A review of small area estimation methods for poverty mapping

Isabel Molina
Dept. of Statistics, Univ. Carlos III de Madrid

Coauthors: María Guadarrama, Balgobin Nandram, J.N.K. Rao
POPULATION AT RISK OF POVERTY
EXAMPLE

Survey on Income and Living Conditions 2006

Total sample size: $n = 34,389$ persons.
Province $\times$ gender sample sizes:

<table>
<thead>
<tr>
<th>Province $\times$ Gender</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Barcelona,F)</td>
<td>1483</td>
</tr>
<tr>
<td>(Córdoba,F)</td>
<td>230</td>
</tr>
<tr>
<td>(Tarragona,M)</td>
<td>129</td>
</tr>
<tr>
<td>(Soria,F)</td>
<td>17</td>
</tr>
</tbody>
</table>
NOTATION

- $U$ finite population of size $N$.
- $U$ partitioned into $D$ subsets $U_1, \ldots, U_D$ of sizes $N_1, \ldots, N_D$, called domains or areas.
- $s$ sample of size $n$ drawn from the population $U$.
- $s_d = s \cap U_d$ sub-sample from domain $d$ of size $n_d$.
- $r_d = U_d - s_d$ out-of-sample elements from domain $d$. 
DOMAIN PARAMETERS

• $y_{dj}$ outcome for unit $j$ in area $d$.
• $y_d = (y_{d1}, \ldots, y_{dN_d})'$ vector of outcomes for area $d$.
• Target quantities: Possibly non-linear function of $y_d$,
  $$\delta_d = h_d(y_d), \quad d = 1, \ldots, D.$$  
• Example: mean of $d$-th area,
  $$\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}.$$
DIRECT ESTIMATION

- Based essentially on the **area-specific** sample data.
- \( \pi_{dj} = P(j \in s_d) \) inclusion prob., \( w_{dj} = 1/\pi_{dj} \) sampling weight.
- Sampling weights \( w_{dj} \) protect against **informative sampling** (probability of selection depending on outcomes).
- **Example:** Horvitz-Thompson direct estimator,

\[
\hat{Y}_d^{DIR} = \frac{1}{N_d} \sum_{j \in s_d} w_{dj} y_{dj}.
\]

- Sampling variance \( V_{\pi}(\hat{Y}_d^{DIR}) \) can be estimated easily with the **area-specific** data.
POVERTY AND INEQ. INDICATORS

- \( E_{dj} \) welfare measure for indiv. \( j \) in domain \( d \).
- \( z \) = poverty line.
- **FGT poverty indicator of order \( \alpha \) for domain \( d \):**

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

- When \( \alpha = 0 \) ⇒ **Poverty incidence** (or at-risk-of-poverty rate)
- When \( \alpha = 1 \) ⇒ **Poverty gap**
- **Other:** Quintile share ratio, Gini coef., Sen index, Theil index, Generalized entropy, Fuzzy monetary/supplementary index.

✓ Foster, Greer & Thornbecke (1984), Econom.
✓ Neri, Ballini & Betti (2005), Stat. in Transition
DIRECT ESTIMATORS

- FGT pov. indicator as a mean:

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

- HT estimator:

\[ \hat{F}^{DIR}_{\alpha d} = \frac{1}{N_d} \sum_{j \in s_d} w_{dj} F_{\alpha dj}. \]
DIRECT ESTIMATORS

ADVANTAGES:

• **No model** assumptions.
• Sampling weights can be used $\Rightarrow$ Approx. **design-unbiased** even under informative sampling.
• Additivity (**Benchmarking** property):

$$\sum_{d=1}^{D} \hat{Y}_{d}^{DIR} = \hat{Y}^{DIR}.$$ 

DISADVANTAGES:

• $V_{\pi}(\hat{Y}_{d}^{DIR}) \uparrow$ as $n_{d} \downarrow$. Very **inefficient** for small domains.
• Estimator of sampling error very **inefficient** for small domains.
• Cannot be calculated for out-of-sample areas.
INDIRECT ESTIMATORS

- **Indirect estimator**: It *borrows strength* from other areas by making some kind of *homogeneity* assumption across areas (model with *common* parameters) that uses *auxiliary information*. 
FAY-HERRIOT (FH) MODEL

(i) Linking model:

\[ \delta_d = x_d' \beta + u_d, \quad u_d \overset{iid}{\sim} (0, \sigma_u^2), \quad d = 1, \ldots, D \]

\[ \sigma_u^2 \text{ unknown} \]

(ii) Sampling model:

\[ \hat{\delta}_d^{DIR} = \delta_d + e_d, \quad e_d \overset{ind}{\sim} (0, \psi_d), \quad d = 1, \ldots, D \]

\[ u_d \text{ and } e_d \text{ indep., } \psi_d = V_\pi(\hat{\delta}_d^{DIR} | \delta_d) \text{ known } \forall d \]

