Third International Conference on Agricultural Statistics -MEXSAI-

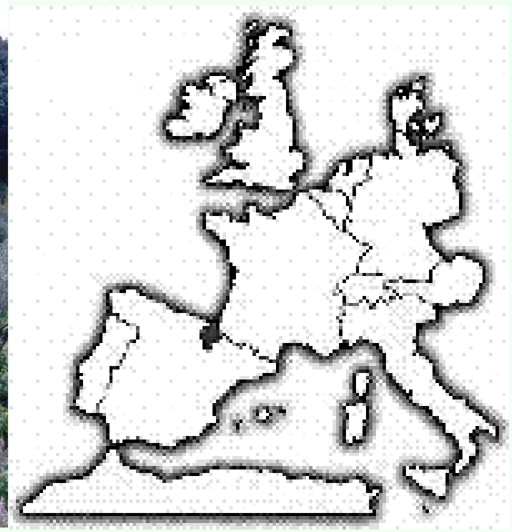
Using Small Area Models to Estimate the Total Area Occupied by Olive Trees

Militino, A.F., Ugarte, M.D. and Goicoa, T.

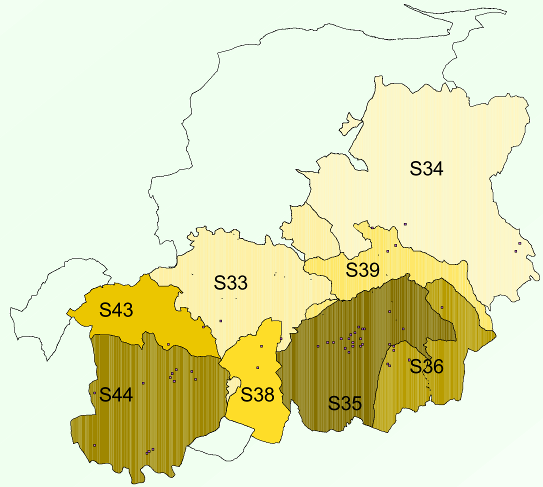
Cancún, 2-4 November 2004

Motivation

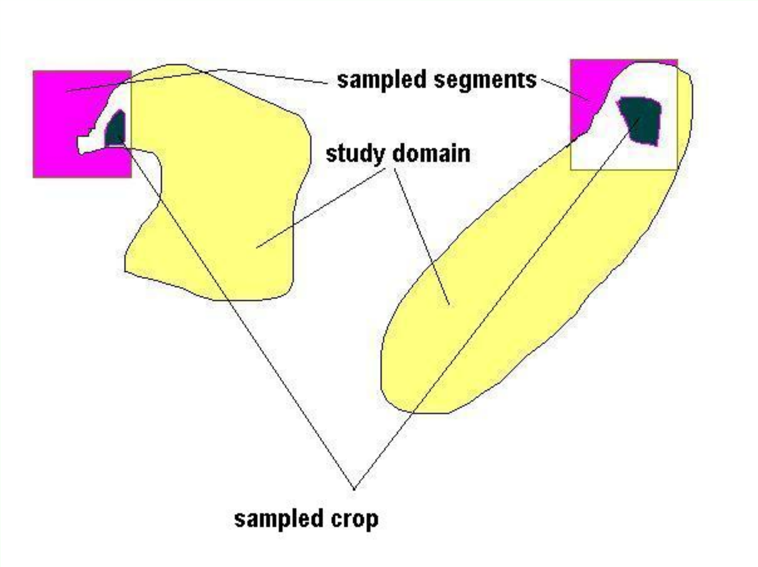
- **To obtain reliable estimates of olive trees in Navarra (Spain)**



- **Small and irregular plots $\Longrightarrow$ domestic consumption**
- **Olive oil is very important in the Mediterranean diet**
- **Development of a modern industry**
- **Sampling process very difficult and expensive**
- **Design based estimators are not appropriate**
- **Model based methods $\Longrightarrow$ Small Area Estimation (Rao, Wiley 2003)**



- **Sample: 39 segments of 4 hectares in 8 non irrigated areas**
- **Plots very irregular and different in size and dispersion**



- **Irregular study domain**
- **Size of sample segments limited by satellite images**
- **Transformation of data**

Goals

- To provide estimates of the small area totals of surface occupied by olive trees
- To provide standard errors of the small area estimators
- To include weights to correct for heteroscedasticity
- To include sampling weights to obtain design-consistent estimators
- To compare the performance of different small area models

