Model-based Estimation of Poverty Indicators for Small Areas: Overview

J. N. K. Rao
Carleton University, Ottawa, Canada

Isabel Molina
Universidad Carlos III de Madrid, Spain

Paper presented at The First Asian ISI Satellite Meeting on Small Area Estimation (SAE), September 1-4, 2013, Bangkok, Thailand

NOTATION

- U **finite** population of size N.
- Population partitioned into D subsets $U_1, \ldots, U_D$ of sizes $N_1, \ldots, N_D$, called **domains** or **areas**.
- Variable of interest Y.
- Y_{dj} value of Y for unit j from domain d.
- **Target:** to estimate domain parameters.

$$\delta_d = h(Y_{d1}, \ldots, Y_{dN_d}), \quad d = 1, \ldots, D.$$

- We want to use data from a sample $S \subset U$ of size n drawn from the whole population.
- $S_d = S \cap U_d$ sub-sample from domain d of size $n_d = |S_d|$.
- **Problem:** n_d **too small** for some domains.

DIRECT ESTIMATORS

- **Direct estimator:** Estimator that uses only the sample data from the corresponding domain.
- Small **area/domain:** subset of the population that is target of inference and for which the direct estimator does not have enough precision.
- **Indirect estimator:** Borrows strength from other areas.

NESTED-ERROR REGRESSION MODEL

- **Model:** $\mathbf{x}_{dj}$ auxiliary variables at unit level,

$$Y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2).$$

- **EBLUP of** $\bar{Y}_d$: Predict non-sample values $\hat{Y}_{dj} = \mathbf{x}'_{dj}\hat{\boldsymbol{\beta}}_{WLS} + \hat{u}_d$,

$$\hat{\bar{Y}}_d^{EBLUP} = \frac{1}{N_d}\left(\sum_{j\in S_d} Y_{dj} + \sum_{j\in U_d - S_d} \hat{Y}_{dj}\right), \quad d = 1, \ldots, D.$$

✓ *Battese, Harter & Fuller (1988), JASA*

SOME POVERTY AND INCOME INEQUALITY INDICATORS

- FGT poverty indicators
- Quintile share
- Gini coefficient
- Sen index
- Theil index
- Generalized entropy
- Fuzzy monetary index

✓ *Neri, Ballini & Betti (2005), Stat. in Transition*

FGT POVERTY INDICATORS

- E_{dj} welfare measure for indiv. j in domain d: for instance, equivalised annual net income.
- z = poverty line.
- **FGT family of poverty indicators for domain d:**

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2.$$

When $\alpha = 0 \Rightarrow$ **Poverty incidence**

When $\alpha = 1 \Rightarrow$ **Poverty gap**

✓ *Foster, Greer & Thornbecke (1984), Econometrica*

FGT POVERTY INDICATORS

- **Complex non-linear** quantities (non continuous): Even if FGT poverty indicators are also means

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} F_{\alpha dj}, \quad F_{\alpha dj} = \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z),$$

we cannot assume normality for the $F_{\alpha dj}$.

SMALL AREA ESTIMATION

- Due to the relative nature of the mentioned poverty line, poverty has usually **low frequency**: Large sample size is needed.

  ✓ In Spain, poverty line for 2006: **6557 euros**, approx. **20 %** population under the line.

- Survey on Income and Living Conditions (EU-SILC) has limited sample size.

  ✓ In the Spanish SILC 2006, $n = 34{,}389$ out of $N = 43{,}162{,}384$ **(8 out 10,000)**.

SAMPLE SIZES OF PROVINCES BY GENDER

- Direct estimators for Spanish provinces are not very precise.
- We want estimates by Gender:
  Small areas: $D = 52$ provinces for each gender.
- CVs of direct and EB estimators of poverty incidences for selected provinces for each gender:

Province	Gender	n_d	Obs. Poor	CV Dir.	CV EB	CV HB
Soria	F	17	6	51.87	16.56	19.82
Tarragona	M	129	18	24.44	14.88	12.35
Córdoba	F	230	73	13.05	6.24	6.93
Badajoz	M	472	175	8.38	3.48	4.24
Barcelona	F	1483	191	9.38	6.51	4.52

EB METHOD FOR POVERTY ESTIMATION

- **Assumption:** there exists a transformation $Y_{dj} = T(E_{dj})$ of the welfare variables E_{dj} which follows a normal distribution (i.e., the nested error model with normal errors u_d and e_{dj}).
- FGT poverty indicator as a function of transformed variables:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left\{ \frac{z - T^{-1}(Y_{dj})}{z} \right\}^{\alpha} I\left\{ T^{-1}(Y_{dj}) < z \right\} = h_\alpha(\mathbf{y}_d),$$

where $\mathbf{y}_d = (Y_{d1}, \ldots, Y_{dN_d})' = (\mathbf{y}'_{ds}, \mathbf{y}'_{dr})'$.

