

Multivariate area level models for small area estimation. ^a

Domingo Morales González

d.morales@umh.es

Universidad Miguel Hernández de Elche

^aIn collaboration with Roberto Benavent

1. Small Area Estimation.
2. Multivariate Fay-Herriot models.
3. Application to Spanish Living Condition Survey data.

- Official surveys are designed to obtain reliable estimates in **planned domains**.
- For example, The Spanish Living Condition Survey (SLCS) has sufficiently large sample sizes in the autonomous communities (planned domains).
- Then, the **direct estimators** have acceptably small mean squared errors in the autonomous communities .
- However, the SLCS sample sizes are too small within provinces (**unplanned domains** or **small areas**) and therefore the direct estimates are not reliable in these domains.

- **Small Area Estimation** is a branch of Statistics that gives procedures to improve the direct estimates in unplanned domains.
- We introduce small area estimators based on
 - Multivariate area-level mixed models.

- Let U be a finite population partitioned into D domains U_1, \dots, U_D .
- Let $\mu_d = (\mu_{d1}, \dots, \mu_{dR})'$ be a vector of characteristics of interest in the domain d .
- Let $y_d = (y_{d1}, \dots, y_{dR})'$ be a vector of direct estimators of μ_d .

The multivariate Fay-Herriot model is defined in two stages.

- The **sampling model** is

$$y_d = \mu_d + e_d, \quad d = 1, \dots, D, \quad (1)$$

- the vectors $e_d \sim N(0, V_{ed})$ are independent,
- the $R \times R$ covariance matrices V_{ed} are known.
- The **linking model** assumes that the μ_{dr} 's are linearly related to
 - $x_{dr} = (x_{dr1}, \dots, x_{drp_r})$ with p_r explanatory variables.
 - $x_d = \text{diag}(x_{d1}, \dots, x_{dR})_{R \times p}$ with $p = \sum_{r=1}^R p_r$.
 - $\beta = (\beta'_1, \dots, \beta'_r)'_{p \times 1}$.

- González-Manteiga et al. (2008b) considered the linking model

$$\mu_d = x_d\beta + 1_R v_d, \quad v_d \stackrel{ind}{\sim} N_1(0, \sigma_v^2), \quad d = 1, \dots, D, \quad (2)$$

where 1_n is the $n \times 1$ vector with all elements equal to 1.

- We introduce a **multivariate Fay-Herriot model** by assuming (1) and substituting the condition (2) by the more realistic linking model

$$\mu_d = x_d\beta + u_d, \quad u_d \stackrel{ind}{\sim} N_R(0, V_{ud}), \quad d = 1, \dots, D, \quad (3)$$

- the vectors u_d 's are independent of the vectors e_d 's.
- The $R \times R$ covariance matrices V_{ud} depend on m unknown parameters, $\theta_1, \dots, \theta_m$, with $1 \leq m \leq \frac{R(R-1)}{2} + R$.

We consider four particularizations of model (3).

- **Model 0** is the product of independent marginal models that assumes (1), (3) and takes

$$V_{e_d} = \text{diag}_{1 \leq r \leq R} (\sigma_{edr}^2), \quad V_{u_d} = \text{diag}_{1 \leq r \leq R} (\sigma_{ur}^2), \quad d = 1, \dots, D, \quad (4)$$

- the sampling error variances σ_{edr}^2 's are known,
 - $m = R$ and $\theta_r = \sigma_{ur}^2, r = 1, \dots, R$.
 - The components of e_d and u_d are independent under Model 0.
- **Model 1** assumes (1) and (3), with a known but not necessarily diagonal matrix V_e , and independent components of u_d , i.e.

$$V_{u_d} = \text{diag}_{1 \leq r \leq R} (\sigma_{ur}^2), \quad d = 1, \dots, D, \quad (5)$$

- $m = R$ and $\theta_r = \sigma_{ur}^2, r = 1, \dots, R$.
- Model 0 is Model 1 with V_e diagonal.