(iii) Combined model: Linear mixed model

\[ \hat{\delta}_d^{DIR} = x_d' \beta + u_d + e_d, \quad d = 1, \ldots, D. \]

✓ Fay & Herriot (1979), JASA
BEST LINEAR UNBIASED PREDICTOR

- Minimizes the MSE among **linear** and **unbiased** estimators.
- Easily obtained by fitting the mixed model:

\[
\tilde{\delta}^{BLUP}_d = x'_d \tilde{\beta} + \tilde{u}_d,
\]

where

\[
\tilde{\beta} = \tilde{\beta}(\sigma^2_u) = \left( \sum_{d=1}^D \gamma_d x_d x'_d \right)^{-1} \left( \sum_{d=1}^D \gamma_d x_d \hat{\delta}^{DIR}_d \right),
\]

\[
\tilde{u}_d = \tilde{u}_d(\sigma^2_u) = \gamma_d (\hat{\delta}^{DIR}_d - x'_d \tilde{\beta}), \quad \gamma_d = \frac{\sigma^2_u}{\sigma^2_u + \psi_d}
\]
GOOD PROPERTY OF THE BLUP

- Weighted combination of direct and “regression synthetic” estimator:

\[ \tilde{\delta}^{BLUP}_d = \gamma_d \hat{\delta}^{DIR}_d + (1 - \gamma_d) x'_d \tilde{\beta}, \quad \gamma_d = \frac{\sigma^2}{\sigma^2_u + \psi_d}. \]

- When \( \hat{\delta}^{DIR}_d \) is reliable (\( \downarrow \psi_d \)) or when area heterogeneity is not well explained by \( x'_d \tilde{\beta} \) (\( \uparrow \sigma^2_u \)), \( \tilde{\delta}^{BLUP}_d \rightarrow \hat{\delta}^{DIR}_d \).

- Otherwise, if \( \hat{\delta}^{DIR}_d \) unreliable or \( x'_d \tilde{\beta} \) reliable, \( \tilde{\delta}^{BLUP}_d \rightarrow x'_d \tilde{\beta} \).
**EMPIRICAL BLUP (EBLUP)**

- \( \tilde{\delta}_{d}^{BLUP} \) depends on unknown \( \sigma_{u}^{2} \) through \( \tilde{\beta} \) and \( \gamma_{d} \):
  \[
  \tilde{\delta}_{d}^{BLUP} = \tilde{\delta}_{d}^{BLUP}(\sigma_{u}^{2})
  \]

- **Empirical** BLUP (EBLUP) of \( \delta_{d} \): \( \hat{\sigma}_{u}^{2} \) estimator of \( \sigma_{u}^{2} \),
  \[
  \hat{\delta}_{d}^{EBLUP} = \tilde{\delta}_{d}^{BLUP}(\hat{\sigma}_{u}^{2}), \quad d = 1, \ldots, D
  \]

- The EBLUP remains **model-unbiased** under certain conditions (typically satisfied).

- MSE of EBLUP under FH model can be approximated with \( o(1/D) \) bias **under normality**.
NESTED ERROR MODEL

- Nested error **unit level** model:

\[ y_{dj} = \mathbf{x}'_{dj} \beta + u_{d} + e_{dj}, \quad j = 1, \ldots, N_{d}, \ d = 1, \ldots, D \]

\[ u_{d} \overset{iid}{\sim} N(0, \sigma^2_u), \quad e_{dj} \overset{iid}{\sim} N(0, \sigma^2_e) \]

✓ Battese, Harter & Fuller (1988), JASA

- The distribution of incomes \( E_{dj} \) is highly right skewed.
- Select a transformation \( T() \) such that the distribution of \( y_{dj} = T(E_{dj}) \) is approximately Normal.
- **Assumption**: \( y_{dj} = T(E_{dj}) \) satisfies the nested error model.
EB METHOD FOR POVERTY ESTIMATION

- Poverty indicators in terms of $y_d = (y_{d1}, \ldots, y_{dN_d})'$:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left\{ \frac{z - T^{-1}(y_{dj})}{z} \right\}^\alpha I \left\{ T^{-1}(y_{dj}) < z \right\} = h_\alpha(y_d).$$

- Partition $y_d$ into sample and out-of-sample: $y_d = (y_{ds}', y_{dr}')'$

- **Best predictor:** Minimizes the MSE

$$\tilde{F}^B_{\alpha d} = E_{y_{dr}} \left[ F_{\alpha d}|y_{ds}; \beta, \sigma^2_u, \sigma^2_e \right].$$

