independent with \( \gamma_i \sim N(0, h_i\sigma^2_i) \) and \( \epsilon_{ij} \sim N(0, k_{ij}\sigma^2_{\epsilon_i}) \), with \( h_i \) and \( k_{ij} \) are known auxiliary variables at area and individual levels, respectively;
- Battese et al. (1988) is a special case of model (3) with \( h_i = 1, \quad k_{ij} = 1, \quad \beta_i = \beta \) and \( \sigma^2_{\epsilon_i} = \sigma^2_{\epsilon} \); \( i = 1, \ldots, m; \quad j = 1, \ldots, N_i \).
The Best Predictor (BP)

Under model (3) the best predictor (BP) of \( \theta_i = \beta_0 + \bar{X}_i' \beta_i + \gamma_i \) is given by

\[
\hat{\theta}_i^{BP} \equiv \hat{\theta}_i(\phi_i) = \beta_0 + \bar{X}_i' \beta_i + (1 - B_i)(\bar{y}_i - \beta_0 - \bar{x}_i' \beta_i) \\
= (\bar{X}_i - \bar{x}_i)' \beta_i + \{ B_i(\beta_0 + \bar{x}_i' \beta_i) + (1 - B_i)\bar{y}_i \}
\]

- \( B_i = \frac{\sigma^2_{\varepsilon_i}/n_i}{\sigma^2_{\varepsilon_i}/n_i + \sigma^2_{\gamma_i}}; \)
- \( \phi_i = (\beta_0, \beta_i, \sigma^2_{\gamma_i}, \sigma^2_{\varepsilon_i})'; \)
- An empirical best predictor (EBP) of \( \theta_i \) can be written as \( \hat{\theta}_i^{EBP} \equiv \hat{\theta}_i(\hat{\phi}_i) \), where \( \hat{\phi}_i \) is a consistent estimator of \( \phi_i \) under the assumed model (3) as \( m \) tends to \( \infty \).
An extension of the nested error regression model

**Estimation of the vector of the parameters**

\[ \phi_i = (\beta_0, \beta_i, \sigma_\gamma^2, \sigma_\epsilon^2)' \]

For model (3), the log-likelihood function has the expression

\[
\ell(\phi) = \text{constant} - \frac{1}{2} \sum_{i=1}^{m} \left[ n_i \log \sigma_\epsilon^2 + \log \left( \frac{\sigma_\epsilon^2 + n_i \sigma_\gamma^2}{\sigma_\epsilon^2} \right) \right] \\
+ \frac{1}{\sigma_\epsilon^2} \left\{ \sum_{j=1}^{n_i} \left( y_{ij} - \beta_0 - x_{ij}' \beta_i \right)^2 - \frac{\sigma_\gamma^2}{\sigma_\epsilon^2 + n_i \sigma_\gamma^2} (n_i \bar{y}_i - n_i \beta_0 - n_i \bar{x}_i \beta_i)^2 \right\}, (4)
\]
An extension of the nested error regression model

**MLE estimators**

MLE of the regression coefficients for area $i$ are:

$$
\hat{\beta}_0 = \left( \sum_{i=1}^{m} \sigma_{\epsilon_i}^{-2} n_i B_i \right)^{-1} \left( \sum_{i=1}^{m} \sigma_{\epsilon_i}^{-2} n_i B_i (\bar{y}_i - \bar{x}_i' \beta_i) \right), \quad (5)
$$

$$
\hat{\beta}_i = \left( \sum_{j=1}^{n_i} x_{ij} x_{ij}' - n_i (1 - B_i) \bar{x}_i \bar{x}_i' \right)^{-1} \left( \sum_{j=1}^{n_i} x_{ij} (y_{ij} - \beta_0) - n_i (1 - B_i) \bar{x}_i (\bar{y}_i - \beta_0) \right), \quad (6)
$$

and the MLE of the sampling variance for area $i$ is:

$$
\hat{\sigma}_{\epsilon_i}^2 = \frac{1}{n_i} \left\{ \sum_{j=1}^{n_i} (y_{ij} - \beta_0 - x_{ij}' \beta_i)^2 - n_i (1 - B_i) (\bar{y}_i - \beta_0 - \bar{x}_i' \beta_i)^2 \right\}. \quad (7)
$$
Neyman and Scott (1948) gave an example which shows that, when the number of nuisance parameters increases with at the same rate as the sample size, the MLEs may not be consistent;

Jiang and Nguyen (2012) considered a heteroscedastic NER model with area-specific error variance, and noted that the MLE of the area-specific error variance is inconsistent.
### Background
- **NERHDP**