- Model 2 assumes (1), (3) with AR(1)-correlated u_d , i.e.

$$V_{ud} = \sigma_u^2 \Omega_d(\rho), \quad \Omega_d(\rho) = \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & \rho & \dots & \rho^{R-1} \\ \rho & 1 & \dots & \rho^{R-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{R-1} & \rho^{R-2} & \dots & 1 \end{pmatrix}, \quad (6)$$

- $m = 2, \theta_1 = \sigma_u^2, \theta_2 = \rho.$

- Model 3 assumes (1), (3) with HAR(1)-correlated u_d , i.e.

$$u_{dr} = \rho u_{dr-1} + a_{dr}, \quad u_{d0} \sim N(0, \sigma_0^2), \quad a_{dr} \stackrel{ind}{\sim} N(0, \sigma_r^2), \quad r = 1, \dots, R, \quad (7)$$

- $\sigma_0^2 = 1,$

- $u_{d0}, a_{dr}, r = 1, \dots, R,$ are independent,

- $m = R + 1$ and $\theta_1 = \sigma_1^2, \dots, \theta_R = \sigma_R^2, \theta_{R+1} = \rho.$

- We are interested in estimating small area poverty proportions and gaps by using data from the 2006 Spanish Living Condition Survey (SLCS).
- We calculate EBLUPs based on multivariate Fay-Herriot models.
- The target domains are the 52 Spanish provinces crossed by sex ($D = 104$).
- The target indicators are the poverty proportion ($\alpha = 0$) and gap ($\alpha = 1$),

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

- z is the poverty line and
 - E_{dj} is the equivalised net income of individual j within domain d , $j = 1, \dots, N_d$, $d = 1, \dots, D$.
- ▶ s is the global sample and s_d is the sample of domain d
 - ▶ The sample sizes are n and n_d respectively, so that
 - ▶ $s = \cup_{d=1}^D s_d$ and $n = \sum_{d=1}^D n_d$.

- The **direct estimator** of the domain **total** $Y_{dr} = \sum_{j=1}^{N_d} y_{drj}$ is

$$\hat{Y}_{dr}^{dir} = \sum_{j \in s_d} w_{dj} y_{drj},$$

where w_{dj} 's are the official calibrated sampling weights.

- The estimated domain size is $\hat{N}_d^{dir} = \sum_{j \in s_d} w_{dj}$.
- A **direct estimator** of the domain **mean** \bar{Y}_{dr} is $\bar{y}_{dr} = \hat{Y}_{dr}^{dir} / \hat{N}_d^{dir}$.
- The \bar{y}_{dr} 's are the responses in the area-level model.
- The design-based covariances of these estimators are approximated by

$$\widehat{\text{COV}}_{\pi}(\hat{Y}_{dr_1}^{dir}, \hat{Y}_{dr_2}^{dir}) = \sum_{j \in s_d} w_{dj}(w_{dj} - 1)(y_{dr_1j} - \bar{y}_{dr_1})(y_{dr_2j} - \bar{y}_{dr_2}),$$

$$\sigma_{\pi, d, r_1, r_2} = \widehat{\text{COV}}_{\pi}(\bar{y}_{dr_1}, \bar{y}_{dr_2}) = \widehat{\text{COV}}_{\pi}(\hat{Y}_{dr_1}^{dir}, \hat{Y}_{dr_2}^{dir}) / \hat{N}_d^2.$$

- We take the $\sigma_{\pi, d, r_1, r_2}$'s as the known elements of the matrix V_{ed} in the multivariate Fay-Herriot models.