- **EB estimator of $F_{\alpha d}$:**

$$\hat{F}_{\alpha d}^{EB} = E_{\mathbf{y}_{dr}}\left[F_{\alpha d} | \mathbf{y}_{ds} \right].$$

- MSE by parametric bootstrap for finite populations (✓ *González-Manteiga, Lombardía, Molina, Morales and Santamaría, 2008, J.Stat.Comp.Simul.*).

MONTE CARLO APPROXIMATION

(a) Generate L non-sample vectors $\mathbf{y}_{dr}^{(\ell)}$, $\ell = 1, \ldots, L$ from the (estimated) conditional distribution of $\mathbf{y}_{dr}|\mathbf{y}_{ds}$.

(b) Attach the sample elements to form a population vector $\mathbf{y}_{d}^{(\ell)} = (\mathbf{y}_{ds}, \mathbf{y}_{dr}^{(\ell)})$, $\ell = 1, \ldots, L$.

(c) Calculate the poverty measure with each population vector $F_{\alpha d}^{(\ell)} = h_\alpha(\mathbf{y}_{d}^{(\ell)})$, $\ell = 1, \ldots, L$. Then take the average over the L Monte Carlo generations:

$$\hat{F}_{\alpha d}^{EB} = E_{\mathbf{y}_{dr}}\left[F_{\alpha d}|\mathbf{y}_{ds}\right] \cong \frac{1}{L}\sum_{\ell=1}^{L} F_{\alpha d}^{(\ell)}.$$

MSE ESTIMATION

- Construct bootstrap populations $\{Y_{dj}^{*(b)}, b = 1, \ldots, B\}$ from

$$Y_{dj}^{*} = \mathbf{x}_{dj}'\hat{\beta} + u_d^{*} + e_{dj}^{*}; \quad j = 1, \ldots, N_d, \ d = 1, \ldots, D.$$
$$u_d^{*} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2); \quad e_{dj}^{*} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2).$$

- Calculate bootstrap population parameters $F_{\alpha d}^{*}(b)$

- From each bootstrap population, take the sample with the same indexes S as in the initial sample and calculate EBs $F_{\alpha d}^{EB*}(b)$ using bootstrap sample data $\mathbf{y}_s^{*}$ and known $\mathbf{x}_{dj}$.

$$mse^{*}(\hat{F}_{\alpha d}^{EB}) = \frac{1}{B}\sum_{b=1}^{B}\{\hat{F}_{\alpha d}^{EB*}(b) - F_{\alpha d}^{*}(b)\}^2$$

WORLD BANK (WB) / ELL METHOD

- Elbers et al. (2003) also used nested error model on transformed variables Y_{dj}, using clusters as d.
- For comparability we take cluster as small area.
- Generate A bootstrap populations $\{Y^*_{dj}(a), a = 1, \ldots, A\}$
- Calculate $F^*_{\alpha d}(a), a = 1, \ldots, A$. Then ELL estimator is:

$$\hat{F}^{(ELL)}_{\alpha d} = \frac{1}{A}\sum_{a=1}^{A} F^*_{\alpha d}(a) = F^*_{\alpha d}(\cdot)$$

WORLD BANK (WB) / ELL METHOD

- MSE estimator:

$$mse(\hat{F}_{\alpha d}^{ELL}) = \frac{1}{A}\sum_{a=1}^{A}\{F_{\alpha d}^{*}(a) - F_{\alpha d}^{*}(\cdot)\}^2$$

- If the mean $\bar{Y}_d$ is the parameter of interest, then

$$\hat{\bar{Y}}_d^{(ELL)} \simeq \bar{X}_d \hat{\boldsymbol{\beta}}$$

- $\hat{\bar{Y}}_d^{(ELL)}$ is a regression synthetic estimator.
- For non-sampled areas, $\hat{F}_{\alpha d}^{ELL}$ is essentially equivalent to $\hat{F}_{\alpha d}^{EB}$.
- But MSE estimators are different for ELL and EB.