Introduction

- **Increasing demand for precise estimates in domains with small sample size**
  - **To produce reliable estimates**
  - **To assess the estimation error**
  - **Specificity:** *borrow information*

$$\Downarrow$$

$$\boxed{\textbf{Auxiliary information} \begin{cases} \textbf{Related areas} \\ \textbf{Past data} \end{cases}}$$

- **Agricultural applications**
  - **Linear Mixed Models (Battese, Harter and Fuller, JASA 1988)**
  - **Auxiliary information: Data provided by satellite images**
  - **Regular segments**

Heteroscedastic Unit Level Model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, n_i$$

- $u_{ij} = v_i + e_{ij}, \quad v_i \sim N(0, \sigma_v^2)$ **y** $e_{ij} \sim N(0, \sigma_e^2/c_{ij})$
- v_i **are assumed to be independent of the random errors** e_{ij}
- y_{ij}**: number of hectares of olive trees in the** j**th segment of the** i**th area**
- n_i **is the number of sampled segments**
- x_{ij}**: number of classified hectares of olive trees in the** j**th segment of the** i**th area**
- c_{ij}**: weights to account for heteroscedasticity**

- **In matrix form**

$$\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \boldsymbol{\epsilon},\;\; \mathbf{v} \sim N(\mathbf{0}, \sigma_v^2 \mathbf{I}_t),\;\; \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{C}^{-1})} \quad (1)$$

- **Quantity of interest**

$$\boxed{\bar{y}_{i(p)} = \bar{\mathbf{x}}'_{i(p)}\boldsymbol{\beta} + v_i = \beta_0 + \beta_1 \bar{x}_{i(p)} + v_i}$$

- **Predictor**

$$\boxed{\hat{\bar{y}}_{ic} = \bar{\mathbf{x}}'_{i(p)}\hat{\boldsymbol{\beta}}_c + \hat{v}_{ic} = \bar{\mathbf{x}}'_{i(p)}\hat{\boldsymbol{\beta}}_c + \hat{\gamma}_{ic}(\bar{y}_{ic} - \bar{\mathbf{x}}'_{ic}\hat{\boldsymbol{\beta}}_c)}$$

$\boxed{\hat{\gamma}_{ic}}$ **is the plug-in estimador of** $\boxed{\gamma_{ic} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/c_{i.})}$

Area Level Model

- **Extension of the Prasad and Rao (Survey M., 1999) area level model**
- **Combining Equation (1) and the design estimators**

$$\boxed{\bar{y}_{iw} = \sum_{j=1}^{n_i} w_{ij} y_{ij}, \quad \bar{\mathbf{x}}_{iw} = \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij}}$$

where $w_{ij} = \tilde{w}_{ij} / \sum_{j=1}^{n_i} \tilde{w}_{ij}$ **and** $\tilde{w}_{ij}$ **are the sampling weights. Then,**

$$\boxed{\bar{\mathbf{Y}}_w = \bar{\mathbf{X}}_w \boldsymbol{\beta} + \mathbf{v} + \bar{\boldsymbol{\epsilon}}_w, \ \mathbf{v} \sim N(\mathbf{0}, \sigma_v^2 \mathbf{I}_t), \ \bar{\boldsymbol{\epsilon}}_w \sim N(\mathbf{0}, \sigma_e^2 \boldsymbol{\delta}_c^2)} \quad (2)$$

$$\boldsymbol{\delta}_c^2 = \mathbf{diag}(\delta_{ic}^2); \quad \delta_{ic}^2 = \sum_{j=1}^{n_i} w_{ij}^2 / c_{ij}, \ i = 1, \ldots, t$$

- **Predictor**

$$\boxed{\hat{\bar{y}}_{iwc} = \bar{\mathbf{x}}'_{i(p)}\hat{\boldsymbol{\beta}}_{wc} + \hat{v}_{iwc} = \bar{\mathbf{x}}'_{i(p)}\hat{\boldsymbol{\beta}}_{wc} + \hat{\gamma}_{iwc}(\bar{y}_{iw} - \bar{\mathbf{x}}'_{iw}\hat{\boldsymbol{\beta}}_{wc})}$$

$\boxed{\hat{\gamma}_{iwc}}$ **is the plug-in estimador of** $\boxed{\gamma_{iwc} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2\delta_{ic}^2)}$

