Panorama reciente de los avances en EAP

Estimación de indicadores generales bajo modelos a nivel de unidad

Isabel Molina

Departamento de Estadística e Investigación Operativa,
Universidad Complutense de Madrid

RIESGO DE POBREZA EN ESPAÑA

- **Datos:** Encuesta de Condiciones de Vida, 2006.
- **Tamaño muestral:** $n = 34,389$ de entre $N = 43,162,384$.
- **Indicador:** Tasa en riesgo de pobreza.
- **Umbral pobreza:** $z = 0.6 \times$ Mediana(renta disp. equivalente): En 2006, $z = 6,557$ euros$\rightarrow$ aprox. **20 %** en riesgo.
- **Dominios:** 52 provincias por género.

Provincia	Género	n_d	Pobre	$\hat{CV}$ Dir.	$\hat{CV}$ EB
Soria	M	**17**	**6**	**51.87**	**16.56**
Tarragona	V	129	**18**	**24.44**	**14.88**
Córdoba	M	230	73	13.05	6.24
Badajoz	V	472	175	8.38	3.48
Barcelona	M	1483	191	9.38	6.51

✓ *Molina & Rao (2010), CJS*

POBLACIÓN

- U población **finita** de tamaño N.
- U particionada en D **dominios** o **áreas** $U_1, \ldots, U_D$ de tamaños $N_1, \ldots, N_D$, con $N = \sum_{d=1}^{D} N_d$.
- Y_{di} valor variable objetivo para indiv. i-ésimo del dominio d.
- $\mathbf{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$ vector para el **área** d.
- **Parámetros objetivo:** Funciones **generales** de $\mathbf{y}_d$,

$$\delta_d = h_d(\mathbf{y}_d), \quad d = 1, \ldots, D.$$

- Ejemplo: Medias de las áreas/dominios,

$$\delta_d = \bar{Y}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} Y_{di}, \quad d = 1, \ldots, D.$$

INDICADORES DE POBREZA Y DESIGUALDAD

- E_{di} **poder adquisitivo** (e.g. renta, gasto) del indiv. i en el dominio d.
- z umbral de pobreza.
- **Indicador de pobreza FGT de orden α para el dominio d:**

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

- Si $\alpha = 0 \Rightarrow$ **Tasa en riesgo de pobreza**
- Si $\alpha = 1 \Rightarrow$ **Brecha de pobreza**
- **Otros:** ratio entre quintiles, coeficiente de Gini, etc.

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

MUESTRA

- s muestra aleatoria de tamaño $n \leq N$ extraída de U.
- $s_d = s \cap U_d$ sub-muestra de tamaño $n_d \leq N_d$ del área/dominio d.
- Tamaño muestral total: $n = \sum_{d=1}^{D} n_d$.
- $c_d = U_d - s_d$ complemento de la muestra en el área d, de tamaño $N_d - n_d$.
- **Estimador directo**: Basado **sólo** en los n_d datos del área d.
- Muy **ineficientes** para n_d pequeño.
- **Estimador indirecto**: Asume alguna hipótesis de **homogeneidad** entre las áreas (e.g. modelo con parámetros **comunes**), que permiten **compartir información** entre todas las áreas.

MODELO CON ERRORES ANIDADOS

- Modelo de regresión lineal con efectos aleatorios de las áreas:

$$Y_{di} = \mathbf{x}'_{di}\boldsymbol{\beta} + u_d + e_{di}, \;\; u_d \overset{iid}{\sim} N(0, \tau^2),$$
$$e_{di} \overset{iid}{\sim} N(0, \sigma^2), \;\; i = 1, \ldots, N_d, \;\; d = 1, \ldots, D.$$

- $\boldsymbol{\theta} = (\boldsymbol{\beta}', \tau^2, \sigma^2)'$ parámetros **comunes** a todas las áreas.
- u_d efecto **específico** del área d.

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

MODELO CON ERRORES ANIDADOS

- En notación matricial:

$$\mathbf{y}_d \overset{ind.}{\sim} N(\mathrm{X}_d\boldsymbol{\beta}, \mathrm{V}_d), \quad d = 1, \ldots, D,$$

donde la matriz de covarianzas es

$$\mathrm{V}_d = \begin{pmatrix} \tau^2 + \sigma^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \tau^2 + \sigma^2 & \cdots & \tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 & \tau^2 & \cdots & \tau^2 + \sigma^2 \end{pmatrix}$$

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

ESTIMADORES ÓPTIMOS (BEST)

- Estimamos indicadores de pobreza monetaria.
- Asumimos el modelo con errores anidados para una transf. biyectiva del poder adquisitivo:

$$Y_{di} = T(E_{di}).$$

- Transformación habitual:

$$Y_{di} = \log(E_{di} + k), \quad k > |\,\text{mín}(E_{di})|.$$

- Para seleccionar k, ajustar el modelo a una malla de valores de k en $[\text{máx}(0, \text{mín}(E_{di})), \text{máx}(E_{di})]$.
- Tomar k^* para el cual el coef. asimetría de Fisher de los residuos $\hat{e}_{di} = Y_{di} - \mathbf{x}'_{di}\hat{\boldsymbol{\beta}} + \hat{u}_d$ sea aprox. cero.

✓ *Molina & Rao (2018), CJS*

ESTIMADORES ÓPTIMOS (BEST)

- Expresar el indicador de interés δ_d como función de $\mathbf{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$ mediante la transf. inversa:

$$E_{di} = T^{-1}(Y_{di}).$$

- Para indicadores FGT, bajo la transformación $Y_{di} = \log(E_{di} + k) \Leftrightarrow E_{di} = e^{Y_{di}} - k$,

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z - e^{Y_{di}} + k}{z} \right)^{\alpha} I(e^{Y_{di}} - k < z) = h_d(\mathbf{y}_d).$$

✓ *Molina & Rao (2018), CJS*

ESTIMADORES ÓPTIMOS (BEST)

- Separamos en parte muestral y no muestral:

$$\mathbf{y}_d = (\mathbf{y}'_{ds}, \mathbf{y}'_{dc})', \quad d = 1, \ldots, D.$$

- Predictor óptimo (best predictor) de $\delta_d = h_d(\mathbf{y}_d)$: Predictor $\tilde{\delta}_d$ que minimiza el ECM bajo el modelo

$$\text{MSE}_{\mathbf{y}}(\tilde{\delta}_d) = E_{\mathbf{y}}\left[(\tilde{\delta}_d - \delta_d)^2\right].$$

- Viene dado por:

$$\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\mathbf{y}_{dc}}[\delta_d | \mathbf{y}_{ds}].$$

- El predictor óptimo depende de $\boldsymbol{\theta} = (\boldsymbol{\beta}', \tau^2, \sigma^2)'$.