POVERTY INCIDENCE

- Bias negligible for all three estimators (EB, direct and ELL).
- EB much more efficient than ELL and direct estimators.
- ELL even less efficient than direct estimators!

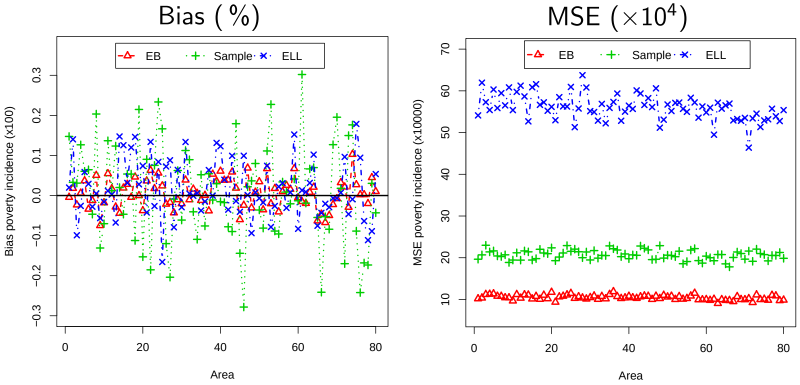


Figure 1. Bias (left) and MSE (right) of EB, direct and ELL estimators of poverty incidences F_{0d} for each area d.

POVERTY GAP

- Same conclusions as for poverty incidence.

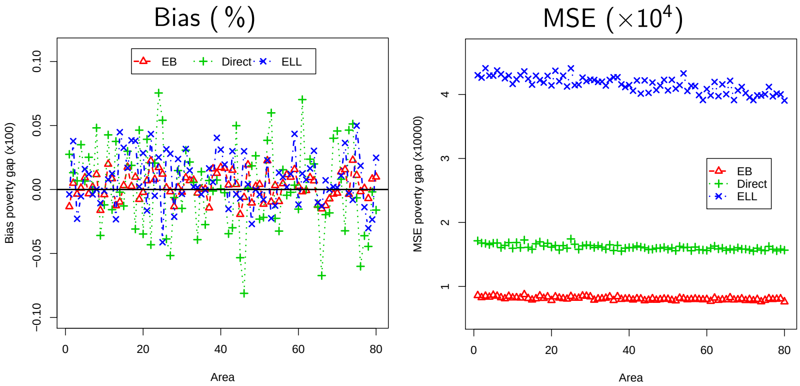


Figure 2. Bias (left) and MSE (right) of EB, direct and ELL estimators of poverty gaps F_{1d} for each area d.

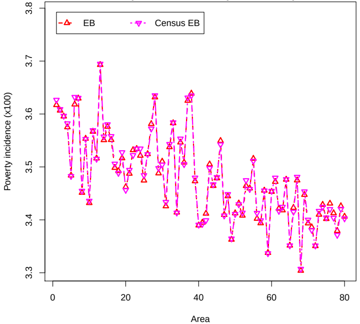
CENSUS EB METHOD

- When sample data cannot be linked with census auxiliary data, in steps (a) and (b) of EB method generate a full census from

$$\mathbf{y}_d = \hat{\mu}_{d|ds} + v_d \mathbf{1}_{N_d} + \epsilon_d, \quad \hat{\mu}_{d|ds} = \mathbf{X}_d \hat{\boldsymbol{\beta}} + \hat{\sigma}_u^2 \mathbf{1}_{N_d} \mathbf{1}'_{n_d} \hat{\mathbf{V}}_{ds}^{-1} (\mathbf{y}_{ds} - \mathbf{X}_{ds} \hat{\boldsymbol{\beta}}).$$

- Practically the same as original EB method.

a) Mean ($\times 100$)



b) MSE ($\times 10^4$)

