Uniform prediction intervals in SAE

Katarzyna Reluga
joint work with Stefan Sperlich and Maria José Lombardía

Université de Genève

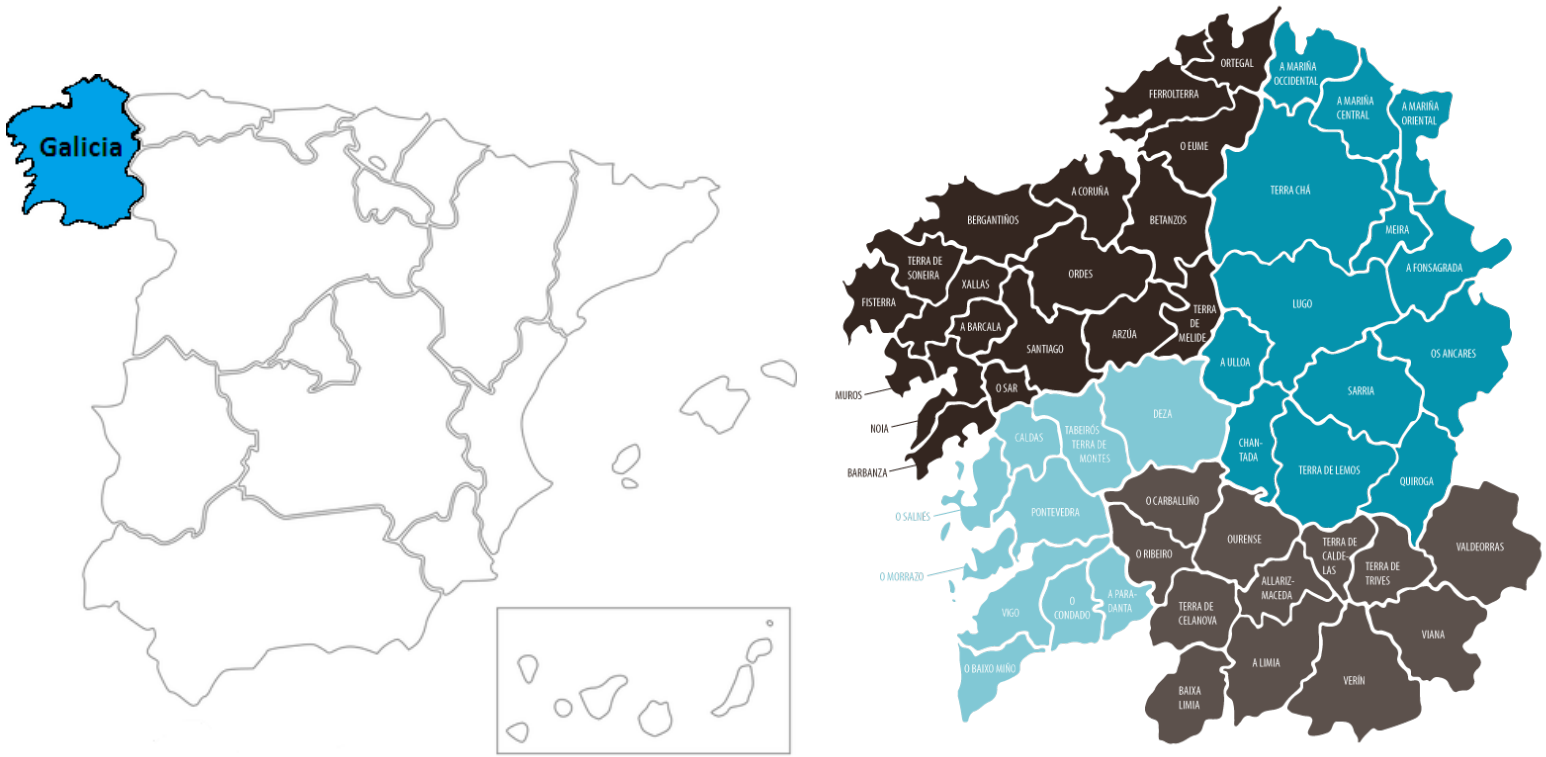
Zurich, 28th August 2018

Table of contents

Outline

Small Area Estimation

Small **A**rea **E**stimation (SAE) - a family of statistical methods which provides tools for an inference and a prediction for *small areas*, i.e. any entity (a district, a school etc.) for which the direct estimates are not feasible due to the large standard errors (or poor precision).



Countour map of Spain and Galicia.
Source: www.getdrawings.com and www.galiciaenteira.com.