### Application
- Model-based simulation experiments

### Conclusions

The estimation algorithm

**New estimation method: Data-driven method for model parameter estimation**

- Generalized Estimating Equation (GEE) approach is used to estimate the parameters of model (3);
- this method allows to borrow strength across areas when estimating each area specific vector of parameters obtaining consistent estimators of the area specific slope parameters ($\beta_i$) and the area specific sampling variance ($\sigma_{ci}^2$);
- for known area specific tuning parameter $\tau_i$, our estimating equation method yields consistent estimators of the model parameters, unlike the maximum likelihood method.
The estimation algorithm

The algorithm

1. **Estimate** $\beta_i$: The regression coefficients $\beta_i$ are estimated.
2. **Estimate** $\sigma^2_{\epsilon_i}$: The residual variance $\sigma^2_{\epsilon_i}$ is estimated.
3. **Estimate** $\sigma^2_\gamma$: The variance of the random effect $\sigma^2_\gamma$ is estimated.
4. **Iterate**: Steps 1-3 are repeated until convergence. Convergence is achieved when the difference between the estimated parameters from two successive iterations is smaller than a predefined threshold.
The estimation algorithm

The algorithm (cont’d)

Step 1 For \( t = 1, 2, \ldots \) define \( r_{l;i}^{(t)} = (U_{l;i}^{(t-1)})^{-1/2}(y_l - \alpha_{0i}^{(t)}1_{n_l} - X_l\beta_i^{(t)}) \), where \( y_l \) is a \( n_l \times 1 \) vector of the response variable and \( X_l \) denotes a matrix \( n_l \times p \) of individual level covariates of the sampling units in area \( l \). Obtain \( (\alpha_{0i}^{(t)} , \beta_i^{(t)}) \) by solving the following system of estimating equations for \( (\alpha_{0i}^{(t)} , \beta_i^{(t)}) \):

\[
\sum_{l=1}^{m} x_l(p+1)(V_{l;i}^{(t-1)})^{-1}(U_{l;i}^{(t-1)})^{1/2} \psi_i(r_{l;i}^{(t)}) = 0, \quad i = 1, \ldots, m,
\]

where \( \psi_i(r_{l;i}^{(t)}) \) is a \( n_l \times 1 \) vector obtained from the vector of residuals \( r_{l;i}^{(t)} \) with its \( j \)th component, say \( r_{lj;i}^{(t)} \), replaced by \( \psi_i(r_{lj;i}^{(t)}) \), a chosen known function of \( r_{lj;i}^{(t)} \). Here \( X_l(p+1) \) denotes a matrix of dimension \( n_l \times (p+1) \) containing the covariates of the sampling units of area \( l \) including the intercept. The solution \( (\alpha_{0i}^{(t)} , \beta_i^{(t)}) \) for \( i = 1, \ldots, m \) can be obtained using an iteratively re-weighted least squares algorithm or the Newton-Raphson algorithm.
The estimation algorithm

Step 1: estimation of the regression coefficients $\beta_i$
The estimation algorithm

Step 1 (cont’d): estimation of the regression coefficients $\beta_i$ when regression coefficients and individual variance vary across the areas
The estimation algorithm

Step 2: estimation of the residual variance $\sigma^2_{\epsilon^i}$
Step 2 (cont’d): estimation of the residual variance $\sigma^2_{\varepsilon_i}$ when regression coefficients and individual variance vary across the areas
Proposition

Suppose that regularity conditions (i)-(viii) (Jiang et al., 2002; Chambers et al., 2014) given in Lahiri and Salvati (2023) hold and \( \psi_i \) functions are completely specified. Then the resulting vector of estimators \( \hat{\phi}_i \) is consistent.

Proposition

Under the model (3) and mild regularity conditions,

\[
RSL(\hat{\theta}_i^{EBP}, \tilde{\theta}_i) = \frac{MSE(\hat{\theta}_i^{EBP}) - MSE(\hat{\theta}_i^{BP})}{MSE(\tilde{\theta}_i) - MSE(\hat{\theta}_i^{BP})} \rightarrow 0, \quad \text{as } m \rightarrow \infty,
\]

where MSE is the mean squared error under model (3).
A data-driven method to estimate $\tau_i$