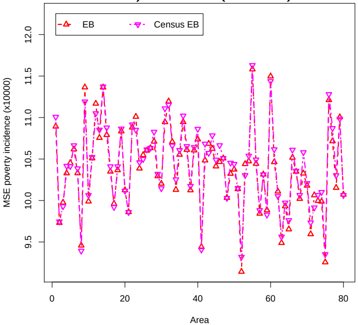
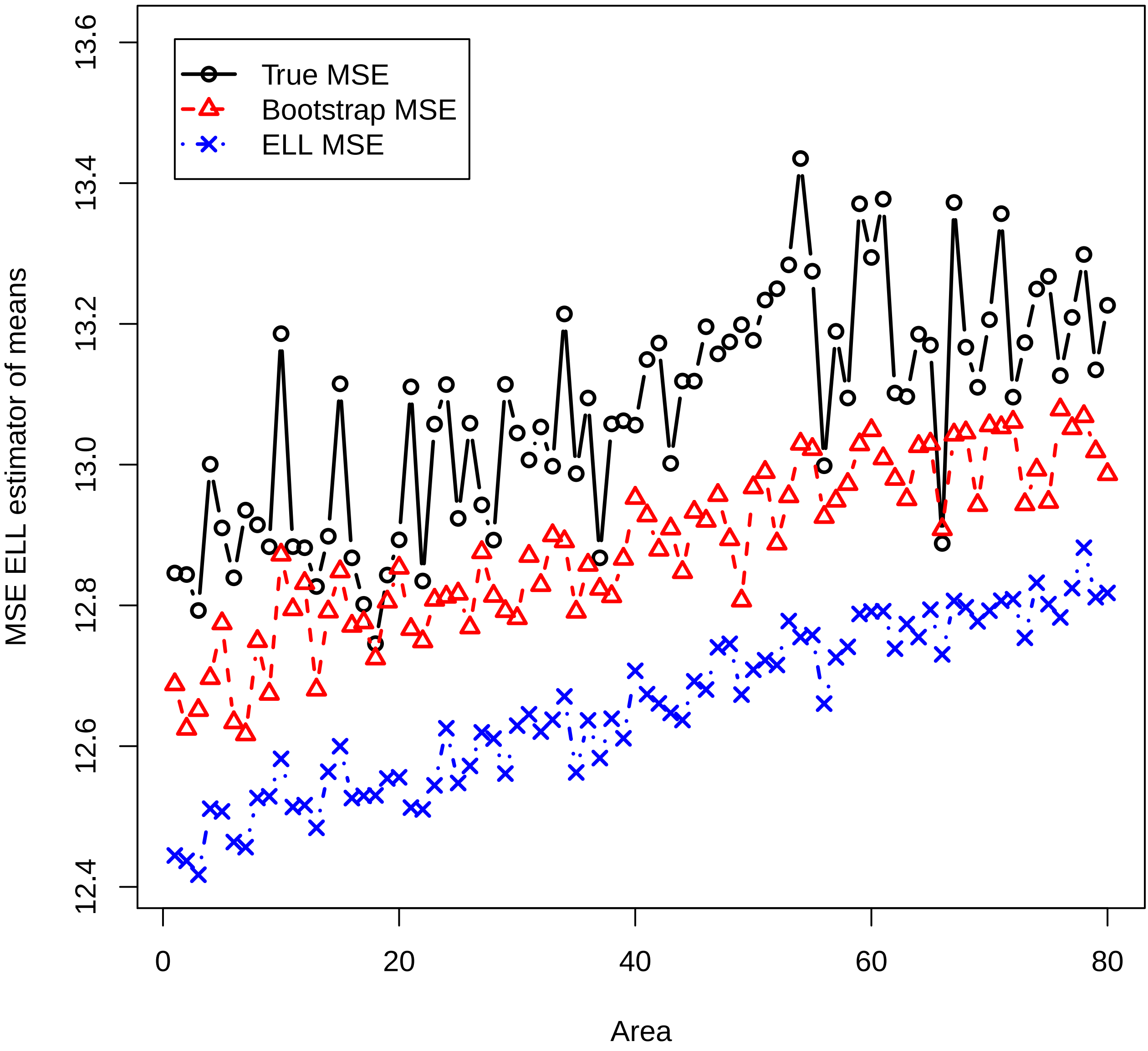


Figure 4. a) Mean and b) MSE of EB and Census EB estimators of poverty gaps F_{1d} for each area d.


True MSE
Bootstrap MSE
ELL MSE
MSE ELL estimator of means
Area
12.4
12.6
12.8
13.0
13.2
13.4
13.6
0
20
40
60
80

BOOTSTRAP MSE

- The bootstrap MSE tracks true MSE.