Inference and prediction in SAE

Focus of SAE

- Provides a **reliable estimate** $\hat{\mu}_d$ of the **small area statistic**, $d = 1, ..., D$, i.e. a mean, a count, a ratio, or any other functional if the distribution of area means is estimated.
- Provides a **measure of the variability** of $\hat{\mu}_d$, the most often $\mathrm{MSE}(\hat{\mu}_d)$ and/or **prediction intervals**.
- $\hat{\mu}_d$ is a linear combination (or a function of it) of **fixed** and **random** effects.

Outline

Motivation

SAE allows us to quantify a validity of the policy intervention, health care or environmental projects etc. $\rightarrow$ one **compares simultaneously** small area statistics among all (or some subset of interest) and chooses those which satisfy certain criteria.

Classical pointwise prediction intervals (PPI) at 95% level

1. Comparison of PPIs is **not valid statistically** as we ignore a part of the variability which arises from a simultaneous statement.
2. Coverage probabilities of pointwise prediction intervals are averaged over the areas giving an "approximate picture".
3. **At least** 5% of μ_d are **out of** their **PPI** — effect mainly for areas with **extreme** area effects..
4. The discussion on simultaneous inference in SAE has hardly started, only Ganesh (2009) created **simultaneous Bayesian credible intervals**.

Valid simultaneous inference

Goal

Construction of the tool for simultaneous inference which enables multiple comparisons and provides uniform coverage probability over the areas of interest.

Our tool

MSE based ***S****imultaneous (uniform)* ***P****rediction* ***I****ntervals* (SPI) from frequentist perspective for small area parameters under Liner Mixed Model:

(a) analytical derivation based on the volume of tube formula

(b) Monte Carlo approximation of (a)

(c) bootstrap estimation

Modeling Framework

- Consider LMM: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$.
- Conditional covariance matrix of $\boldsymbol{y}$: $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^t$, $\boldsymbol{\theta}$ variance parameter, e.g. under NERM $\boldsymbol{\theta} = (\sigma_e^2, \sigma_u^2)$.
- BLUE for $\boldsymbol{\beta}$: $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}$.
- BLUP for $\boldsymbol{u}$: $\tilde{\boldsymbol{u}}(\boldsymbol{\theta}) = \boldsymbol{G}\boldsymbol{Z}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$.
- *Empirical* counterparts $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{u}}$ with $\hat{\boldsymbol{\theta}}$ are estimated using MM, ML, REML, Henderson method etc.

Target parameter: $\bar{Y}_d = N_d^{-1}\sum_{j=1}^{N_d} y_{dj}$

$$\bar{Y}_d \approx \mu_d = \bar{\boldsymbol{X}}_d^t\boldsymbol{\beta} + \bar{\boldsymbol{Z}}_d^t\boldsymbol{u}_d$$
$$\Downarrow$$
EBLUP
$$\Downarrow$$
$$\hat{\mu}_d = \bar{\boldsymbol{X}}_d^t\hat{\boldsymbol{\beta}} + \bar{\boldsymbol{Z}}_d^t\hat{\boldsymbol{u}}_d$$

Outline

General construction of SPI in SAE

Object of interest: **a confidence region** $I_{1-\alpha}$ such that
$P(\mu_d \in I_{1-\alpha} \; \forall d \in \{1, \ldots, D\}) = 1 - \alpha$

- Equivalent with choosing a critical value $c_{1-\alpha}$ such that

$$\begin{aligned} \alpha &= \mathrm{P}\left(|\hat{\mu}_d - \mu_d| > c_{1-\alpha} \cdot \mathrm{MSE}^{1/2}(\hat{\mu}_d) \; \forall d \in \{1, \ldots, D\}\right) \\ &= \mathrm{P}\left(\sup_{1 \leqslant d \leqslant D} \left| \frac{\hat{\mu}_d - \mu_d}{\mathrm{MSE}^{1/2}(\hat{\mu}_d)} \right| > c_{1-\alpha}\right) \end{aligned}$$

- Estimate MSE and approximate quantile $c_{1-\alpha}$ from the distribution of

$$S_D = \sup_{1 \leqslant d \leqslant D} \left| \frac{\hat{\mu}_d - \mu_d}{\mathrm{MSE}^{1/2}(\hat{\mu}_d)} \right|$$