- The available **auxiliary variables** are the domain proportions of people in the categories of the following classification variables:
 - **Age** ($age1$: ≤ 15 , $age2$: $16 - 24$, $age3$: $25 - 49$, $age4$: $50 - 64$, $age5$: ≥ 65),
 - **Education** ($edu0$: less than primary, $edu1$: primary, $edu2$: secondary, $edu3$: university),
 - **Citizenship** ($cit1$: Spanish, $cit2$: not Spanish),
 - **Labor situation** ($lab0$: ≤ 15 , $lab1$: employed, $lab2$: unemployed, $lab3$: inactive).
- As the proportions of people in the categories of a given variable sum up to one, we take the reference categories out of the auxiliary data file.
- The reference categories are $age5$, $edu3$, $cit2$ and $lab3$.

We present two applications.

- The **first application** jointly estimates 2006 poverty proportions and gaps for provinces crossed by sex.
- The **second application** jointly estimates 2005 and 2006 poverty proportions for provinces crossed by sex.

- For jointly estimating 2006 poverty proportions ($\alpha = 0$) and gaps ($\alpha = 1$), we fit Model 3 to a subset of auxiliary variables.

Variables	<i>constant</i>	<i>age1</i>	<i>age2</i>	<i>edu1</i>	<i>cit1</i>	<i>lab2</i>
β_1	-0.70357	0.95490	1.45541	0.74745	0.30873	1.50050
<i>p</i> -value	0.00000	0.00066	0.00165	0.00000	0.00137	0.00006

Table 1. Regression parameters and *p*-values for Model 3, $\alpha = 0$, 2006.

Variables	<i>constant</i>	<i>edu0</i>	<i>edu1</i>	<i>edu2</i>	<i>cit1</i>	<i>lab1</i>
β_2	-0.37458	0.97049	0.34255	0.16551	0.152031	-0.06384
<i>p</i> -value	0.00001	0.00000	0.00001	0.11197	0.00104	0.02502

Table 2. Regression parameters and *p*-values for Model 3, $\alpha = 1$, 2006.

- By observing the signs of the regression parameters we conclude that provinces having larger proportions of population in categories *age1*, *age2*, *edu1*, *cit1* and *lab2* have greater poverty proportion.
- On the other side, provinces having larger proportions of population in categories *edu0*, *edu1*, *edu2*, and *cit1* and smaller proportions of population in the category *lab1* have greater poverty gaps.

- The estimates of the variance component parameters are $\hat{\sigma}_{u1}^2 = 0.00138$, $\hat{\sigma}_{u2}^2 = 0.00037$ and $\hat{\rho} = 0.01859$.
- We test $H_0 : \sigma_{u1}^2 = \sigma_{u2}^2$. The test statistics is

$$T_{12} = \frac{\hat{\sigma}_{u1}^2 - \hat{\sigma}_{u2}^2}{\sqrt{\nu_{11} + \nu_{22} - 2\nu_{12}}} = 3.34588,$$

- ν_{ij} , $i, j = 1, 2, 3$ are the elements of the inverse of the REML Fisher information matrix of Model 3 evaluated at $\hat{\theta} = (\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\rho})$.
- As $T_{12} \underset{asym}{\sim} N(0, 1)$ under H_0 , the p -value is 0.00082.
- We conclude that random effects variances are different and we prefer Model 3 instead of Model 2.
- We also test $H_0 : \rho = 0$. The test statistics is $T_\rho = \frac{\hat{\rho}}{\sqrt{\nu_{33}}} = 1.96464$.
 - As $T_\rho \underset{asym}{\sim} N(0, 1)$ under H_0 , the p -value is 0.049456.
 - We conclude that both components (poverty proportion and gap) are positively correlated and we prefer Model 3 instead of Model 1.