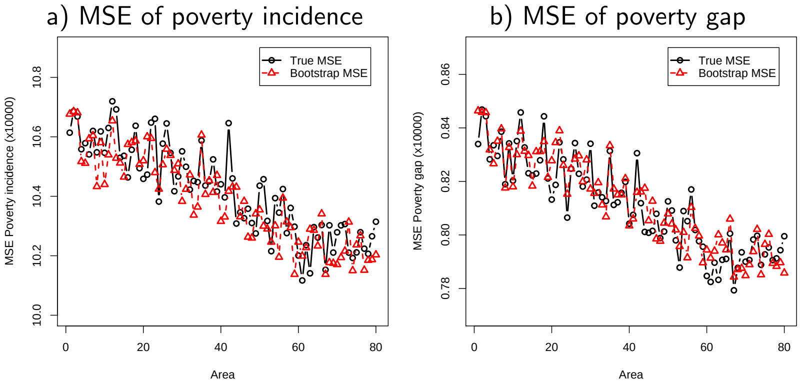


Figure 3. True MSEs and bootstrap estimators ($\times 10^4$) of EB estimators with $B = 500$ for each area d.

HIERARCHICAL BAYES METHOD

- Reparameterized nested-error model:

$$y_{di}|u_d, \boldsymbol{\beta}, \sigma^2 \overset{ind}{\sim} N(\mathbf{x}'_{di}\boldsymbol{\beta} + u_d, \sigma^2)$$
$$u_d|\rho, \sigma^2 \overset{ind}{\sim} N\left(0, \frac{\rho}{1-\rho}\sigma^2\right), \; i = 1, \ldots, N_d, \; d = 1, \ldots, D.$$

- $\rho \sim U(0,1)$ shrinkage or reference prior under a constant mean model, with good frequentist properties.
- $\pi(\sigma^2) \propto 1/\sigma^2$ Jeffreys objective or reference prior.
- $\pi(\boldsymbol{\beta}, \sigma^2, \rho) \propto 1/\sigma^2$ noninformative prior.

✓ *Rao, Nandram & Molina, Work in progress*

HIERARCHICAL BAYES METHOD

- Proper posterior density (provided $\mathbf{X}$ full column rank):

$$\pi(\mathbf{u}, \boldsymbol{\beta}, \sigma^2, \rho|\mathbf{y}_s) = \pi_1(\mathbf{u}|\boldsymbol{\beta}, \sigma^2, \rho, \mathbf{y}_s)\, \pi_2(\boldsymbol{\beta}|\sigma^2, \rho, \mathbf{y}_s)\, \pi_3(\sigma^2|\rho, \mathbf{y}_s)\, \pi_4(\rho|\mathbf{y}_s)$$

- $u_d|\boldsymbol{\beta}, \sigma^2, \rho, \mathbf{y}_s \overset{ind}{\sim}$ Normal.
- $\boldsymbol{\beta}|\sigma^2, \rho, \mathbf{y}_s \sim$ Normal.
- $\sigma^{-2}|\rho, \mathbf{y}_s \sim$ Gamma.
- $\pi_4(\rho|\mathbf{y}_s)$ not simple but ρ-values from it can be generated using a grid method.

✓ *Rao, Nandram & Molina, Work in progress*

HIERARCHICAL BAYES METHOD

- $\boldsymbol{\theta} = (\mathbf{u}', \boldsymbol{\beta}, \sigma^2, \rho)'$ vector of parameters.
- Distribution of out-of-sample values given parameters:

$$Y_{di}|\boldsymbol{\theta} \stackrel{ind}{\sim} N(\mathbf{x}_{di}'\boldsymbol{\beta} + u_d, \sigma^2), \quad i \in r_d, \quad d = 1, \ldots, D.$$

- Hierarchical Bayes estimator of $F_{\alpha d} = h_\alpha(\mathbf{y}_d)$:

$$\hat{F}_{\alpha d}^{HB} = E_{\mathbf{y}_{dr}}(F_{\alpha d}|\mathbf{y}_s),$$

with

$$f(\mathbf{y}_{dr}|\mathbf{y}_s) = \int \prod_{i \in r_d} f(Y_{di}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}_s)d\boldsymbol{\theta}.$$

✓ *Rao, Nandram & Molina, Work in progress*

APPLICATION WITH SILC DATA

- We assume the nested error model for the log-equivalized annual net income ($Y_{dj} = T(E_{dj}) = \log E_{dj}$).
- We fit the nested error model separately for each gender with provinces as areas ($D = 52$).
- We take as explanatory variables, the indicators of 5 age groups, of having Spanish nationality, of 3 education levels and of labor force status (unemployed, employed or inactive).

HIERARCHICAL BAYES METHOD

- HB estimates practically the same as EB ones.

- The same result in simulations under the frequential setup (frequential validity).

