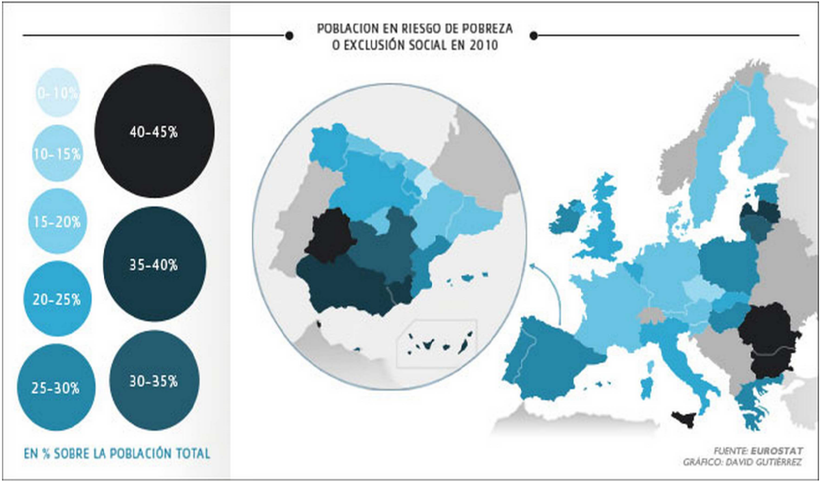
Poverty mapping in small areas: complex sampling problems

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

Coauthors: María Guadarrama, J.N.K. Rao

POPULATION AT RISK OF POVERTY

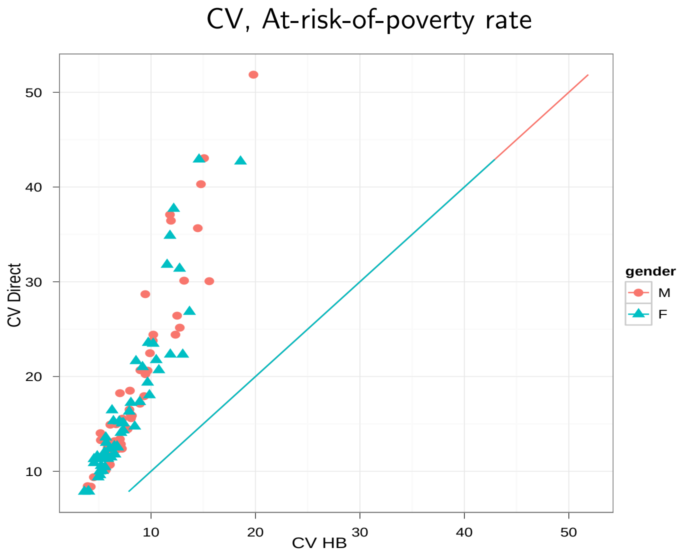


EXAMPLE: RISK OF POVERTY IN SPAIN

- Data: Survey on Income and Living Conditions, 2006.
- Sample size: $n = 34,389$ out of $N = 43,162,384$.
- Parameters: At-risk-of-poverty rates for the 52 provinces by gender.
- Poverty line $z = 0.6 \times$ Median(disposable equivalent income): In 2006, $z = 6,557$ euros→ approx. **20 %** at risk.

Province	Gender	n_d	At risk	$\hat{CV}$ Dir.	$\hat{CV}$ EB	$\hat{CV}$ 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

EXAMPLE: RISK OF POVERTY IN SPAIN



POVERTY AND INEQ. INDICATORS

- E_{dj} **welfare** measure for indiv. j in domain d.
- z = poverty line.
- **FGT poverty indicator of order α 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 $\alpha = 0 \Rightarrow$ **Poverty incidence** (or at-risk-of-poverty rate)
- When $\alpha = 1 \Rightarrow$ **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:

$$\hat{F}_{\alpha d}^{DIR} = \frac{1}{N_d} \sum_{j \in s_d} w_{dj} F_{\alpha dj}, \quad \hat{F}_{\alpha d}^{S} = \frac{1}{n_d} \sum_{j \in s_d} F_{\alpha dj}.$$

- Highly inefficient for areas d with small sample size n_d.