- $m = 4$ small areas. Each dot of the same color corresponds to a small area.
- For a fine grid of $\tau \in \Omega$ we obtain a collection of fitted values for the entire sample using the standard quantile or M-quantile methods.
- For each observation $y_{ij}$ in the sample, we find the fitted line with minimum prediction error defined as the difference between $y_{ij}$ and the predicted value by the fitted regression at $x_{ij}$. Let $\hat{\tau}_{ij}$ denote the value of $\tau$ in the grid for this best line.
An estimate of the area-\(i\)-specific M-quantile coefficient is the sample average of the estimated M-quantile coefficients for that area, \(\tilde{\tau}_i = n_i^{-1} \sum_{j \in s_i} \hat{\tau}_{ij}\) (Chambers and Tzavidis, 2006). Since \(n_i\) is typically small, \(\tilde{\tau}_i\) is likely to be unstable, we propose the following ELB:

(a) \(E[\hat{\tau}_{ij}|\tau_i] = \tau_i, V[\hat{\tau}_{ij}|\tau_i] = \nu^2\),
(b) \(E[\tau_i] = \mu, V[\tau_i] = \eta^2\).

Assuming \(E[\tau_i|\hat{\tau}] = \alpha + \beta \tilde{\tau}_i\), the linear best (LB) predictor of \(\tau_i\) can be written as

\[
\hat{\tau}_i^{LB} = (1 - B_i) \tilde{\tau}_i + B_i \mu,
\]

where \(B_i = \frac{\nu^2/n_i}{\nu^2/n_i + \eta^2}\). Substituting \(\mu\), \(\eta^2\) and \(\nu^2\) by their consistent estimators \(\hat{\mu}\), \(\hat{\eta}^2\) and \(\hat{\nu}^2\), we obtain the following empirical linear best (ELB) predictor (Ghosh and Lahiri, 1987):

\[
\hat{\tau}_i^{ELB} = (1 - \hat{B}_i) \tilde{\tau}_i + \hat{B}_i \hat{\mu},
\]

where \(\hat{B}_i = \frac{\hat{\nu}^2/n_i}{\hat{\nu}^2/n_i + \hat{\eta}^2}\).
The algorithm presented in previous slides could be computational intensive if the sample size $n$ is greater than 10,000 observations and if the number of small areas increases. For this reason we are working on a simplified algorithm that uses functions present in R and very rapid in execution. From the initial tests, it appears that the new algorithm is very fast with big sample size (over 100,000 units) and big number of small areas.
An estimator of RMSE for EBP

When \( \tau_i \) is known, an estimator of RMSE for EBP, under the assumed model (3), is given by

\[
\sqrt{g_{1i}(\hat{\phi}_i)},
\]

where

\[
g_{1i}(\hat{\phi}_i) = \frac{\sigma^2 \gamma \sigma^2_{\epsilon_i} / n_i}{\sigma^2 + \sigma^2_{\epsilon_i} / n_i}, \quad (11)
\]

and \( \hat{\phi}_i \) is a consistent estimator of \( \phi_i \), for large \( m \).

- Parametric bootstrap method
- McJack method (Jiang et al., 2018).
Outline

1. Background

2. Nested Error Regression Model with High Dimensional Parameter

3. EBP estimates of the ecological condition of lakes in the northeastern USA

4. Model-based simulation experiments

5. Conclusions and future developments
Estimates of slope coefficient and sampling variance

**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 EBLUP and EBP under NERHDP

N. Salvati
A Nested Error Regression Model with High Dimensional Parameter for Small Area Estimation
### Outline

1. Background
2. Nested Error Regression Model with High Dimensional Parameter
3. EBP estimates of the ecological condition of lakes in the northeastern USA
4. Model-based simulation experiments
5. Conclusions and future developments

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

---

**N. Salvati**
Simulation setting

- Model-based simulation study considering 1,000 replications
- $m = 100$ small areas
- $N_i = 100$ and $n_i = 4$, $i = 1, \ldots, m$
- $x \sim \text{LogN}(1.0, 0.5^2)$
- $y$ are generated using the following linear mixed model:

$$y_{ij} = 10 + \beta_i x_{ij} + \gamma_i + \epsilon_{ij} \quad i = 1, \ldots, 100; j = 1, \ldots, 4$$

- three different scenarios
Different scenarios

(i) \((0, 0)\): \(\beta_i = 5\) for all the small areas, \(\gamma_i \sim N(0, 3)\) and \(\epsilon_{ij} \sim N(0, 6)\) – this model is essentially the nested error regression model (1) under which EBLUP is developed;

(ii) \((\beta, 0)\): \(\beta_i = 5\) for \(i = 1, \ldots, 50\) and \(\beta_i = -5\) for \(i = 51, \ldots, 100\) and it is kept fixed over the simulations, \(\gamma_i \sim N(0, 3)\) and \(\epsilon_{ij} \sim N(0, 6)\) – this model violates assumptions of the NER model (1) because slopes vary across small areas;