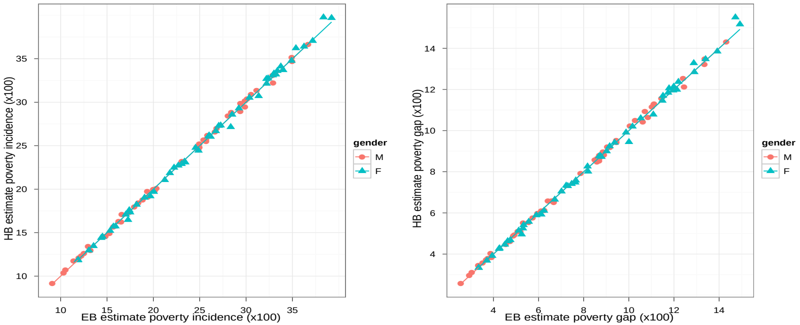


Figure 3 HB estimates of poverty incidence F_{0d} (left) and of poverty gap F_{1d} (right) against EB estimates for each province d.

HIERARCHICAL BAYES METHOD

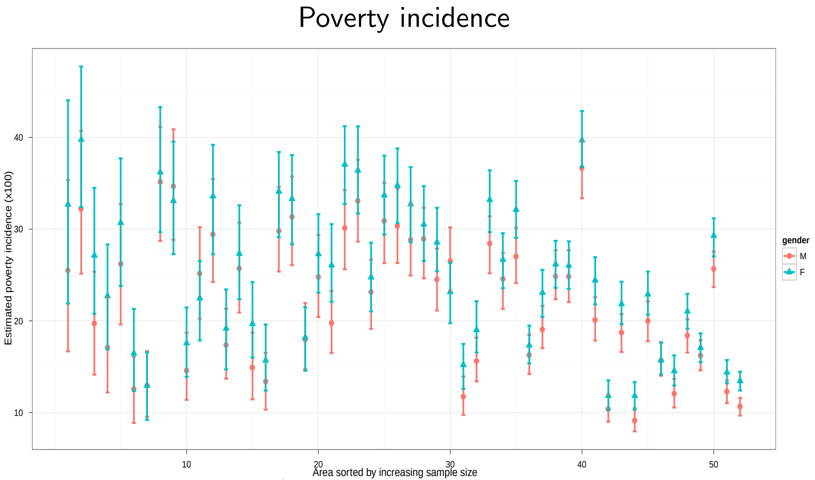


Figure 4 HB estimates of poverty incidences with HPD intervals by gender, for each province d. Provinces sorted by increasing sample size.

HIERARCHICAL BAYES METHOD

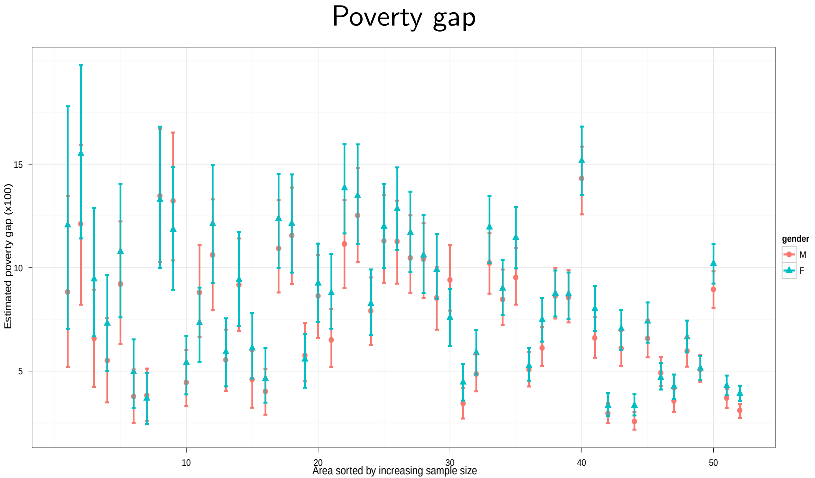


Figure 5 HB estimates of poverty gaps with HPD intervals by gender, for each province d. Provinces sorted by increasing sample size.