- Construct $I^S_{1-\alpha}$ based on an order statistic of S_D

$$I^S_{1-\alpha} : \left[\hat{\mu}_d - c_{1-\alpha} \times \mathrm{MSE}^{1/2}(\hat{\mu}_d), \hat{\mu}_d + c_{1-\alpha} \times \mathrm{MSE}^{1/2}(\hat{\mu}_d)\right] \; \forall d \in \{1, \ldots, D\}$$

Analytical derivation based on the volume of tube I

Following Sun and Loader (1994), a reasonable formula for the prediction bands for small area means in LMM is

$$\begin{aligned}
\alpha &= \mathrm{P}\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \frac{\left|\boldsymbol{x}^t\boldsymbol{\beta} + \boldsymbol{z}^t\boldsymbol{u} - \hat{\boldsymbol{l}}^t\boldsymbol{y}\right|}{\hat{\sigma}_e||\hat{\boldsymbol{l}}_M||} > c_{1-\alpha} \right) \\
&\leqslant \mathrm{P}\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \left\{ \left[\left| \frac{\langle \boldsymbol{l}_M, \boldsymbol{e}_M \rangle}{\sigma_e||\boldsymbol{l}_M||} \right| + \frac{\eta(\boldsymbol{x})}{\sigma_e||\boldsymbol{l}_M||} \right] \frac{||\boldsymbol{l}_M||}{||\hat{\boldsymbol{l}}_M||} \right\} > c_{1-\alpha}\frac{\hat{\sigma}_e}{\sigma_e} \right) \\
&\leqslant \mathrm{P}\left(\sup_{\boldsymbol{x}\in\mathcal{X}} |\mathcal{Z}| > c_{1-\alpha}\frac{\hat{\sigma}_e}{\sigma_e}\xi - \eta \right) = 2\mathrm{P}\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \mathcal{Z} > c_{1-\alpha}\frac{\hat{\sigma}_e}{\sigma_e}\xi - \eta \right)
\end{aligned}$$

where $\xi = \inf_{\boldsymbol{x}\in\mathcal{X}} \frac{||\hat{\boldsymbol{l}}_M||}{||\boldsymbol{l}_M||}$ and $\eta = \sup_{\boldsymbol{x}\in\mathcal{X}} \frac{\eta(\boldsymbol{x})}{\sigma_e||\boldsymbol{l}_M||}$ with $\eta(\boldsymbol{x})$ a complex remainder.

For consistent $\hat{\boldsymbol{\theta}}$ we have $\xi = 1 + o_p(1)$ and $\eta = o_p(1)$ as $n \to \infty$.

Analytical derivation based on the volume of tube II

Proposition

Suppose that σ_e^2 is estimated by some consistent estimator and that $\exists\ \xi_0 > 0$, $\eta_0 > 0$ such that $\mathrm{P}(\xi \leqslant \xi_0) = o(\alpha)$ and $\mathrm{P}(\eta \geqslant \eta_0) = o(\alpha)$ as $n \to \infty$ and $\alpha \to 0$.
Then, for $\boldsymbol{p = 1}$, α can be approximated by:

$$\alpha \approx \frac{\kappa_0}{\pi}\left[\left(1+\frac{c_{1-\alpha}^2\xi_0^2}{\nu}\right)^{-\nu/2} + \eta_0\frac{2^{1/2}c_{1-\alpha}\xi_0\Gamma\left(\frac{\nu+1}{2}\right)}{\nu^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{c_{1-\alpha}^2\xi_0^2}{\nu}\right)^{-(\nu+1)/2}\right]$$
$$+\,\mathcal{E}\cdot P(|t_\upsilon| \geqslant c_{1-\alpha}\xi_0)$$

where t_ν is a t-distributed r.v. with ν d.o.f., $\kappa_0 = \int_{x\in\mathcal{X}} \|\mathcal{Q}'(\boldsymbol{x})\|d\boldsymbol{x}$ with $\mathcal{Q} = \frac{\boldsymbol{l}_M}{\|\boldsymbol{l}_M\|}$ and $\mathcal{E}$ is the Euler-Poincaré characteristic of the manifold $\mathcal{M} = \{\mathcal{Q}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\}$.

Remarks:

- ξ_0 and η_0 needs to be estimated and geometrical constants — approximated numerically.
- For $p = 2$ and $p \geqslant 3$ the expressions contain more constants to approximate.
- **The volume of tube approximation is not operational and involves substantial numerical and large sample approximation errors**.