- **The estimator is design-consistent assuming**

$$\boxed{\delta_{ic}^2 \longrightarrow 0} \quad \textbf{as} \quad \boxed{n_i \longrightarrow \infty}$$

Extended Pseudo-EBLUP

- **Extension of the You and Rao (Canadian J. Statistics, 2002) Pseudo-EBLUP**

Steps

1. **Assume $\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2$ are known in the area level model (2). Then, the BLUP is**

$$\tilde{\bar{y}}_{iwc} = \bar{\mathbf{x}}'_{i(p)}\boldsymbol{\beta} + \gamma_{iwc}(\bar{y}_{iw} - \bar{\mathbf{x}}'_{iw}\boldsymbol{\beta})$$

2. **The variance components are estimated from the heteroscedastic unit level model (1)**

3. **Obtain the BLUP of v_{iwc} from Expression(2)**

$$\tilde{v}_{iwc} = \gamma_{iwc}(\bar{y}_{iw} - \bar{\mathbf{x}}'_{iw}\boldsymbol{\beta})$$

Solving the weighted estimating equations

$$\sum_{i=1}^{t}\sum_{j=1}^{n_i} \tilde{w}_{ij} c_{ij} \mathbf{x}_{ij}[y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta} - \tilde{v}_{iwc}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2)] = \mathbf{0}$$

it is obtained

$$\boxed{\hat{\boldsymbol{\beta}}_{wcYR} = \left\{\sum_{i=1}^{t}\sum_{j=1}^{n_i} \tilde{w}_{ij} c_{ij} \mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\gamma}_{iwc}\bar{\mathbf{x}}_{iw})'\right\}^{-1} \left\{\sum_{i=1}^{t}\sum_{j=1}^{n_i} \tilde{w}_{ij} c_{ij} \mathbf{x}_{ij}(y_{ij} - \hat{\gamma}_{iwc}\bar{y}_{iw})\right\}}$$

- **Predictor**

$$\boxed{\hat{\bar{y}}_{iwcYR} = \bar{\mathbf{x}}'_{i(p)}\hat{\boldsymbol{\beta}}_{wcYR} + \hat{\gamma}_{iwc}(\bar{y}_{iw} - \bar{\mathbf{x}}'_{iw}\hat{\boldsymbol{\beta}}_{wcYR})}$$

- **The estimator is design-consistent assuming**

$$\boxed{\delta^2_{ic} \longrightarrow 0} \quad \textbf{as} \quad \boxed{n_i \longrightarrow \infty}$$

Variance Components Estimation

- **Fitting of constants (Searle, Casella and McCullogh, Wiley 1992).**

$$\boxed{\hat{\sigma}_e^2 = \frac{1}{n-t-k} \sum_{i=1}^{t} \sum_{j=1}^{n_i} c_{ij} \hat{\epsilon}_{ij}^2}$$

$\hat{\epsilon}_{ij}^2$ **: weighted regression of Y on X introducing v as a dummy variable**

$$\boxed{\hat{\sigma}_v^2 = \mathbf{max} \left(\frac{1}{n_{*c}} \left\{ \sum_{i=1}^{t} \sum_{j=1}^{n_i} c_{ij} \hat{s}_{ij}^2 - (n-k-1)\hat{\sigma}_e^2 \right\}, 0 \right)}$$

$\hat{s}_{ij}^2$ **: residuals from the weighted regression of Y on X**

Mean Squared Error

- **Kackar and Harville, JASA 1984, showed, under normality**

$$\boxed{\mathbf{MSE}[\hat{\tilde{y}}_{i(p)}(\hat{\boldsymbol{\sigma}}^2, \mathbf{Y})] = \mathbf{MSE}[\tilde{\tilde{y}}_{i(p)}(\boldsymbol{\sigma}^2, \mathbf{Y})] + E[\hat{\tilde{y}}_{i(p)}(\hat{\boldsymbol{\sigma}}^2, \mathbf{Y}) - \tilde{\tilde{y}}_{i(p)}(\boldsymbol{\sigma}^2, \mathbf{Y})]^2}$$