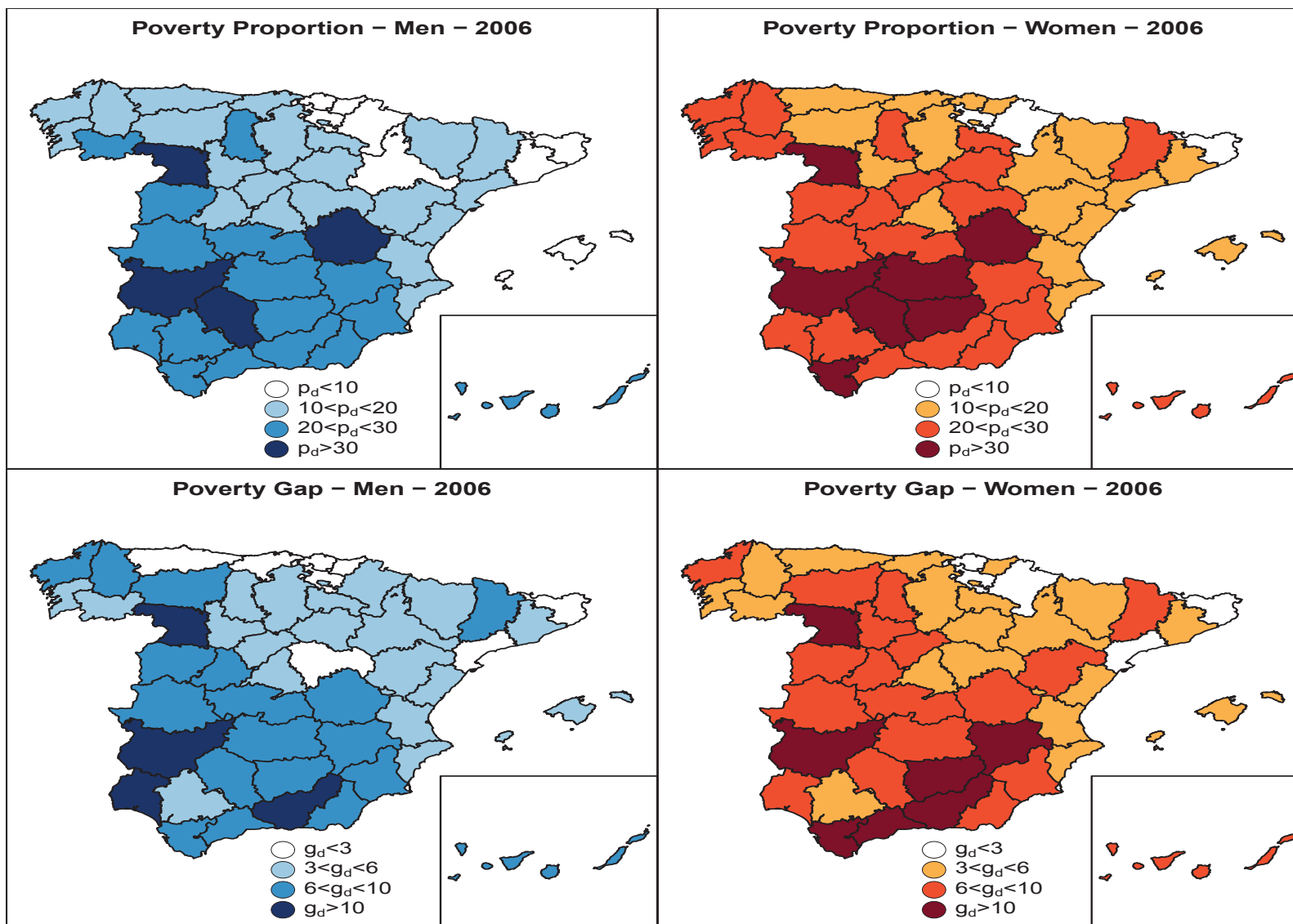


Figure 1. Poverty proportions (top) and gaps (bottom) for men (left) and women (right) in Spanish provinces during 2006.