Monte Carlo Approximation

Idea

Use Monte Carlo methods to approximate the distribution of the random vector of
$$\hat{\mu}_d - \mu_d = \bar{\boldsymbol{X}}_d^t \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \bar{\boldsymbol{Z}}_d^t \left(\hat{\boldsymbol{u}}_d - \boldsymbol{u}_d\right)$$

- Under the assumption of normality, it follows that
$$\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{u}} - \boldsymbol{u} \end{bmatrix} \approx N\left[\mathbf{0}, \left(\boldsymbol{C}^t \boldsymbol{R}^{-1} \boldsymbol{C} + \boldsymbol{G}^{0+}\right)^{-1}\right]$$
- Define $\bar{\boldsymbol{C}}_d = (\bar{\boldsymbol{X}}_d^t, \bar{\boldsymbol{Z}}_d^t)^t$ and approximate S_D by its simulated counterpart $\hat{S}_D$
$$S_D = \max_{d=1,\ldots,D} \frac{|\hat{\mu}_d - \mu_d|}{\mathrm{MSE}^{1/2}(\hat{\mu}_d)} \approx \max_{d=1,\ldots,D} \frac{\left|\bar{\boldsymbol{C}}_d^t \begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{u}}_d - \boldsymbol{u}_d \end{bmatrix}\right|}{\mathrm{MSE}^{1/2}(\hat{\mu}_d)} =: \hat{S}_D$$
- Construct $I_{1-\alpha}^{MC}$ using an order statistic $\hat{c}_{1-\alpha}$ of $\hat{S}_D$

$$I_{1-\alpha}^{MC} : \left[\hat{\mu}_d - \hat{c}_{1-\alpha} \times \mathrm{MSE}^{1/2}(\hat{\mu}_d), \hat{\mu}_d + \hat{c}_{1-\alpha} \times \mathrm{MSE}^{1/2}(\hat{\mu}_d)\right] \quad \forall d \in \{1, \ldots, D\}$$

Bootstrap Approximation

Idea

Use bootstrapping to approximate the distribution of the statistic S

$$S_D^{*(b)} = \max_{d=1,\ldots,D} \left| \frac{\hat{\mu}_d^{*(b)} - \mu_d^{*(b)}}{\mathrm{MSE}^{*1/2}(\hat{\mu}_d)} \right|$$

- Construct $I_{1-\alpha}^B$ based on an order statistic of S_D^*

$$I_{1-\alpha}^B : \left[\hat{\mu}_d - c_{1-\alpha}^* \times \mathrm{MSE}^{1/2}(\hat{\mu}_d), \hat{\mu}_d + c_{1-\alpha}^* \times \mathrm{MSE}^{1/2}(\hat{\mu}_d)\right] \quad \forall d \in \{1, \ldots, D\}$$

- Coverage probability of $I_{1-\alpha}^B$ is asymptotically correct

Theorem

Under necessary regularity conditions it holds that

$$P\left(\mu_d \in I_{1-\alpha}^B \ \forall d \in 1, \ldots, D\right) = 1 - \alpha + O(D^{-1})$$

Outline

Estimation of MSE

Classical decomposition

$$\mathrm{MSE}[\hat{\mu}_d] = \mathrm{MSE}[\tilde{\mu}_d] + \mathrm{E}[\hat{\mu}_d - \tilde{\mu}_d]^2 + 2 \cdot \mathbb{E}\left[(\tilde{\mu}_d - \mu_d)(\hat{\mu}_d - \tilde{\mu}_d)\right]$$

- $\mathrm{MSE}[\tilde{\mu}_d]$ accounts for the variability of μ_d when $\boldsymbol{\theta}$ known; in LMM

$$\mathrm{MSE}[\tilde{\mu}_d] = g_{1d}(\boldsymbol{\theta}) + g_{2d}(\boldsymbol{\theta})$$

  with $g_{1d}(\boldsymbol{\theta})$ and $g_{2d}(\boldsymbol{\theta})$ explicit in $\boldsymbol{X}$, $\boldsymbol{V}$, $\boldsymbol{Z}$, $\boldsymbol{G}$

- Generally, two last terms in above decomposition are intractable.
- We can estimate $\mathrm{MSE}[\hat{\mu}_d]$ using:
  1. Taylor approximation;
  2. different types of bootstrap estimators (approximate the whole expression vs. approximate the intractable terms, correct for bias etc.);
  3. various bootstrap schemes are available (i.e. parametric, wild).