- **An adequate estimator (Prasad and Rao, JASA 1990)**

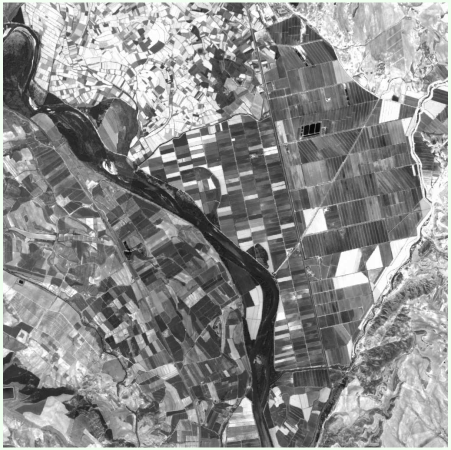
$$\boxed{\widehat{\mathbf{MSE}}[t_i(\hat{\boldsymbol{\sigma}}^2, \mathbf{Y})] = g_{1ic}(\hat{\boldsymbol{\sigma}}^2) + g_{2ic}(\hat{\boldsymbol{\sigma}}^2) + 2g_{3ic}(\hat{\boldsymbol{\sigma}}^2)}$$

- g_{1ic} **is associated to random effects**
- g_{2ic} **is associated to fixed effects**
- g_{3ic} **is associated to variance components**

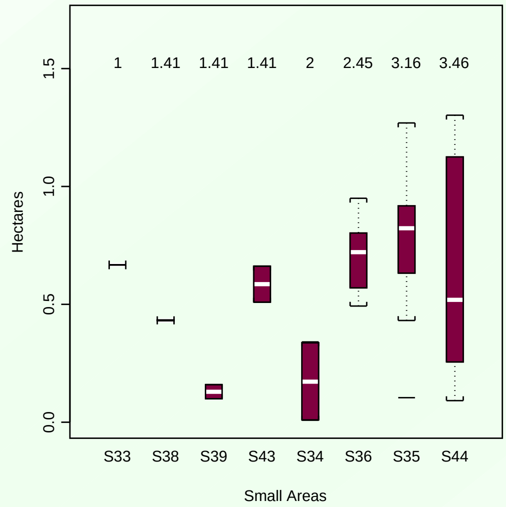
Application

- **Complex project: several scientific disciplines**
- **Study domain determined by a Navarra map and aerial photos**

- **Auxiliary information: satellite images**
  - **Two kind of images: panchromatic and multispectral**
  - **New methods of merging images**



Sampled Segments



- **Variability increases with sample size**
- **Weights:** $c_{ij} = 1/\sqrt{n_i},\ i = 1, \ldots, t;\ j = i, \ldots, n_i$

Models

- **Model 1: Homoscedastic unit level model,** $c_{ij} = 1,\ \tilde{w}_{ij} = 1$.
- **Model 2: Heteroscedastic unit level model** $c_{ij} = 1/\sqrt{n_i},\ \tilde{w}_{ij} = 1$.
- **Model 3: Area level model (Prasad y Rao, Survey Methodology, 1999).** $c_{ij} = 1,\ \tilde{w}_{ij} = N_i/n_i$.
- **Modelo 4: Area level model.** $c_{ij} = 1/\sqrt{n_i},\ \tilde{w}_{ij} = N_i/n_i$.
- **Model 5: Pseudo-EBLUP estimator (You y Rao, Canadian J. Statistics, 2002).** $c_{ij} = 1,\ \tilde{w}_{ij} = N_i/n_i$.
- **Modelo 6: Extended Pseudo-EBLUP .** $c_{ij} = 1/\sqrt{n_i},\ \tilde{w}_{ij} = N_i/n_i$.

Table 1. Results for Unit Level Models

				Model 1 $(c_{ij} = 1)$			**Model 2** $(c_{ij} = 1/\sqrt{n_i})$		
Area	n_i	N_i	S_i	$\hat{y}_{iw}$	$s.e.$	$c.v$	$\hat{y}_{iw}$	$s.e.$	$c.v$
S33	**1**	**32**	**26.560**	**10.380**	**3.389**	**0.326**	**13.593**	**3.598**	**0.265**
S38	**2**	**97**	**87.199**	**34.839**	**10.455**	**0.300**	**39.940**	**9.682**	**0.242**
S39	**2**	**115**	**170.224**	**31.543**	**12.516**	**0.397**	**26.525**	**11.491**	**0.433**
S43	**2**	**81**	**67.010**	**31.557**	**8.722**	**0.276**	**40.053**	**8.090**	**0.202**
S34	**4**	**227**	**226.286**	**67.084**	**24.301**	**0.362**	**50.143**	**20.112**	**0.401**
S36	**6**	**284**	**280.085**	**125.992**	**29.460**	**0.234**	**135.801**	**24.075**	**0.177**
S35	**10**	**697**	**791.867**	**400.333**	**63.608**	**0.159**	**413.477**	**51.769**	**0.125**
S44	**12**	**731**	**935.936**	**347.611**	**64.120**	**0.184**	**349.560**	**53.449**	**0.153**
Total	**39**	**2264**	**2585.168**	**1049.339**	**99.846**	**0.095**	**1069.092**	**82.615**	**0.077**