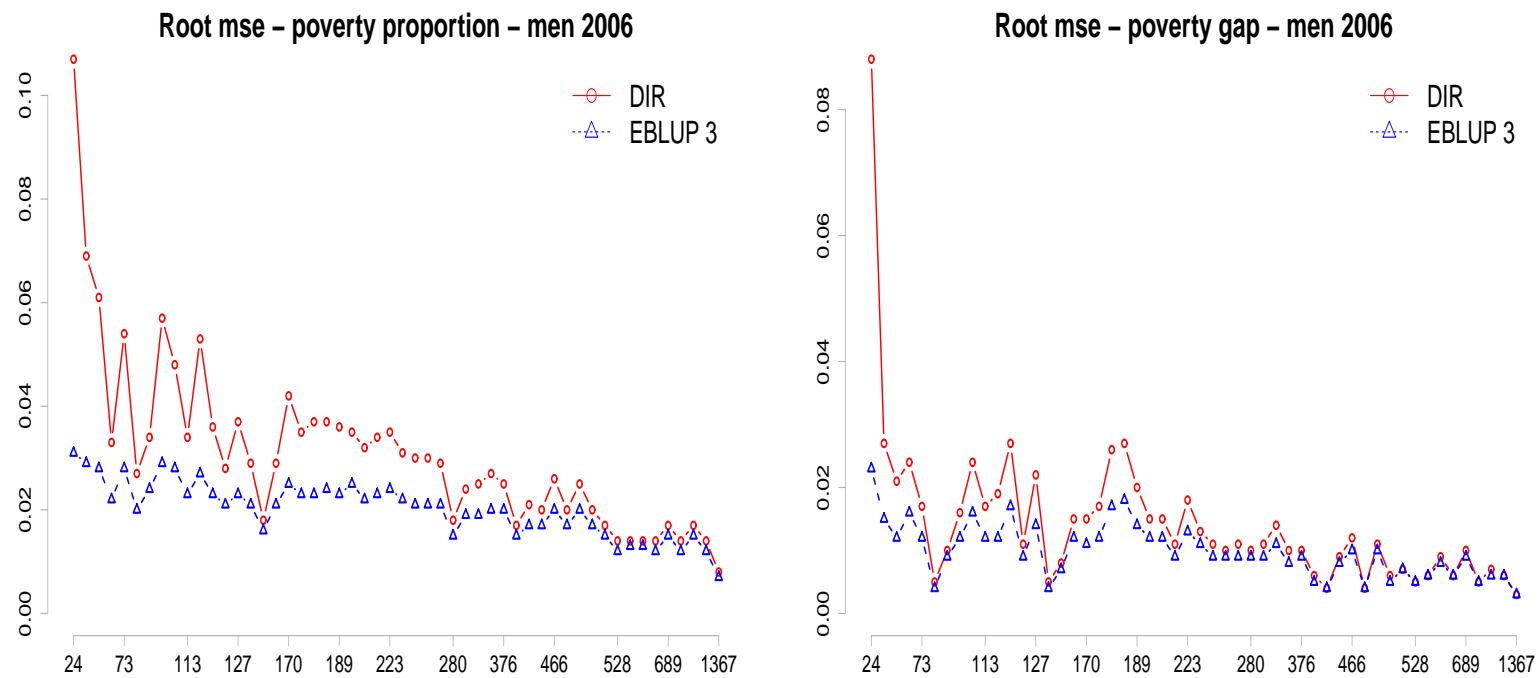


Figure 2. Root-MSEs of direct and EBLUP (under Model 3) estimators of poverty proportions (left) and gaps (right) in Spanish provinces during 2006.

- For jointly estimating 2005 and 2006 poverty proportions ($\alpha = 0$), we fit Model 3 to a subset of auxiliary variables.

Variables	<i>constant</i>	<i>age1</i>	<i>age2</i>	<i>edu1</i>	<i>cit1</i>	<i>lab2</i>
β	-0.65428	0.69780	2.38240	0.71074	0.25924	0.71268
<i>p</i> -value	0.00010	0.06540	0.00049	0.00000	0.08960	0.15129

Table 3. Regression parameters and *p*-values for Model 3, $\alpha = 0$, **2005**.