✓ *Molina & Rao (2018), CJS*

ESTIMADORES EMPÍRICOS ÓPTIMOS (EB)

- $\mathbf{y}_s = (\mathbf{y}'_{1s}, \ldots, \mathbf{y}'_{Ds})'$ datos de todas las áreas.
- Obtener estimador consistente $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ basado en $f(\mathbf{y}_s; \boldsymbol{\theta})$.
- Si no hay sesgo de selección, obtenemos $f(\mathbf{y}_{ds}; \boldsymbol{\theta})$ marginalizando en $f(\mathbf{y}_d; \boldsymbol{\theta})$ y el modelo para la muestra es el mismo que para la población.
- Predictor empírico óptimo:

$$\hat{\delta}_d^{EB} = \tilde{\delta}_d^B(\hat{\boldsymbol{\theta}}).$$

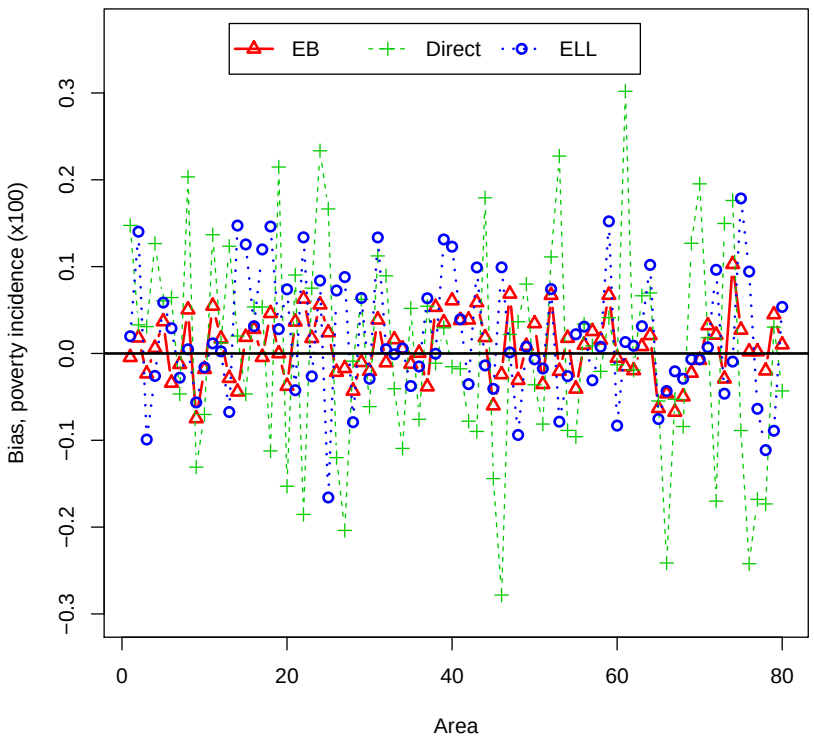
- Para indicadores complicados, un algoritmo de simulación MC nos aproxima la esperanza que define los estimadores EB.
- Método bootstrap paramétrico para la estimación del ECM.

✓ *Molina & Rao (2018), CJS*

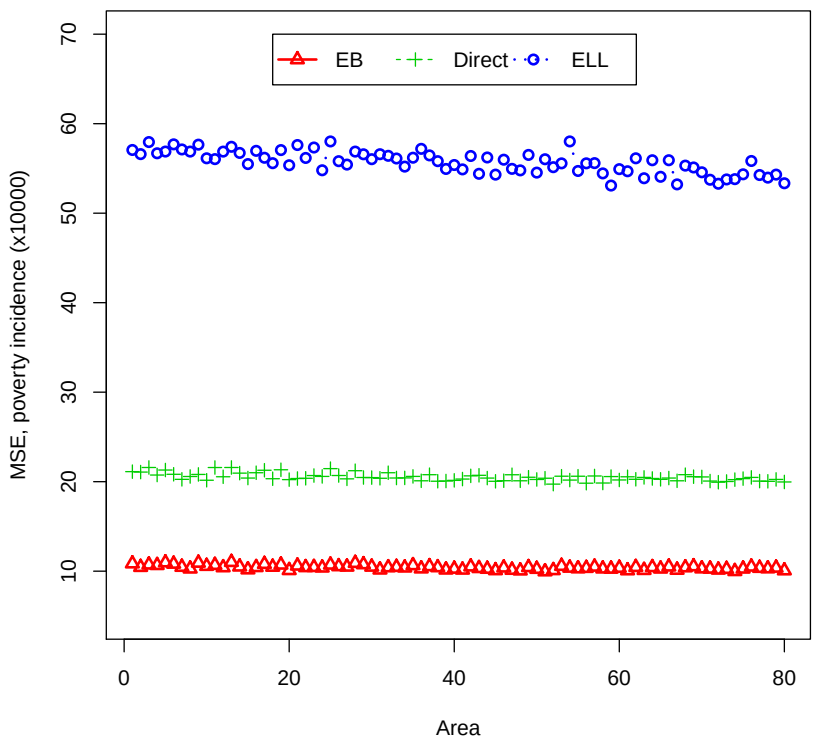
RIESGO DE POBREZA

- EB mucho más eficiente que ELL y estim. directos.
- ELL incluso menos eficientes que estim. directos!

Sesgo (%)



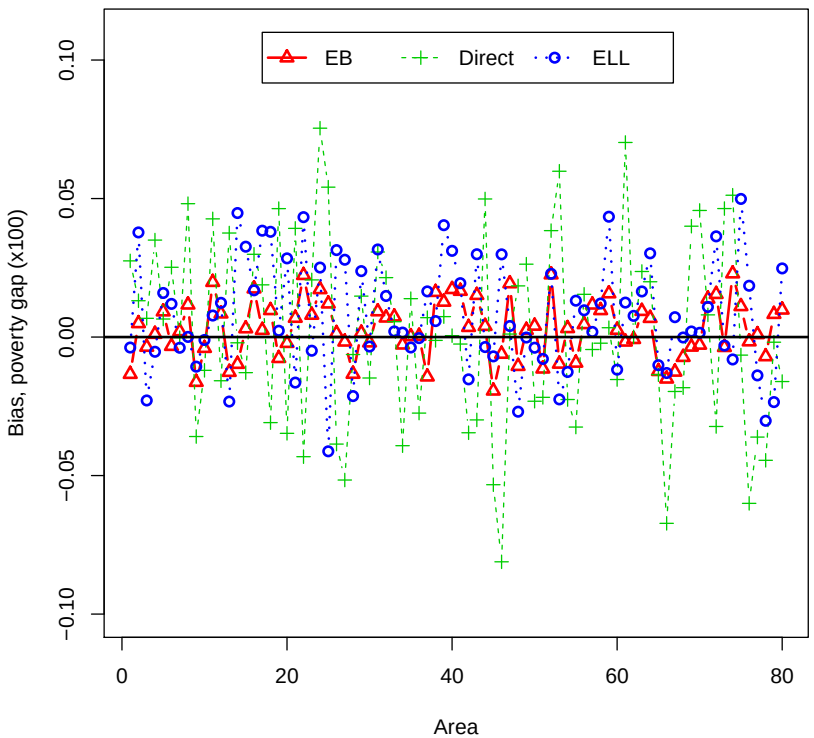
ECM ($\times 10^4$)