- **Empirical best (EB) predictor:**

$$\hat{F}^{EB}_{\alpha d} = \tilde{F}^B_{\alpha d}(\hat{\beta}, \hat{\sigma}^2_u, \hat{\sigma}^2_e).$$

✓ Molina and Rao (2010), CJS
MODEL-BASED EXPERIMENT

- Simulate \( I = 10^4 \) populations from the nested-error model.
- For each population \( i = 1, \ldots, 10^4 \), compute true domain FGT poverty indicators \( F_{\alpha d}^{(i)} \) for \( \alpha = 0, 1 \) and \( d = 1, \ldots, D \).
- Take the sample part of each population (assuming SRS within each domain) and compute EB, direct and ELL estimates (Elbers, Lanjouw and Lanjouw (2003), Econometrica).
- Approximate true MSEs of EB estimators as

\[
\text{MSE}(\hat{F}_{\alpha d}^{EB}) = \frac{1}{I} \sum_{i=1}^{I} \left( \hat{F}_{\alpha d}^{EB}(i) - F_{\alpha d}^{(i)} \right)^2, \ \alpha = 0, 1, \ d = 1, \ldots, D.
\]

- Similarly for direct and ELL estimators.
MODEL-BASED EXPERIMENT

• Population and sample sizes:

\[ N = 20000, \quad D = 80 \]
\[ N_d = 250, \quad n_d = 50, \quad d = 1, \ldots, D \]

• Variance components:

\[ \sigma_e^2 = (0.5)^2, \quad \sigma_u^2 = (0.15)^2 \]

• Explanatory variables: 2 dummies:

\[ X_1 \in \{0, 1\}, \quad p_{1d} = 0.3 + 0.5d/80, \quad d = 1, \ldots, D. \]
\[ X_2 \in \{0, 1\}, \quad p_{2d} = 0.2, \quad d = 1, \ldots, D. \]

• Coefficients:

\[ \beta = (3, 0.03, -0.04)' \]
POVERTY INCIDENCE

- **EB much more efficient** than ELL and direct estimators.
- **ELL even less efficient** than direct estimators!

**Bias (%)**

**MSE \((\times 10^4)\)**
POVERTY GAP

- Same conclusions as for poverty incidence.
HIERARCHICAL BAYES METHOD

- Reparameterization:
  \[ \rho = \frac{\sigma^2_u}{(\sigma^2_u + \sigma^2_e)} \]

- Reparameterized nested-error model:
  \[ y_{dj} \mid u_d, \beta, \sigma^2_e \overset{\text{ind}}{\sim} N(x'_{dj} \beta + u_d, \sigma^2_e), \]
  \[ u_d \mid \rho, \sigma^2_e \overset{\text{ind}}{\sim} N\left(0, \frac{\rho}{1 - \rho} \sigma^2_e\right) \]

- Noninformative prior:
  \[ \pi(\beta, \sigma^2_e, \rho) \propto 1/\sigma^2_e \]

HIERARCHICAL BAYES METHOD

- **Proper** posterior density (provided $X$ full column rank and $\rho$ is in a closed interval from $(0, 1)$):

  $$\pi(u, \beta, \sigma^2_e, \rho|y_s) = \pi_1(u|\beta, \sigma^2_e, \rho, y_s) \pi_2(\beta|\sigma^2_e, \rho, y_s) \pi_3(\sigma^2_e|\rho, y_s) \pi_4(\rho|y_s)$$

- Conditional distributions:

  $$u_n|\beta, \sigma^2_e, \rho, y_s \sim^{\text{ind}} \text{Normal},$$
  $$\beta|\sigma^2_e, \rho, y_s \sim \text{Normal},$$
  $$\sigma^{-2}_e|\rho, y_s \sim \text{Gamma}$$

- $\pi_4(\rho|y_s)$ not simple but $\rho$-values can be generated using a grid method.

COMPARISON WITH FH ESTIMATES

- FH estimates **biased** because of **linearity** problems.
- HB≈EB estimates nearly unbiased and **much more efficient**.

![Graph showing relative bias and root MSE](image-url)
INFORMATIVE SAMPLING

- HB and EB estimates **biased** under **informative sampling**.
- FH estimates with known dependency of inclusion probabilities on true responses **less biased**.

Rel. Bias (%)  
![Graph showing relative bias (%) with markers for different methods (DIR, FH, HB)]

Rel. Root MSE (%)  
![Graph showing relative root MSE (%) with markers for different methods (DIR, FH, HB)]
PSEUDO EB

- Best predictor for additive area parameters:

\[
\tilde{F}_{\alpha_d}^B = \frac{1}{N_d} \left[ \sum_{j \in s_d} F_{\alpha dj} + \sum_{j \in r_d} E(F_{\alpha dj} | \bar{y}_{ds}) \right],
\]

- Under the nested-error model:

\[
E(F_{\alpha dj} | y_{ds}) = E(F_{\alpha dj} | \bar{y}_d) \rightarrow E(F_{\alpha dj} | \bar{y}_{dw}).
\]

- Pseudo Best predictor for additive parameters:

\[
\tilde{F}_{\alpha_d}^{PB} = \frac{1}{N_d} \left[ \sum_{j \in s_d} F_{\alpha dj} + \sum_{j \in r_d} E(F_{\alpha dj} | \bar{y}_{dw}) \right].
\]
PSEUDO EB

- Including **sampling weights** reduces the design bias!
- Pseudo EB estimators do not lose much efficiency.