(iii) \((\beta, \sigma^2)\): \(\beta_i = 5\) for \(i = 1, \ldots, 50\) and \(\beta_i = -5\) for \(i = 51, \ldots, 100\) and it is kept fixed over the simulations, \(\gamma_i \sim N(0, 3)\), \(\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon i})\), \(\sigma^2_{\epsilon i} \sim N(6, 2)\) for \(i = 1, \ldots, 50\) and \(\sigma^2_{\epsilon i} \sim N(12, 2)\) for \(i = 51, \ldots, 100\) – this model violates assumptions of NER model (1) because both slopes and sampling variances vary across small areas.
Results – Absolute Relative Bias – RRMSE

| Predictor  | Results (%) for the following scenarios | Median absolute relative bias | Median RRMSE |
|------------|-----------------------------------------|-----------------------------|---------------|
|            | (0, 0)                                   | (β, 0)                      | (β, \(\sigma^2\)) |
| Direct     | 0.535                                    | 0.927                       | 1.083         |
| EBLUP      | 0.132                                    | 9.129                       | 9.155         |
| MQ         | 0.127                                    | 5.070                       | 7.315         |
| EBLUP-RS   | 0.120                                    | 0.672                       | 0.783         |
| EBLUP-H    | 0.130                                    | 8.876                       | 8.903         |
| EBP        | 0.136                                    | 0.634                       | 0.671         |
| Direct     | 16.640 (17.887)                          | 44.259 (1.074)              | 45.770 (1.083) |
| EBLUP      | 3.922 (1.000)                            | 43.119 (1.000)              | 44.188 (1.000) |
| MQ         | 4.105 (1.103)                            | 14.774 (0.132)              | 20.101 (0.186) |
| EBLUP-RS   | 3.931 (1.006)                            | 13.991 (0.103)              | 18.283 (0.144) |
| EBLUP-H    | 3.924 (1.003)                            | 50.139 (1.359)              | 53.093 (1.291) |
| EBP        | 4.002 (1.047)                            | 12.065 (0.087)              | 15.596 (0.118) |

N. Salvati

A Nested Error Regression Model with High Dimensional Parameter for Small Area Estimation
# Results – performance of RMSE estimators

| RMSE estimator | Results (%) for scenarios | Median relative bias | Median RRMSE | Median coverage rate |
|----------------|---------------------------|----------------------|--------------|---------------------|
|                | (0, 0)                    | (β, 0)               | (β, σ²)      |                     |
| Naive          | -14.9                     | -23.9                | -25.2        |                     |
| Bootstrap      | -3.6                      | -8.1                 | -11.7        |                     |
| McJack         | -4.9                      | -4.8                 | -4.9         |                     |
| Naive          | 18.4                      | 30.2                 | 32.5         |                     |
| Bootstrap      | 14.5                      | 26.4                 | 28.7         |                     |
| McJack         | 14.1                      | 27.4                 | 26.8         |                     |

N. Salvati

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

1. Background
2. Nested Error Regression Model with High Dimensional Parameter
3. EBP estimates of the ecological condition of lakes in the northeastern USA
4. Model-based simulation experiments
5. Conclusions and future developments
Conclusions

- We have demonstrated, through simulations and data analysis, unsuitability of the well-known nested error regression model for small area estimation when combining a large number of small areas.
- We develop a robust area specific estimating equations approach where different estimating equations are used for different small areas.
- Performances of the estimators of area specific regression coefficients and sampling variances $\phi_i$ and the associated EBP depend very much on the area specific tuning parameters $\tau_i$ of the estimating equations.
- Design consistency
- Proposed methods perform well in our Monte Carlo
Correlated data within the same PSU.
Method is extendable to estimate nonlinear finite population parameters. EBP under NERHDM for poverty mapping.
Discrete data.
Theory when $\tau_i$ is estimated.
R package: NERHD

The R package is at:

https://github.com/nicolasalvati73/NERHD/blob/main/NERHD_0.1.1.tar.gz
Thank You!
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.

Das, S. and Chambers, R. (2024) Small area poverty estimation under heteroskedasticity *Journal of Survey Statistics and Methodology*, 12, 369–403.
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.

Ghosh, M. and Meeden, G. (1997) *Bayesian Methods for Finite Population Sampling From Stratified Sampling*. *Journal of the American Statistical Association*, 82, 1153–1162.

Ghosh, M. and Lahiri, P. (1987) *Robust Empirical Bayes Estimation of Means From*. 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.

N. Salvati

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

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.

Maiti, T., Ren, H., and Sinha, S. (2014) Prediction Error of Small Area Predictors Shrinking Both Means and Variances *Scandinavian Journal of Statistics, 41*, 775-790.
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.