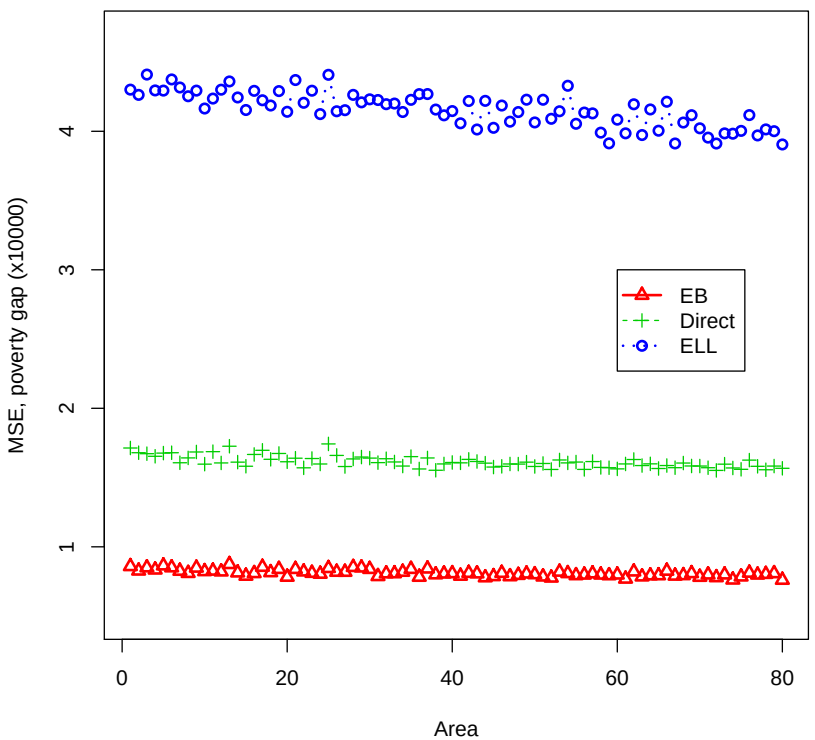
BRECHA DE POBREZA

- Mismas conclusiones

Sesgo (%)

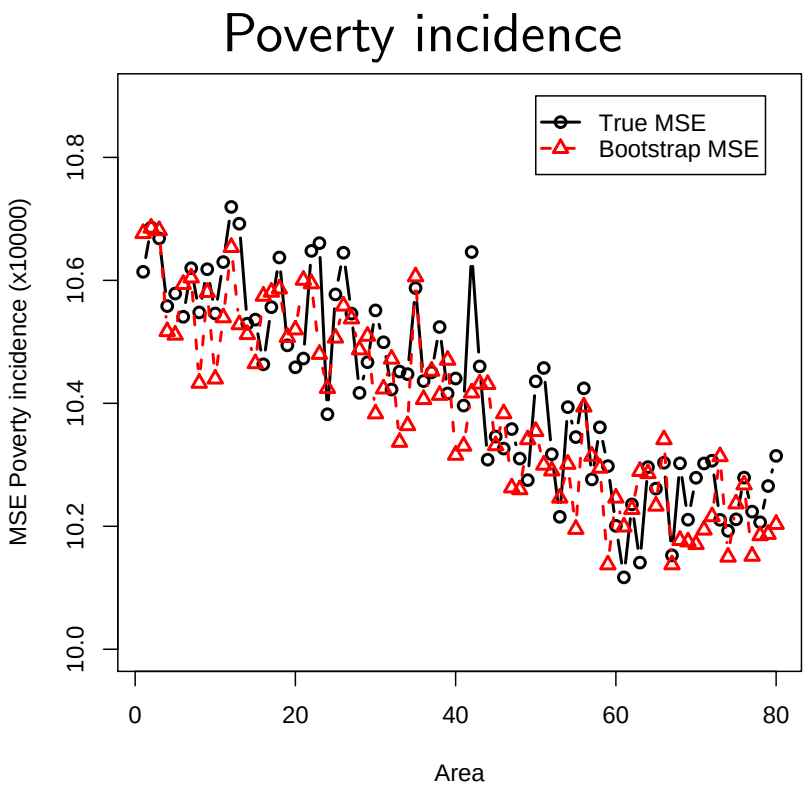


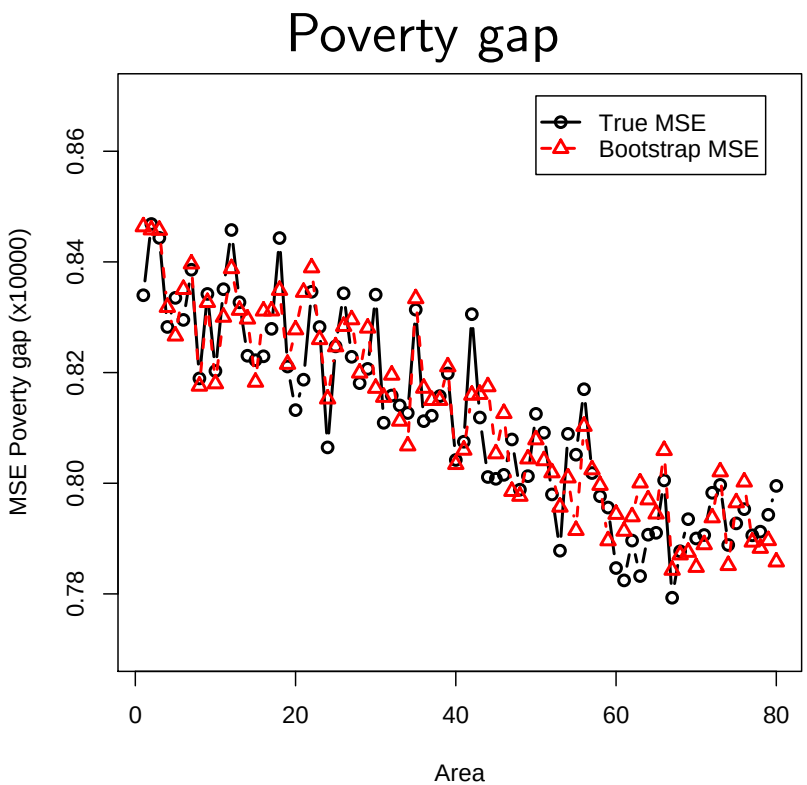
ECM ($\times 10^4$)



ECM POR BOOTSTRAP

- El ECM estimado por bootstrap ($B = 500$) aproxima el verdadero ECM.