HIERARCHICAL BAYES METHOD

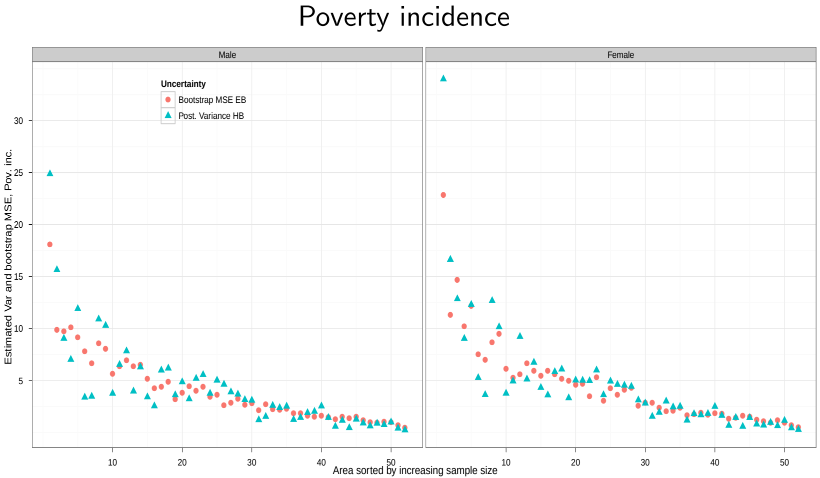
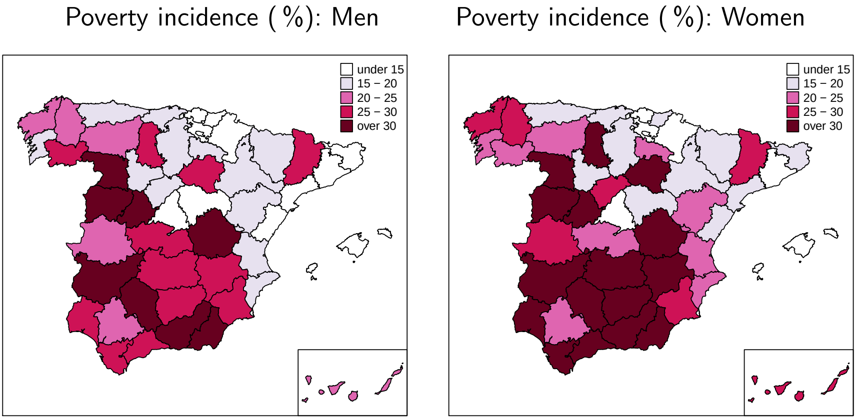


Figure 6 Posterior variance of HB estimators and bootstrap MSE of EB estimators of poverty incidence for each province for Males (left) and Females (right). Provinces sorted by increasing sample size.

RESULTS

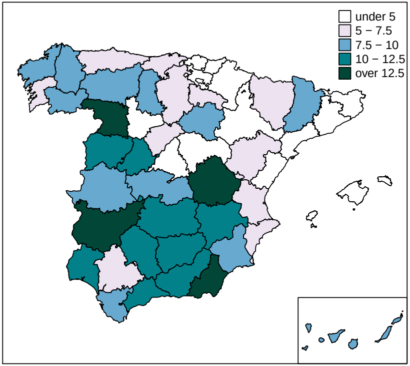


Pov.inc.$\geq$ 30 %, Men: Almería, Granada, Córdoba, Badajoz, Ávila, Salamanca, Zamora, Cuenca.

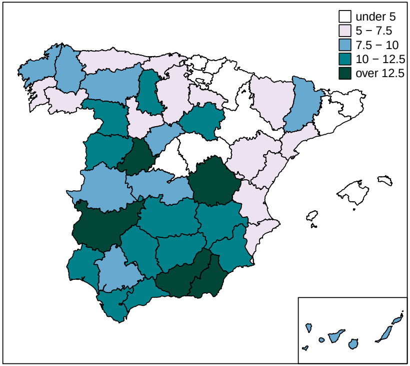
Women: also Jaén, Albacete, Ciudad Real, Palencia, Soria.

RESULTS

Poverty gap (%): Men



Poverty gap (%): Women



Pov.gap ≥ 12.5 %, **Men:** Almería, Badajoz, Zamora, Cuenca.
Women: Granada, Amería, Badajoz, Ávila, Cuenca.

M-quantile SAE estimation: Giusti et al. (2012)

- Model: M-quantile of order q of the conditional distribution of y given x, $Q_\psi(x,q) = x'\beta_\psi(q)$.

- $\hat{\beta}_\psi(q)$: Estimator of $\beta_\psi(q)$ for specified q. Solve $y_{dj} = \hat{Q}_\psi(x'_{dj}, q)$ to get $\hat{q}_{dj}$ for $j \in s_d$ and take their mean $\hat{\theta}_d$.