Table 2. Results for Area Level Models

				Model 3 $(c_{ij} = 1)$			**Model 4** $(c_{ij} = 1/\sqrt{n_i})$		
Area	n_i	N_i	S_i	$\hat{y}_{i\tilde{w}}$	$s.e.$	$c.v$	$\hat{y}_{i\tilde{w}}$	$s.e.$	$c.v$
S33	**1**	**32**	**26.560**	**9.997**	**3.927**	**0.393**	**13.303**	**3.941**	**0.296**
S38	**2**	**97**	**87.199**	**33.989**	**11.345**	**0.334**	**39.623**	**9.841**	**0.248**
S39	**2**	**115**	**170.224**	**30.367**	**13.923**	**0.458**	**26.098**	**11.732**	**0.450**
S43	**2**	**81**	**67.010**	**30.860**	**9.441**	**0.306**	**39.767**	**8.243**	**0.207**
S34	**4**	**227**	**226.286**	**65.428**	**25.776**	**0.394**	**49.231**	**20.735**	**0.421**
S36	**6**	**284**	**280.085**	**123.584**	**31.999**	**0.259**	**133.139**	**28.227**	**0.212**
S35	**10**	**697**	**791.867**	**397.182**	**65.674**	**0.165**	**409.336**	**56.620**	**0.138**
S44	**12**	**731**	**935.936**	**342.442**	**69.500**	**0.203**	**343.391**	**63.430**	**0.185**
Total	**39**	**2264**	**2585.168**	**1033.851**	**106.107**	**0.103**	**1053.889**	**93.669**	**0.089**

Table 3. Results for Pseudo-EBLUP Estimators

				Model 5 $(c_{ij} = 1)$			Model 6 $(c_{ij} = 1/\sqrt{n_i})$		
Area	n_i	N_i	S_i	$\hat{y}_{i\tilde{w}}$	$s.e.$	$c.v$	$\hat{y}_{i\tilde{w}}$	$s.e.$	$c.v$
S33	**1**	**32**	**26.560**	**9.965**	**3.403**	**0.342**	**13.213**	**3.605**	**0.273**
S38	**2**	**97**	**87.199**	**33.744**	**10.487**	**0.311**	**39.085**	**9.695**	**0.248**
S39	**2**	**115**	**170.224**	**30.203**	**12.557**	**0.416**	**25.499**	**11.506**	**0.451**
S43	**2**	**81**	**67.010**	**30.646**	**8.749**	**0.285**	**39.334**	**8.101**	**0.206**
S34	**4**	**227**	**226.286**	**64.914**	**24.355**	**0.375**	**48.446**	**20.136**	**0.416**
S36	**6**	**284**	**280.085**	**123.481**	**29.522**	**0.239**	**133.524**	**24.119**	**0.181**
S35	**10**	**697**	**791.867**	**395.825**	**63.696**	**0.161**	**409.161**	**51.836**	**0.127**
S44	**12**	**731**	**935.936**	**342.791**	**64.228**	**0.187**	**344.797**	**53.542**	**0.155**
Total	**39**	**2264**	**2585.168**	**1031.569**	**100.015**	**0.097**	**1053.060**	**82.740**	**0.079**

- **Diagnosis: it is very important to check model hypothesis**
  - **Significance of the variance components: a parametric bootstrap test is conducted**
  - **Normality: it is a necessary condition to estimate the mean squared error**
  - **There are some simulation studies to show the robustness of the models to small deviations from normality when the variance components are estimated by the fitting of constants method**
- **It is possible to use standard software such as SAS, S-PLUS, R to fit small area models, but extra programming is needed to obtain the small area predictor and the mean squared error**