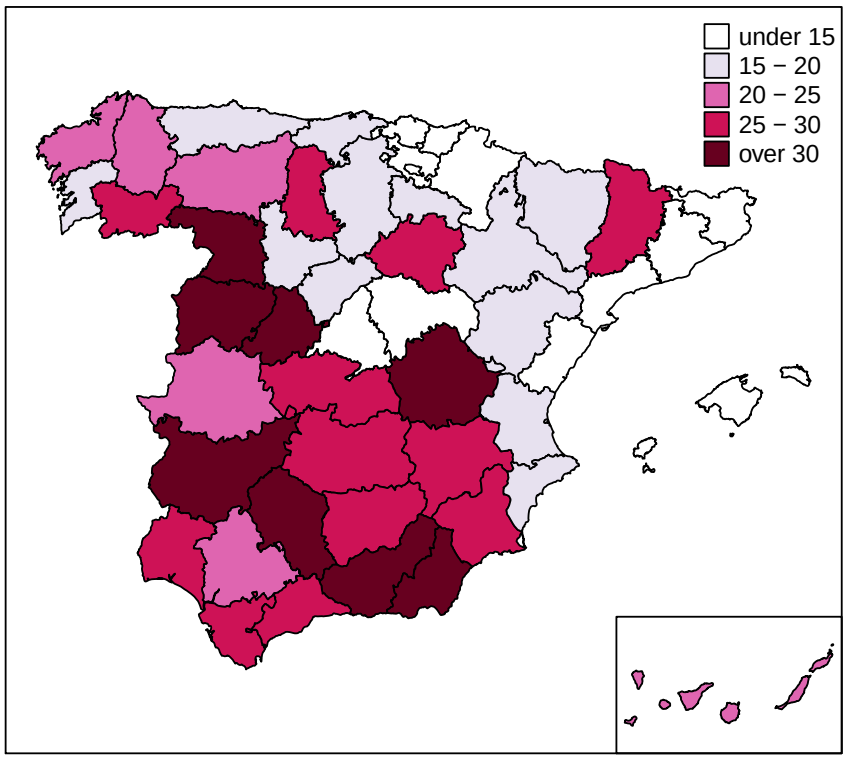


RIESGO DE POBREZA EN ESPAÑA

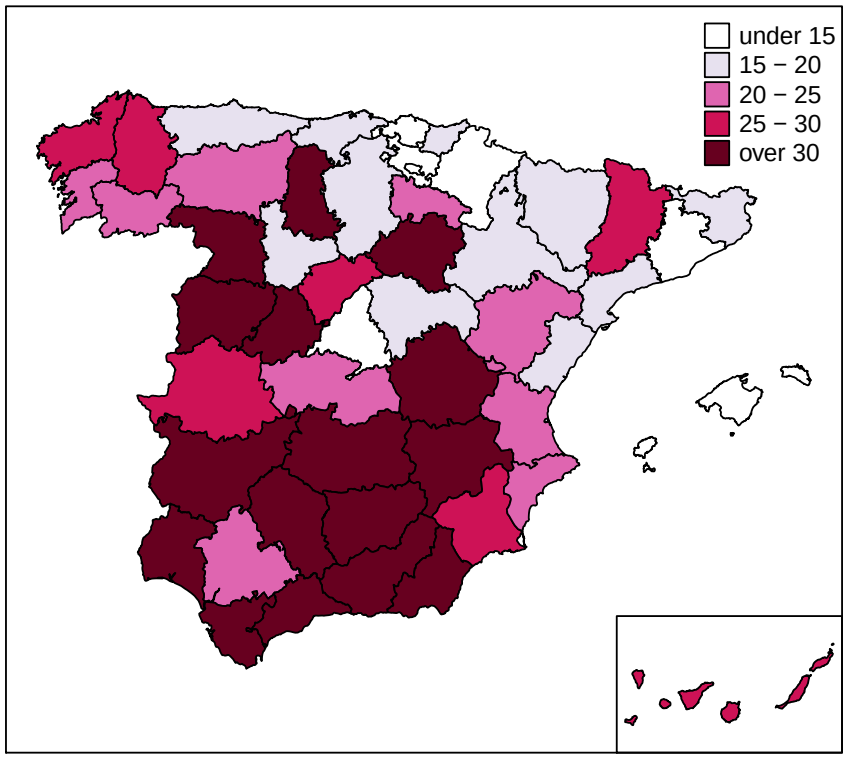
- Se asume el modelo con errores anidados para $Y_{di} = T(E_{di}) = \log(E_{di} + k)$.
- Ajustamos un modelo con errores anidados separado para cada género, con las provincias como áreas ($D = 52$).
- Variables explicativas: indicadores de
  - ✓ 5 grupos de edad;
  - ✓ poseer nacionalidad española;
  - ✓ 3 niveles educativos;
  - ✓ estado laboral (desempleado, empleado o inactivo).

RESULTADOS: RIESGO DE POBREZA (%)

Hombres



Mujeres

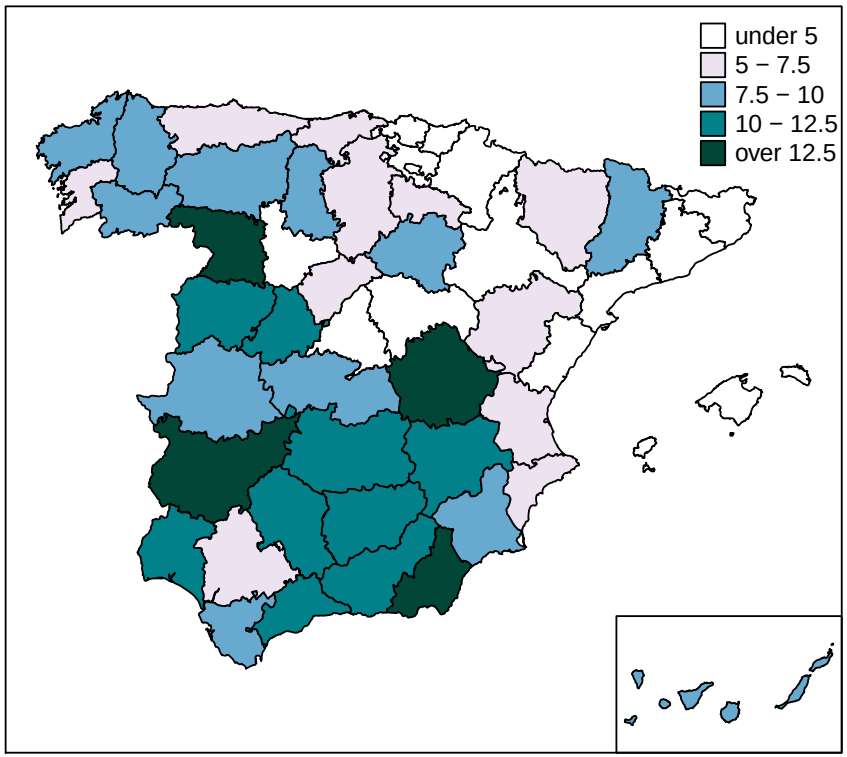


Inc. Pob.$\geq$ 30 %, **Hombres:** Almería, Granada, Córdoba, Badajoz, Ávila, Salamanca, Zamora, Cuenca.