![Rel. Bias (%)](image1)

![Rel. Root MSE (%)](image2)
OTHER EXTENSIONS

• Existence of **two grouping levels** $\rightarrow$ EB method under a **two-fold** nested-error model.
  ✓ Marhuenda, Molina, Morales and Rao (2017), JRSSA

• Particular non linear parameter: **Area mean** under a model for the **log-transformation** of the target variable $\rightarrow$ **Explicit exact EB** estimator and **asymptotic MSE**.
  ✓ Molina and Martín (2018), AOS

• EB method for poverty estimation assumes **normality** for some transformation of the variable of interest $\rightarrow$ Extension to **skewed** distributions.
  ✓ Graf, Marín and Molina (2018), Test
POVERTY MAPPING IN SPAIN

• **Data source:** Spanish Survey on Income and Living Conditions (EU-SILC) of 2006.

• **Target:** Calculate EB and HB estimates of poverty incidences and gaps for Spanish provinces by gender.

• **Areas:** $D = 52$ provinces for each gender. We fit a separate model for each gender.

• **Transformation:** We consider the nested-error model for the log-equivalized disposable income:

$$y_{dj} = T(E_{dj}) = \log(E_{dj} + k).$$

• **Explanatory variables:** 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 in simulations under the frequentist setup (frequentist validity).

Poverty incidence ( %)

Poverty gap ( %)
# Poverty Mapping in Spain

- Estimated CVs of direct, EB and HB estimators of poverty incidences for selected provinces crossed with gender:

<table>
<thead>
<tr>
<th>Province</th>
<th>Gender</th>
<th>$n_d$</th>
<th>Obs. Poor</th>
<th>$\hat{C}V$ Dir.</th>
<th>$\hat{C}V$ EB</th>
<th>$\hat{C}V$ HB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soria</td>
<td>F</td>
<td>17</td>
<td>6</td>
<td>51.87</td>
<td>16.56</td>
<td>19.82</td>
</tr>
<tr>
<td>Tarragona</td>
<td>M</td>
<td>129</td>
<td>18</td>
<td>24.44</td>
<td>14.88</td>
<td>12.35</td>
</tr>
<tr>
<td>Córdoba</td>
<td>F</td>
<td>230</td>
<td>73</td>
<td>13.05</td>
<td>6.24</td>
<td>6.93</td>
</tr>
<tr>
<td>Badajoz</td>
<td>M</td>
<td>472</td>
<td>175</td>
<td>8.38</td>
<td>3.48</td>
<td>4.24</td>
</tr>
<tr>
<td>Barcelona</td>
<td>F</td>
<td>1483</td>
<td>191</td>
<td>9.38</td>
<td>6.51</td>
<td>4.52</td>
</tr>
</tbody>
</table>
**RESULTS**

Poverty incidence ( %): Men

Poverty incidence ( %): Women

**Pov.inc. ≥ 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 $\geq 12.5\%$, Men: Almería, Badajoz, Zamora, Cuenca.
Women: Granada, Amería, Badajoz, Ávila, Cuenca.
COMPARISON WITH DIRECT ESTIMATORS

DISADVANTAGES:

• Require **model** assumptions (model checking important!).
• **Not design-unbiased** in general $\Rightarrow$ Sampling weights can be incorporated to reduce design bias.
• Require **adjustment** to satisfy the benchmarking property:

$$\sum_{d=1}^{D} \hat{Y}_d = \hat{Y}^{DIR}.$$ 

ADVANTAGES:

• Very **efficient** for small domains.
• Estimator of model MSE **efficient** for small domains as well.
• Can be calculated for **out-of-sample** areas.
SOFTWARE

The R package `sae` contains functions:

- FH model: `eblupFH`, `mseFH`.
- Spatial FH model: `eblupSFH`, `mseSFH`, `pbmseSFH`, `npbmseSFH`.
- Spatio-temporal FH model: `eblupSTFH`, `pbmseSTFH`.
- Nested-error model: `eblupBHF`, `pbmseBHF`.
- EB method: `ebBHF`, `pbmseebBHF`.
- Other: `direct`, `pssynt`, `ssd`.
- Data sets and examples.
MERCI BEAUCOUP!!