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**.

NESTED ERROR MODEL

- 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:

$$y_{dj} = \mathbf{x}'_{dj}\beta + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \; d = 1, \ldots, D$$

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

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

EB METHOD FOR POVERTY ESTIMATION

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

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

- Partition $\mathbf{y}_d$ into sample and out-of-sample: $\mathbf{y}_d = (\mathbf{y}_{ds}', \mathbf{y}_{dr}')'$
- **Best predictor:** Minimizes the MSE

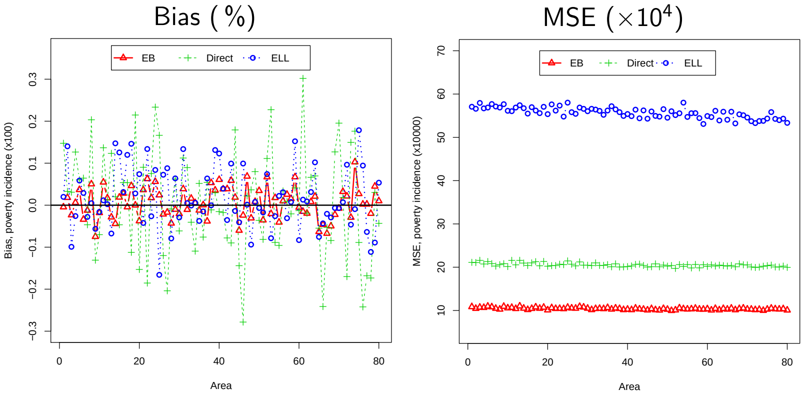
$$\tilde{F}_{\alpha d}^{B} = E_{\mathbf{y}_{dr}}\left[F_{\alpha d} | \mathbf{y}_{ds}; \boldsymbol{\beta}, \sigma_u^2, \sigma_e^2 \right].$$

- **Empirical best (EB) predictor:** $\hat{F}_{\alpha d}^{EB} = \tilde{F}_{\alpha d}^{B}(\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$.

✓ *Molina and Rao (2010), CJS*

POVERTY RATE

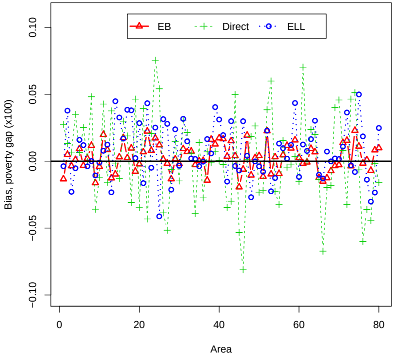
- EB **much more efficient** than ELL and direct estimators.
- ELL even **less efficient** than direct estimators!



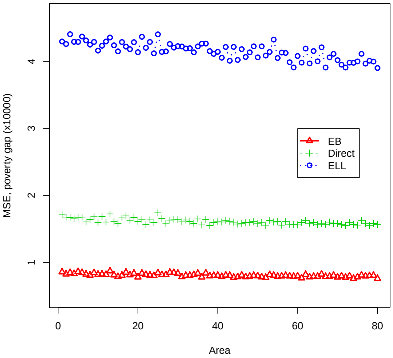
POVERTY GAP

- Same conclusions as for poverty incidence.

Bias (%)



MSE ($\times 10^4$)



PSEUDO EB

- Best predictor for additive area parameters:

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

- Under the nested-error model:

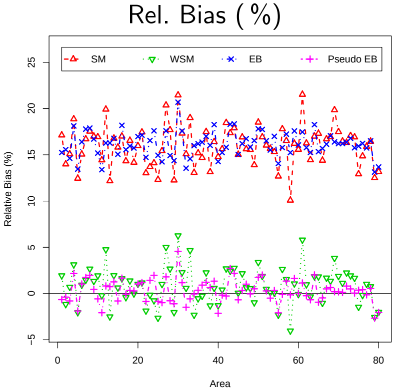
$$E(F_{\alpha dj}|\mathbf{y}_{ds}) = E(F_{\alpha dj}|\bar{y}_d) \longrightarrow E(F_{\alpha dj}|\bar{y}_{dw}).$$