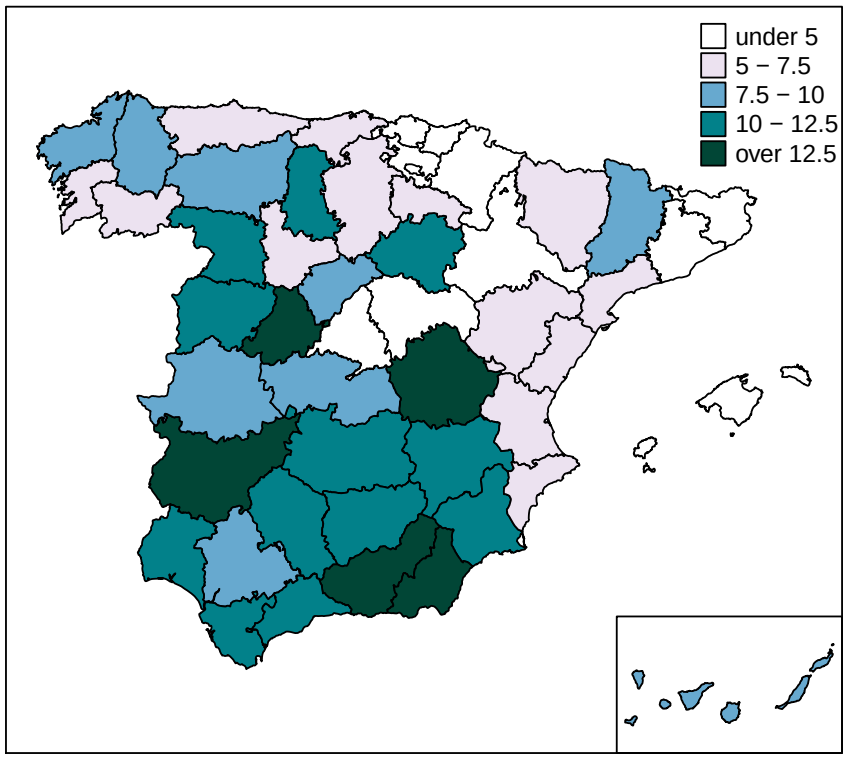
Mujeres: también Jaén, Albacete, Ciudad Real, Palencia, Soria.

RESULTADOS: BRECHA DE POBREZA (%)

Hombres



Mujeres



Brecha Pob. $\geq$ **12.5 %**, **Hombres:** Almería, Badajoz, Zamora, Cuenca.
Mujeres: Granada, Almería, Badajoz, Ávila, Cuenca.

REQUERIMIENTOS DE DATOS

- Aunque se deseen estimar medias $\delta_d = \bar{E}_d$, si tomamos como variable respuesta del modelo una transformación de E_{di}, el EBLUP no tiene sentido $\rightarrow$ EB.
- Para estimar de forma óptima indicadores no lineales $\delta_d = h_d(\mathbf{y}_d)$ con datos a nivel de individuo, se necesita:
  - ✓ Una **encuesta** con microdatos de la variable de interés y variables auxiliares;
  - ✓ Un **censo** con microdatos de las variables auxiliares.
- EB original asume que las unidades de la encuesta están en el censo y requiere identificarlas en el fichero del censo.

MÉTODO CENSUS EB

- **Census EB** (CEB) no asume que la muestra de la encuesta es parte de las unidades del censo.
- El estimador Census EB es el EB para $\delta_d = h_d(\mathbf{y}_d)$, si los vectores aumentados $\mathbf{y}_{d,a} = (\mathbf{y}'_d, \mathbf{y}'_{ds})'$, $d = 1, \ldots, D$, siguen el modelo con errores anidados.
- ECM del Census EB estimado por una variación del método bootstrap paramétrico para el EB.

✓ *Correa, Molina & Rao (2012)*
✓ *Molina (2019), CEPAL*

MUESTREO INFORMATIVO

- Muestra sin reemplazo: $\mathbf{I}_d = (I_{d1}, \ldots, I_{dN_d})'$, donde, para cada $i = 1, \ldots, N_d$,

$$I_{di} = \begin{cases} 1, & \text{si unidad } i \text{ del área } d \text{ es seleccionada} \\ 0, & \text{en otro caso} \end{cases}$$

- Entonces,

$$s_d = \{i \in U_d : I_{di} = 1\}.$$

- Muestreo no informativo:

$$P(\mathbf{I}_d = \mathbf{i}_d | \mathbf{y}_d) = P(\mathbf{I}_d = \mathbf{i}_d), \quad \forall \mathbf{y}_d \in \boldsymbol{R}^{N_d}, \ \forall \mathbf{i}_d \in \{0,1\}^{N_d}.$$

- Una vez extraída la muestra, por el Teorema de Bayes:

$$f(\mathbf{y}|\mathbf{I} = \mathbf{i}) = f(\mathbf{y}) \frac{P(\mathbf{I} = \mathbf{i}|\mathbf{y})}{\int P(\mathbf{I} = \mathbf{i}|\mathbf{y}) f(\mathbf{y}) d\mathbf{y}}.$$

MUESTREO INFORMATIVO

- Verosimilitud: $f(\mathbf{y}_s|\mathbf{I}=\mathbf{i})$, se obtiene marginalizando en la conjunta $f(\mathbf{y}|\mathbf{I}=\mathbf{i})$.
- Si el muestreo es informativo, $f(\mathbf{y}_s|\mathbf{I}=\mathbf{i})$ no necesariamente tiene la misma forma que $f(\mathbf{y})$.
- Predictor óptimo de $\delta_d = h_d(\mathbf{y}_d)$: Predictor $\tilde{\delta}_d$ que minimiza

$$\text{MSE}_{(\mathbf{y},\mathbf{I})}(\tilde{\delta}_d) = E_{(\mathbf{y},\mathbf{I})}\left[(\tilde{\delta}_d - \delta_d)^2\right].$$

- Viene dado por:

$$\tilde{\delta}_d^{B\text{-}I}(\boldsymbol{\theta}) = E_{\mathbf{y}_{dc}}\left[\delta_d|\text{Datos}, \mathbf{I}_d = \mathbf{i}_d\right].$$

✓ *Pfeffermann & Sverchkov (2007), JASA*

✓ *Berg, Cho, Eideh, Guadarrama & Molina (2023), Enviado*

MUESTREO INFORMATIVO

- Se asume un modelo para $E(w_{di}|I_{di}=1,\mathbf{x}_{di},Y_{di},u_i)$ en función de $\mathbf{x}_{di}$ y de Y_{di}.
- Fórmulas explícitas para el EB bajo diseño informativo (EB_I), para tasas de riesgo y brechas de pobreza bajo transf. logaritmo.
- Método de simulación MC para indicadores generales.
- Método bootstrap paramétrico para el ECM bajo el modelo.