Variables	<i>constant</i>	<i>age1</i>	<i>age2</i>	<i>edu1</i>	<i>cit1</i>	<i>lab2</i>
β	-0.75278	0.88497	1.89752	0.79734	0.31471	2.04460
<i>p</i> -value	0.00000	0.00609	0.00047	0.00000	0.00414	0.00000

Table 4. Regression parameters and *p*-values for Model 3, $\alpha = 0$, **2006**.

- By observing the signs of the regression parameters we conclude that provinces having larger proportions of population in categories *age1*, *age2*, *edu1*, *cit1* and *lab2* have greater poverty proportion in 2005 and 2006.

- The estimates of the variance component parameters are $\hat{\sigma}_{u1}^2 = 0.00256$, $\hat{\sigma}_{u2}^2 = 0.00193$ and $\hat{\rho} = 0.02105$.
- We test $H_0 : \sigma_{u1}^2 = \sigma_{u2}^2$. The test statistics is

$$T_{12} = \frac{\hat{\sigma}_{u1}^2 - \hat{\sigma}_{u2}^2}{\sqrt{\nu_{11} + \nu_{22} - 2\nu_{12}}} = 1.0756,$$

- where ν_{ij} , $i, j = 1, 2, 3$ are the elements of the inverse of the REML Fisher information matrix of Model 3 evaluated at $\hat{\theta} = (\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\rho})$.
- As $T_{12} \underset{asym}{\sim} N(0, 1)$ under H_0 , the p -value is 0.28208.
- We cannot conclude that random effects variances are different and we prefer Model 2 instead of Model 3.
- Therefore, we fit Model 2 to the subset of auxiliary variables.

Variables	<i>constant</i>	<i>age1</i>	<i>age2</i>	<i>edu1</i>	<i>cit1</i>	<i>lab2</i>
β_{2005}	-0.53822	0.67365	1.74785	0.60288	0.23672	0.99025
<i>p</i> -value	0.00040	0.03876	0.00209	0.00000	0.08998	0.02351
β_{2006}	-0.74083	0.90128	1.69006	0.68294	0.37468	1.78575
<i>p</i> -value	0.00000	0.00595	0.00127	0.00000	0.00163	0.00007

Table 5. Regression parameters and *p*-values for Model 2 and $\alpha = 0$.

- We test $H_0 : \rho = 0$ under model 2. The test statistics is

$$T_\rho = \frac{\hat{\rho}}{\sqrt{\nu_{22}}} = 16.72633,$$

- ν_{ij} , $i, j = 1, 2$ are the elements of the inverse of the REML Fisher information matrix of Model 2 evaluated at $\hat{\theta} = (\hat{\sigma}^2, \hat{\rho})$.
- As $T_\rho \underset{asym}{\sim} N(0, 1)$ under H_0 , the *p*-value is 0.00.
- We conclude that both components (2005 and 2006 poverty proportions) are positively correlated and we prefer Model 2 instead of Model 1.