Outline

Simulation settings

- LMM - NERM: $\beta_0 = 1$, $\beta_1 = 1$, $x_{dj} \sim U[0,1]$, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$.
- $n_d = 5, 10, 15$, $D = 25, 50, 75$, different distributions for u_d, e_{dj}.
- $B = 1000$ bootstrap samples with $C = 1$ for double bootstrap.
- 500 simulation runs for evaluating 95% PPI & SPI.
- Results only for random effect bootstrap (REB).

Remarks:

- Since BLUP is unbiased under the assumption of symmetry of F_e and F_u, most methods work only for symmetric distributions
- Strong disturbance if F_e is not normal (CLT); F_u matters less

MC SPI

Empirical coverage probability using $I^{MC}_{1-\alpha}$ and different MSE estimators.
The nominal coverage probability is 95%.

			mse_L		mse^*_{B1}		mse^*_{BC2}		mse^*_{3T}		mse^*_{SP}		mse^*_{SPA}	
e_{dj}	u_d	D/n_d	25/5	75/15	25/5	75/15	25/5	75/15	25/5	75/15	25/5	75/15	25/5	75/15
$t_6(0.5)$	$N(1)$	REML	89.6	92.0	89.4	92.2	89.8	91.8	89.4	92.4	89.6	92.0	89.4	92.0
$t_6(0.5)$	$N(1)$	MM	89.6	92.2	88.8	92.6	88.8	92.6	89.2	92.4	89.6	92.2	89.8	92.2
$\chi^2_5(1)$	$N(0.5)$	REML	89.8	90.4	90.4	90.6	90.8	90.2	90.4	90.6	89.8	90.2	89.8	90.4
$\chi^2_5(1)$	$N(0.5)$	MM	89.6	90.6	90.6	90.4	90.2	90.2	90.2	90.6	89.6	90.6	89.6	90.4
$N(1)$	$N(0.5)$	REML	92.4	94.6	92.2	94.4	91.6	94.6	91.6	94.4	92.4	94.6	92.4	94.6
$N(1)$	$N(0.5)$	MM	92.6	94.8	91.6	94.8	92.0	95.6	92.0	95.0	92.6	94.8	92.6	94.8
$t_6(1)$	$t_6(0.5)$	REML	87.8	91.2	87.8	91.6	88.8	91.8	88.2	91.6	87.8	91.2	87.4	90.8
$t_6(1)$	$t_6(0.5)$	MM	88.2	91.0	88.0	91.8	89.0	92.8	87.8	91.8	88.2	91.0	87.8	91.2

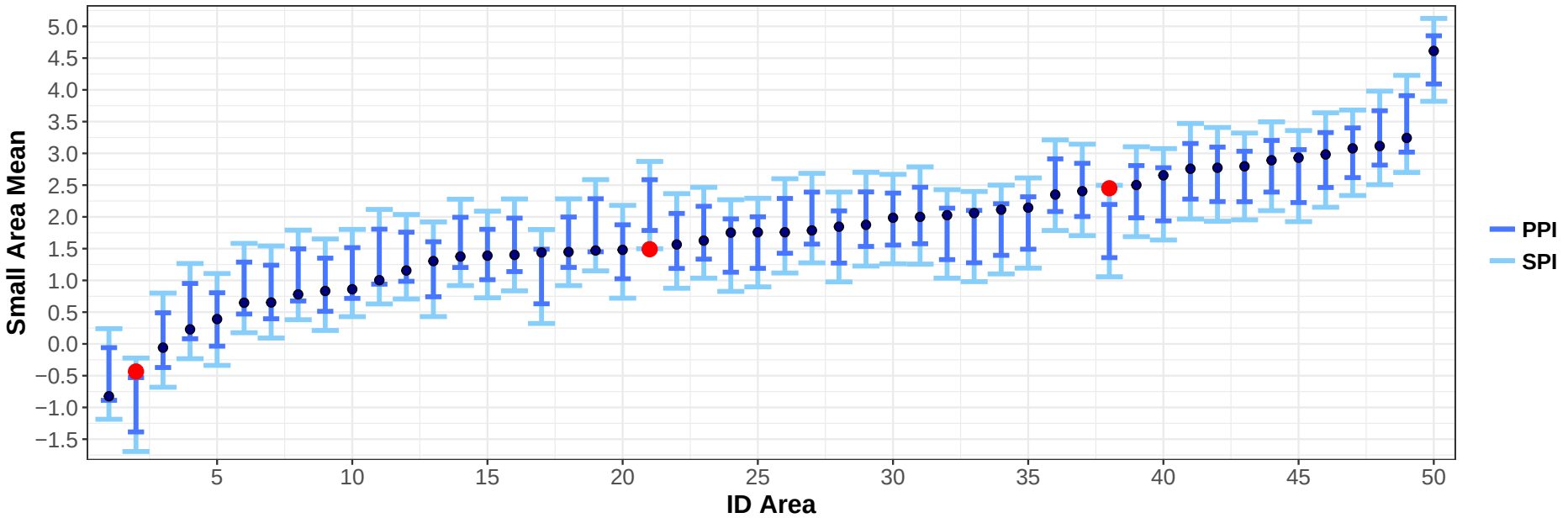
REB SPI

Empirical coverage probability using $I^B_{1-\alpha}$ with REB and different MSE estimators. The nominal coverage probability is 95%.