✓ *Pfeffermann & Sverchkov (2007), JASA*
✓ *Berg, Cho, Eideh, Guadarrama & Molina (2023), Enviado*

PSEUDO EB

- El estimador óptimo de $F_{\alpha d}$ bajo el modelo con errores anidados depende de $\mathbf{y}_{ds}$ sólo a través de la media muestral $\bar{y}_{ds} = \frac{1}{n_d}\sum_{i\in s_d} Y_{di}$:

$$\tilde{F}^{B}_{\alpha d} = E_{\mathbf{y}_{dc}}\left[F_{\alpha d}|\mathbf{y}_{ds}\right] = E_{\mathbf{y}_{dc}}\left[F_{\alpha d}|\bar{y}_{ds}\right].$$

- **Estimador Pseudo Óptimo**: En la distrib. de $\mathbf{y}_{dc}|\bar{y}_{ds}$, reemplazar $\bar{y}_{ds}$ por el estimador de expansión (media ponderada) $\bar{y}_{dw}$:

$$\tilde{F}^{PB}_{\alpha d} = E_{\mathbf{y}_{dc}}\left[F_{\alpha d}|\bar{y}_{dw}\right].$$

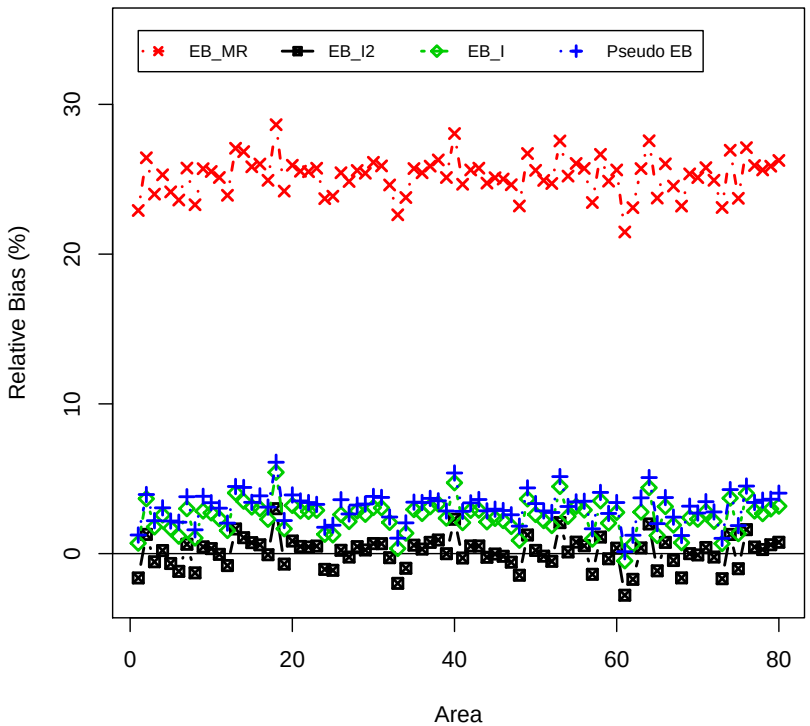
- Extensión del bootstrap paramétrico para la estimación del ECM del Pseudo EB bajo el modelo.

✓ *Guadarrama, Molina & Rao (2018), CSDA*

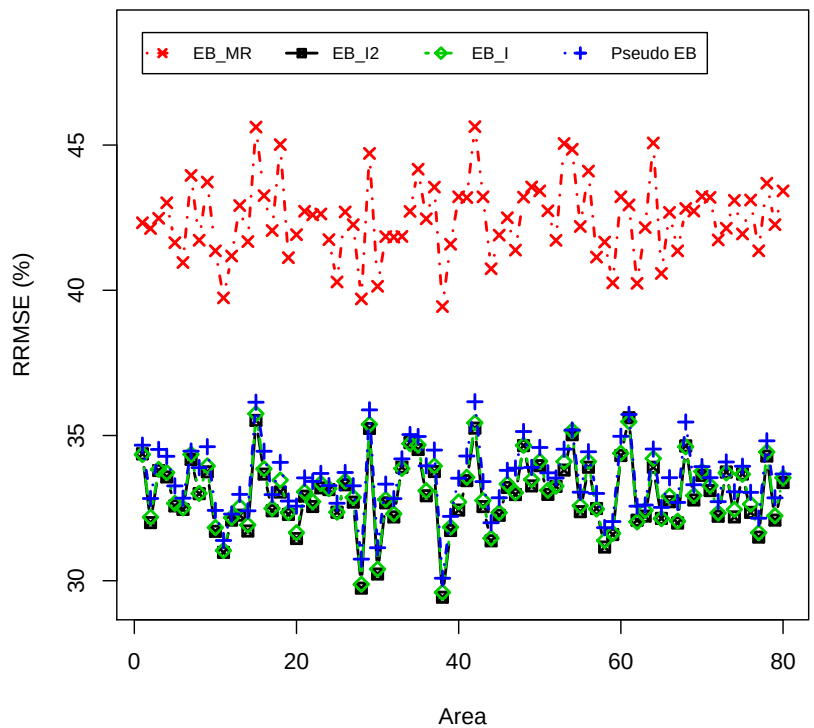
BRECHA POBREZA: MUESTREO INF

- Los estimadores que usan los pesos del diseño reducen el sesgo.

Sesgo (%)



ECM ($\times 10^4$)



MODELO DE MIXTURAS MULTIVARIATE

- Consideramos K grupos (latentes) de áreas con parámetros distintos $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k', \tau_k^2, \sigma_k^2)'$, $k = 1, \ldots, K$.
- Para $K = 2$ es un modelo para áreas atípicas.
- Modelo de mixtura multivariate:

$$\mathbf{y}_d \overset{ind.}{\sim} \sum_{k=1}^{K} \pi_k N(\mathrm{X}_d \boldsymbol{\beta}_k, \mathrm{V}_{kd}), \quad d = 1, \ldots, D.$$

donde $\pi_1 > \pi_2 > \cdots > \pi_K$, with $\sum_{k=1}^{K} \pi_k = 1$, y

$$\mathrm{V}_{kd} = \mathrm{V}_d(\tau_k^2, \sigma_k^2), \quad k = 1, \ldots, K,$$

✓ *Bikauskaite, Molina & Morales (2022), JRSSA*

MEDICIÓN POBREZA EN PALESTINA

- **Datos:** Encuesta de Consumo y Gasto Palestino (PECS) de 2016/2017 y Censo de Popublación Census de 2017.
- **Indicadores:** Tasas en riesgo y brechas de pobreza para localidades Palestinas.
- **Áreas:** En censo, 319 **localidades**$\rightarrow D = 162$ en encuesta. Estimamos en las localidades **muestreadas**.
- **Poder adquisitivo:** E_{dj} gasto mensual por adulto equivalente (ILS).
- **Umbral de pobreza:** $z = 10,027$ ILS $\rightarrow$ aprox. **26 %** bajo umbral.