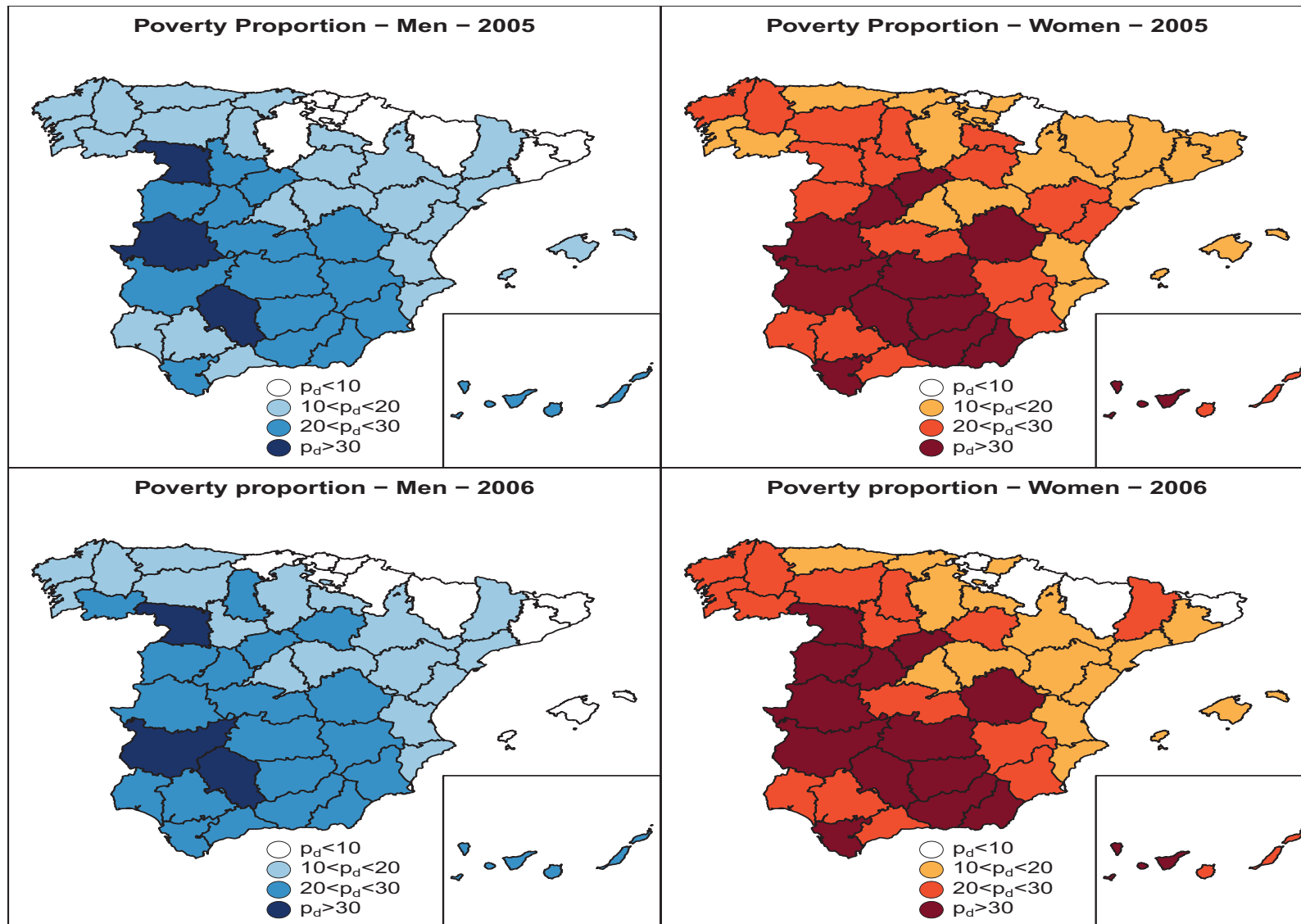


Figure 3. Poverty proportions in 2005 (top) and 2006 (bottom) for men (left) and women (right) in Spanish provinces during 2006.