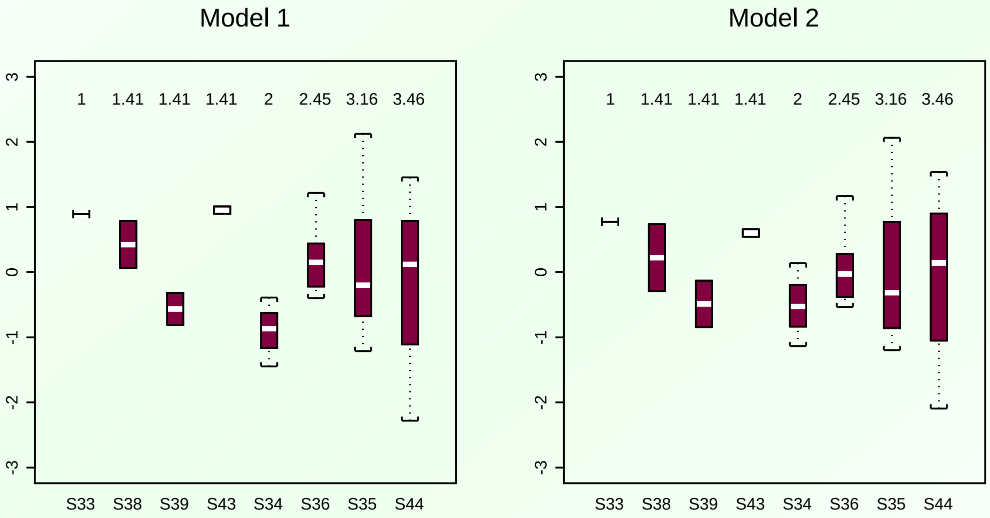
Table 4. Variance components estimates, their standard errors and parametric bootstrap test p-value

	Fitting of Constants				
Model	$\hat{\sigma}_e^2$	$s.e.(\hat{\sigma}_e^2)$	$\hat{\sigma}_v^2$	$s.e.(\hat{\sigma}_v^2)$	**Bootstrap p-value**
Model 1, 3 and 5	**0.051**	**0.013**	**0.005**	**0.010**	**0.164**
Model 2, 4 and 6	**0.016**	**0.004**	**0.015**	**0.013**	**0.019**

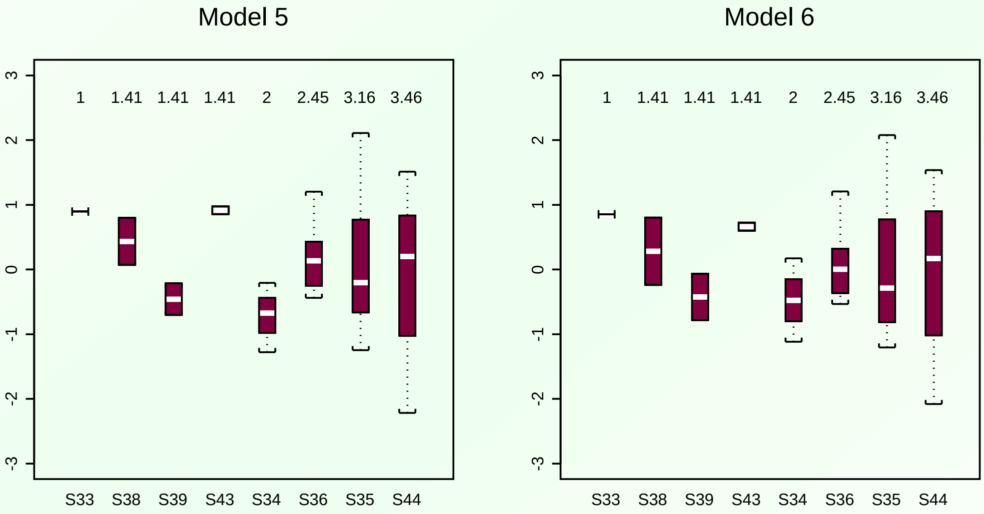
Table 5. p-value of the Shapiro-Wilk statistic for testing the normality of the residuals

	Shapiro-Wilk p-value	
Model	**Transformed residuals**	**Eblup residuals**
Model 1	**0.998**	**0.993**
Model 2	**0.857**	**0.989**
Model 3	**0.704**	**—**
Model 4	**0.862**	**—**
Model 5	**—**	**0.993**
Model 6	**—**	**0.994**

Unit level models. Boxplots of residuals



Pseudo-EBLUP estimators. Boxplots of residuals



Conclusions

- There is a claer necessity of using specific methodologies to obtain accurate estimates in small areas
- We provide small area model that use model weights to correct for heteroscedasticity and sampling weights to obtain design consistency.
- We obtain good results in the real application considered here.