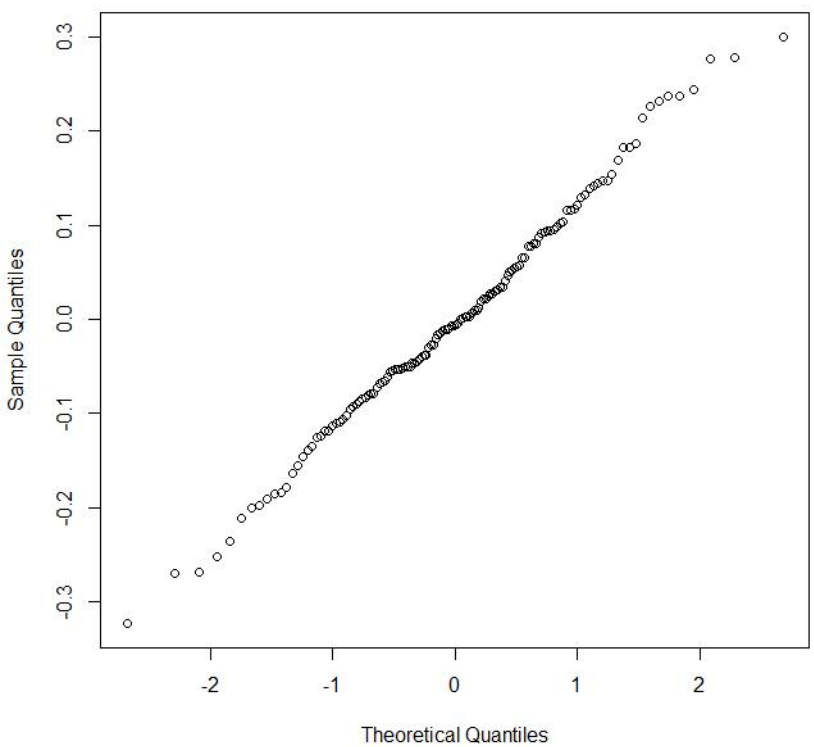
MODELO AJUSTADO

- **Variables explicativas:**
  - ✓ Indicadores de región (Gaza, Cisjordania), tipo de localidad (rural/urbana, campamento).
  - ✓ Características del hogar (tamaño, prop. mujeres, ratio de empleo).
  - ✓ Características de la persona de referencia del hogar (unemployed, employisrasett, employnatgov, refugstat, diff, neverschool, secondabove).
  - ✓ Características de la vivienda (tipo, tenencia, num. habitaciones).
  - ✓ Suministros (agua, basura, calefacción, nevera, etc.)

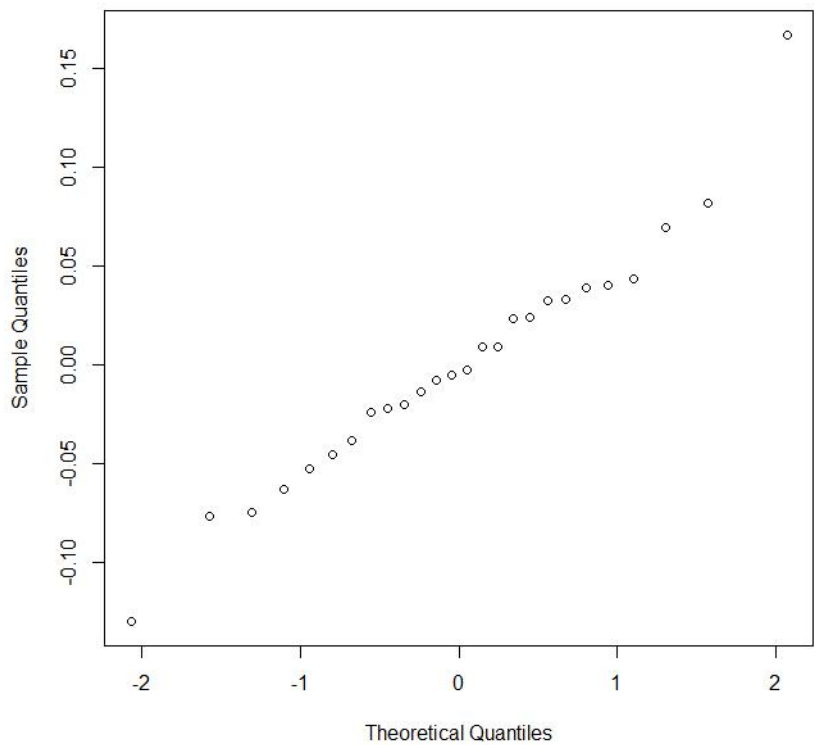
✓ *García-Portugués & Molina (2020), ESCWA.*

Gráficos QQ-normales de efectos de localidad predichos: Ajustes separados por región