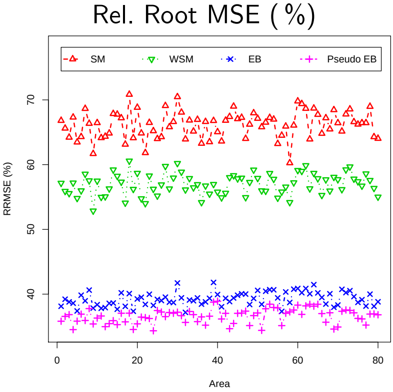
- **Pseudo Best predictor** for additive parameters:

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

PSEUDO EB

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



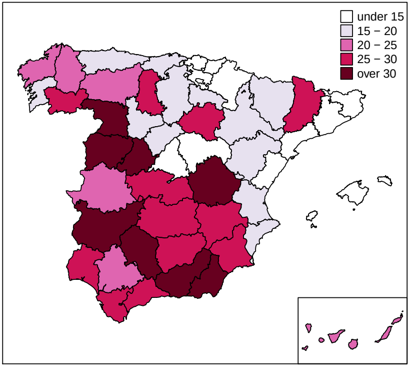


POVERTY MAPPING IN SPAIN

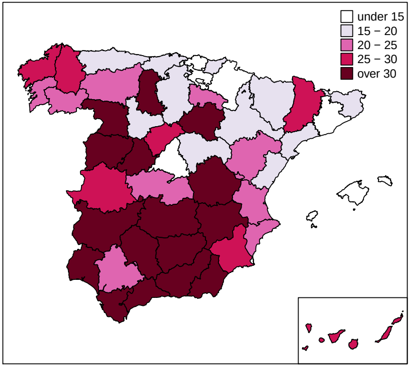
- **Data:** Spanish Survey on Income and Living Conditions (EU-SILC) of 2006.
- **Target:** Calculate EB and HB estimates of poverty incidences and gaps for Spanish provinces by gender.
- **Areas:** $D = 52$ **provinces** for each **gender**. We fit a separate model for each gender.
- **Transformation:** 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).

POVERTY RATES (%)

Men



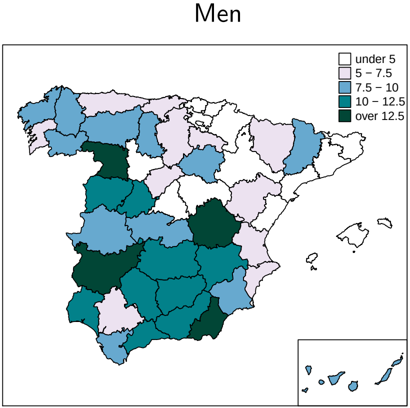
Women

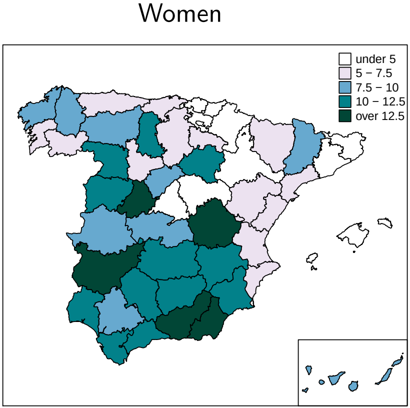


Pov.inc.$\geq$ 30 %, **Men:** Almería, Granada, Córdoba, Badajoz, Ávila, Salamanca, Zamora, Cuenca.

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

POVERTY GAPS (%)





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

SOURCES OF INFORMATION

- **Survey:** Unit level values of target variable and aux. variables for sampled units.
- **Census/Admin. records:** Values of aux. variables for each population unit $\longrightarrow$ confidentiality issues.
- **Aggregated aux. information:** Counts, totals/means of aux. variables from census/admin. registers at the area level or other aggregation level $\longrightarrow$ avoid confidentiality issues.
- **Larger surveys:** Estimated counts/totals of aux. variables $\longrightarrow$ Measurement error in covariates.
- **Non probability surveys, social media, Satellite/Images:** Counts/totals of aux. variables $\longrightarrow$ Potential bias.

THANK YOU VERY MUCH!!