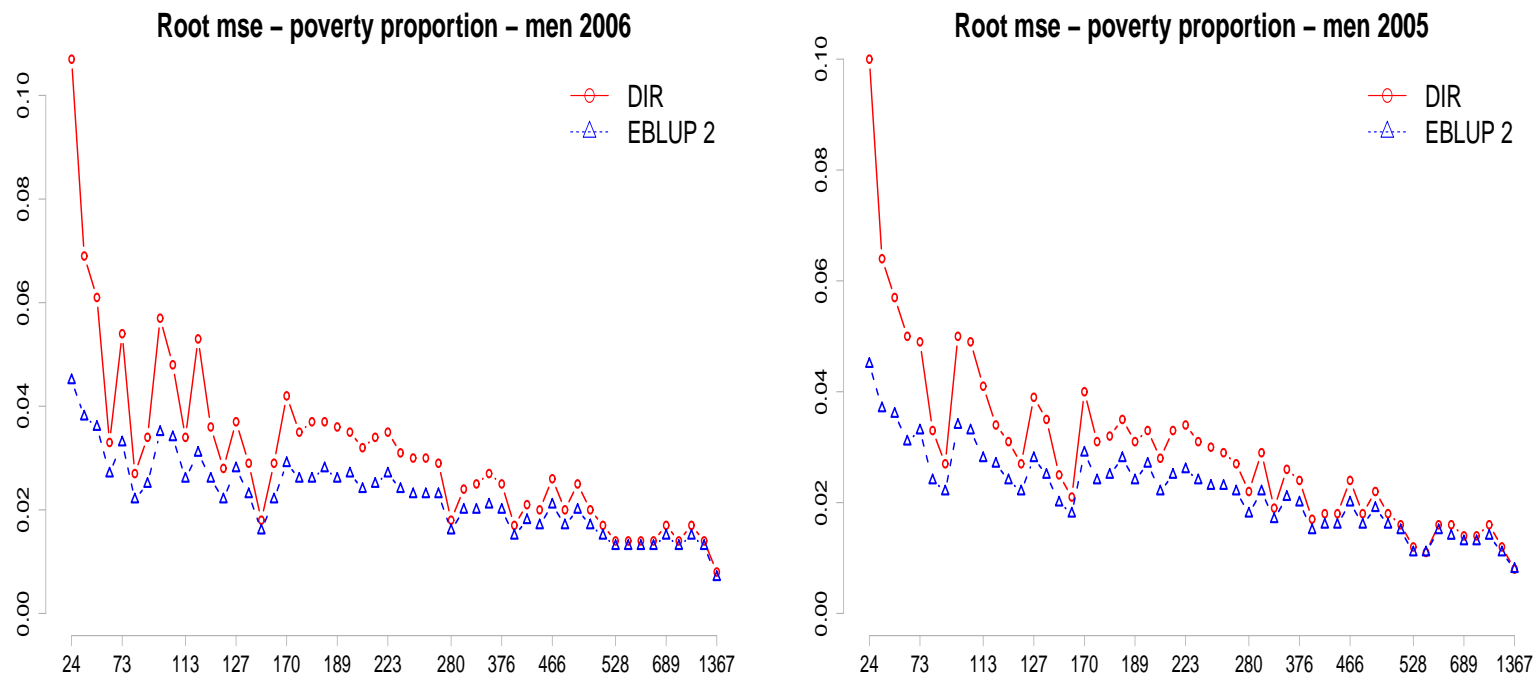


Figure 4. Root-MSEs of direct and EBLUP (under Model 2) estimators of poverty proportions for 2006 (left) and 2005 (right) in Spanish provinces.

- Benavent R., Morales D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis*, 94, 372-390.
- Esteban, M.D., Morales, D., Pérez, A., Santamaría, L., 2011. Two Area-Level Time Models for Estimating Small Area Poverty Indicators. *Journal of the Indian Society of Agricultural Statistics*, 66, 11, 75-89.
- Esteban, M.D., Morales, D., Pérez, A., Santamaría, L., 2012. Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, 56, 2840-2855.
- Fay, R.E., Herriot, R.A., 1979. Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74, 269-277.
- González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., Santamaría, L., 2008b. Analytic and Bootstrap Approximations of Prediction Errors under a Multivariate Fay-Herriot Model. *Computational Statistics and Data Analysis*, 52 (12), 5242-5252.
- Särndal C.E., Swensson B., Wretman J., 1992. Model assisted survey sampling. Springer-Verlag.

Thank you
for your attention