Cisjordania

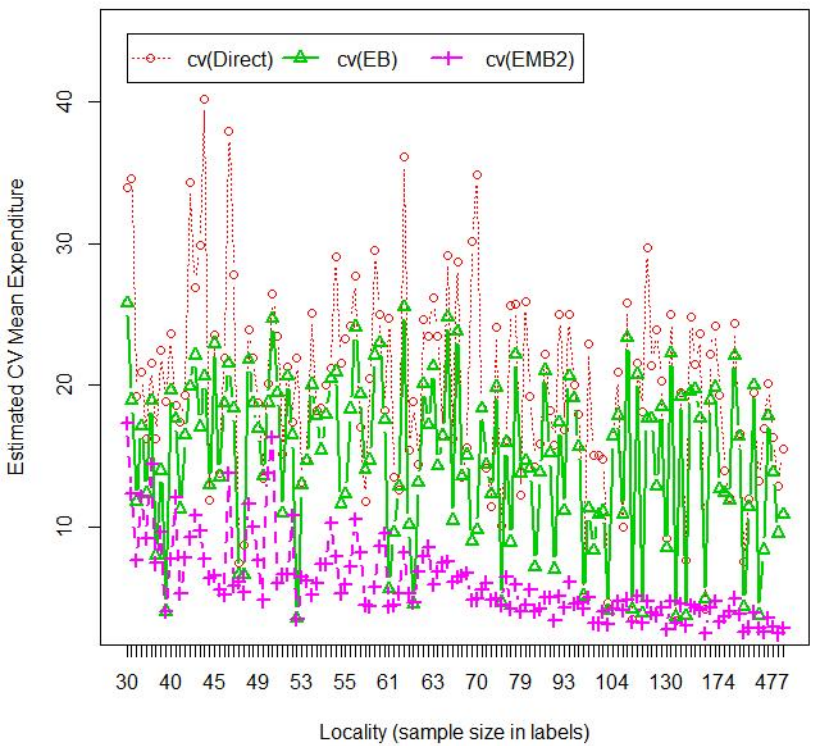


Gaza

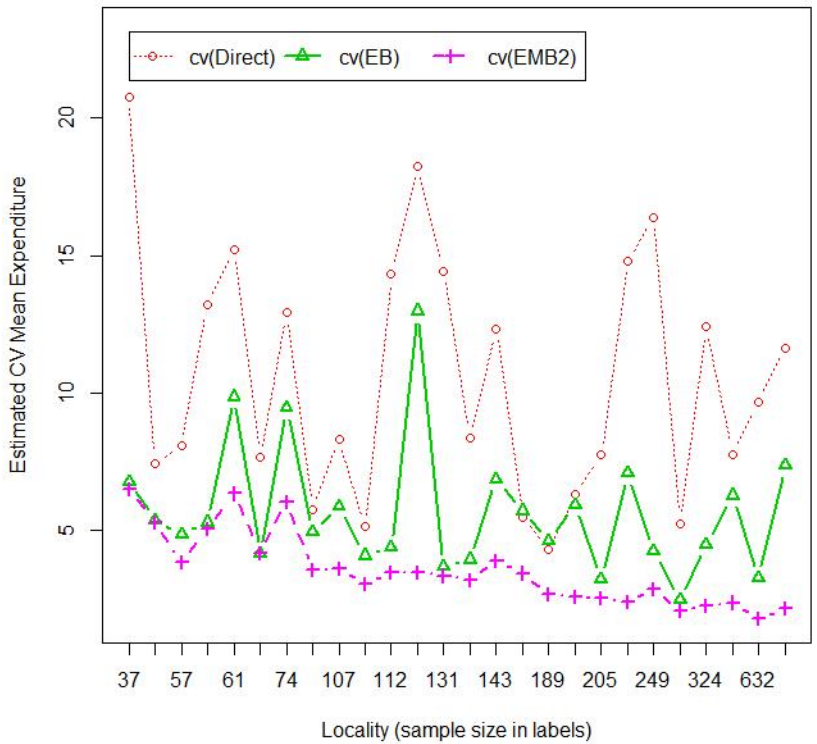


CV: GASTO MEDIO

Cisjordania

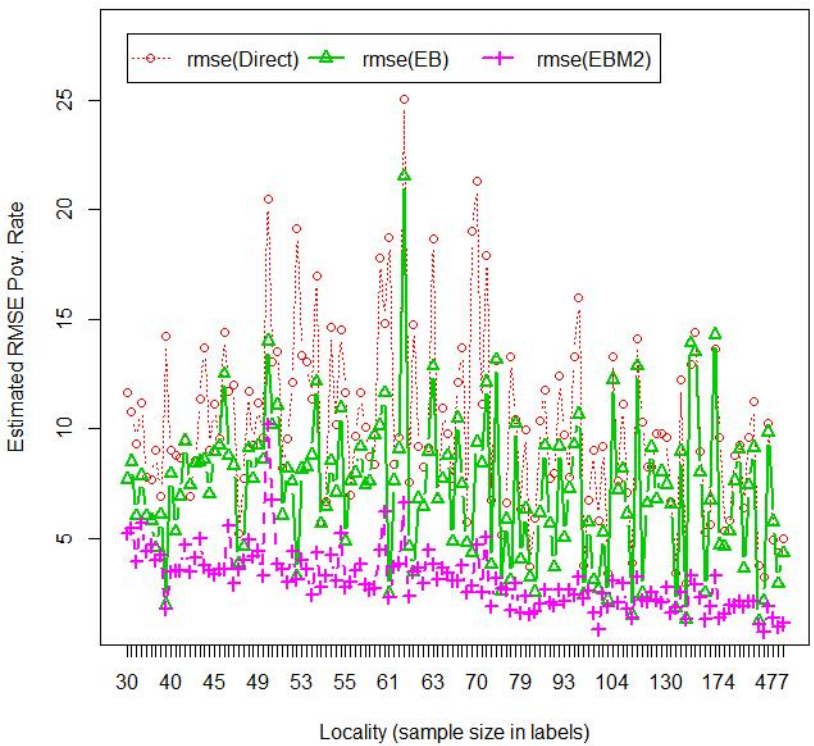


Gaza

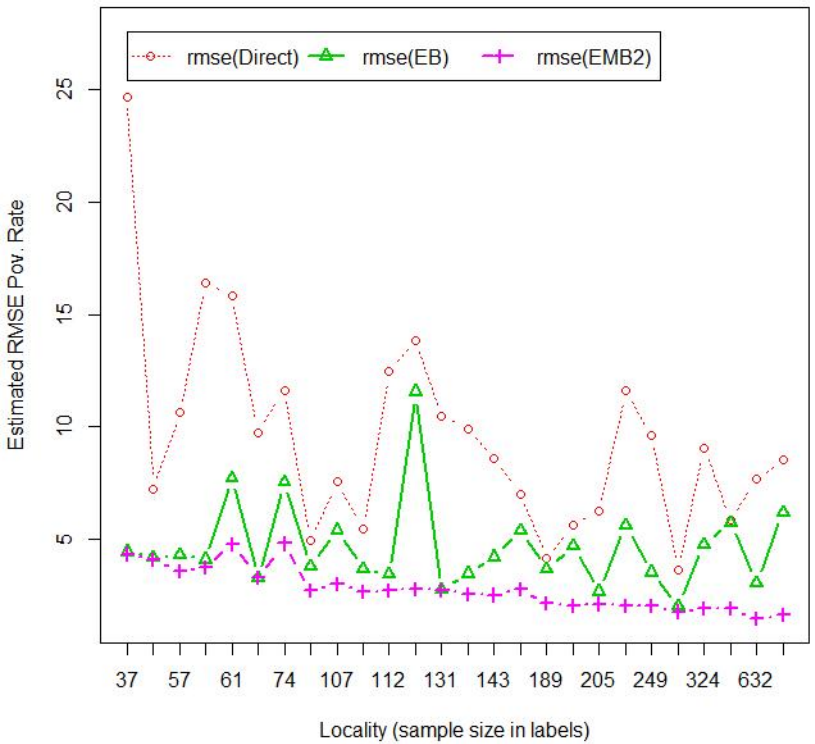


CV: RIESGO DE POBREZA

Cisjordania



Gaza



OTRAS EXTENSIONES DEL EB

- Versión **HB**: ✓ *Molina, Nandram & Rao (2014), AoAS*
- **Fast** EB: ✓ *Ferretti & Molina (2011), JISAS*
- EB bajo un modelo con errores anidados **a dos niveles**:
  ✓ *Marhuenda, Molina, Morales & Rao (2017), JRSSA*
- EB bajo distribuciones **asimétricas**:
  GB2: ✓ *Graf, Marín & Molina (2018), Test*
  Skew Normal: ✓ *Diallo & Rao (2018), Scand. J. Stat.*
- EB con **correlación temporal**:
  ✓ *Guadarrama, Morales & Molina (2020), CSDA*

PROBLEMAS ABIERTOS

- Estimación en **años intercensales**, cuando el último censo de variables auxiliares está **obsoleto**.
- Predictores óptimos bajo **falta de respuesta no ignorable**.
- Estimación eficiente del **ECM bajo el diseño** de los estimadores basados en modelos:
  ✓ *Molina & Strzalkowska-Kominiak (2020), JRSSA*
- **Métodos de selección de variables** explicativas específico para estimación en áreas pequeñas.
- Estimación de indicadores de pobreza **multidimensional**.
- Métodos de **machine-learning**.

La union hace la fuerza!!