- Predictor of y_{dj} for $j \in r_d$ is $\hat{y}_{dj} = x'_{dj}\hat{\beta}_\psi(\hat{\theta}_q)$.

- M-quantile estimator of $\bar{Y}_d$ for small n_d and relatively large N_d is approximately given by

$$\bar{y}_d^{MQ} \approx \bar{y}_d + (\bar{X}_d - \bar{x}_d)'\hat{\beta}_\psi(\hat{\theta}_d)$$

- Note that MQ estimator looks similar to sample regression estimator of Fuller which is a component of the EBLUP under nested error model. Hence MQ can be considerably less efficient than EBLUP under nested error model with significant area effects

- Similar comments apply to the MQ estimator of the poverty measure $F_{\alpha d}$ considered by Giusti et al. (2012).

SKEW-NORMAL EB

- Nester error model with e_{dj} skew normal

$$u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{dj} \overset{iid}{\sim} SN(0, \sigma_e^2, \lambda_e)$$

$$\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2, \lambda_e)'$$

$$\lambda_e = 0 \text{ corresponds to Normal}$$

- As in the Normal case, EB estimator can be computed by generating only **univariate** normal variables, conditionally given a half-normal variable $T = t$.
- SN-EB was computed assuming $\boldsymbol{\theta}$ is known.

SKEW-NORMAL EB SIMULATION

- EB biased under significant skewness ($\lambda > 1$) unlike SN EB.

a) Bias of SN-EB estimator   b) Bias of EB estimator

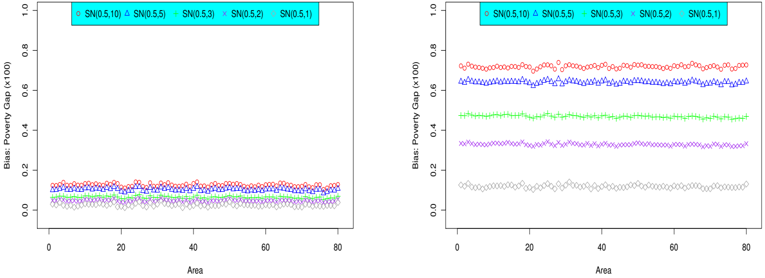


Figure 6. Bias of a) SN-EB estimator and b) EB estimator under skew normal distributions for error term for $\lambda = 1, 2, 3, 5, 10$.

✓ *Diallo & Rao, Work in progress*

SKEW-NORMAL EB SIMULATION

- RMSE = MSE(EB)/MSE(SN-EB)
- SN-EB significantly more efficient than EB when $\lambda > 1$.

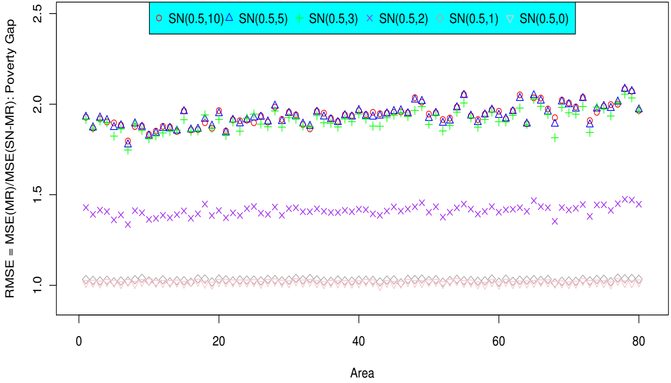


Figure 7. RMSE for skewness parameter $\lambda = 1, 2, 3, 5, 10$.

CONCLUSIONS

- We studied **EB and HB** estimation of **complex** small area parameters.
- Method applicable to **unit level** data.
- EB method assumes **normality** for some transformation of the variable of interest. EB work extended to **skew normal** distributions.
- It requires the knowledge of **all population values** of the auxiliary variables.
- It requires **computational effort** because large number of populations are generated. **Fast EB method** available.

CONCLUSIONS

- Original EB method, unlike ELL method, requires **linking** sample with census data for the auxiliary variables. **Census EB** method avoids the linking and is practically the same as original EB.
- Both EB and ELL methods assume that the sample is **non-informative**, that is, the model for the population holds good for the sample. Under informative sampling, probably both methods are biased. Currently an extension of EB method accounting for **informative** sampling is being studied.