			mse^*_{B1}		mse^*_{BC2}		mse^*_{3T}		mse^*_{SP}		mse^*_{SPA}	
e_{dj}	u_d	D/n_d	25/5	75/15	25/5	75/15	25/5	75/15	25/5	75/15	25/5	75/15
$t_6(0.5)$	$N(1)$	REML	92.6	95.4	92.2	95.4	92.8	95.6	93.6	95.8	93.6	95.8
$t_6(0.5)$	$N(1)$	MM	92.4	95.4	91.2	95.2	92.6	95.4	93.4	95.6	93.4	95.6
$\chi^2_5(1)$	$N(0.5)$	REML	92.2	93.2	91.8	93.2	92.4	93.4	93.0	93.4	92.6	93.4
$\chi^2_5(1)$	$N(0.5)$	MM	93.0	93.2	92.4	93.0	92.6	93.2	92.8	93.6	93.6	93.6
$N(1)$	$N(0.5)$	REML	92.2	94.2	92.0	94.4	92.4	94.2	92.6	94.6	92.8	94.8
$N(1)$	$N(0.5)$	MM	92.4	95.0	92.6	95.6	92.8	94.6	92.2	95.0	92.4	94.8
$t_6(1)$	$t_6(0.5)$	REML	90.8	94.2	91.0	93.6	91.2	94.0	91.2	95.0	91.2	95.0
$t_6(1)$	$t_6(0.5)$	MM	90.8	95.2	90.8	95.2	90.8	95.2	91.0	95.0	90.6	95.0

Comparison of PPI and REB SPI

PPI and REB bootstrap SPI for small area means estimated using REML, mse^*_{SPA}, $e_{dj} \sim t(0.5)$, $u_d \sim N(1)$, $D = 50$.



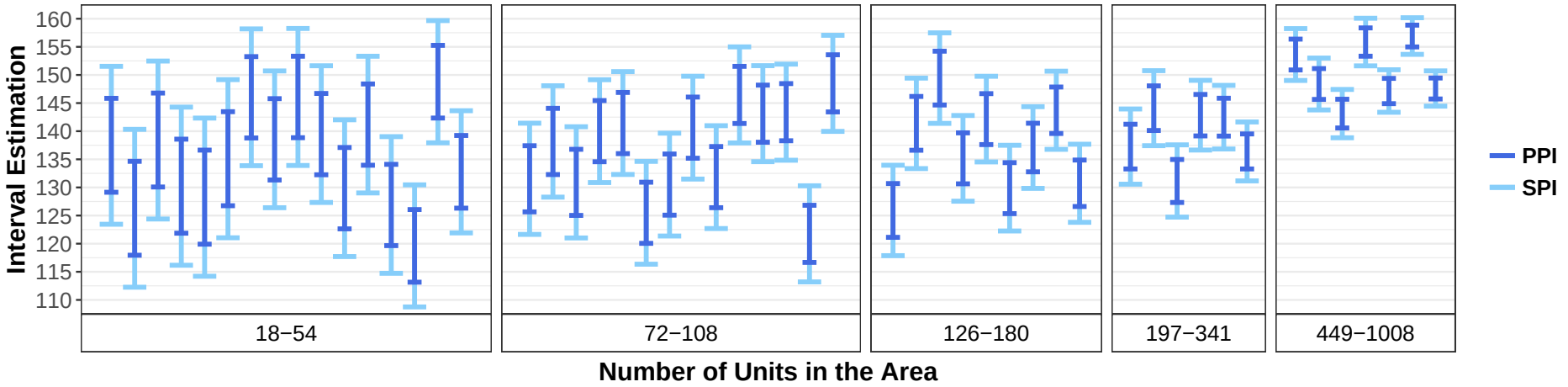
Outline

Modelling income in Galicia: general setting

Structural Survey for Homes in Galicia (NW of Spain)

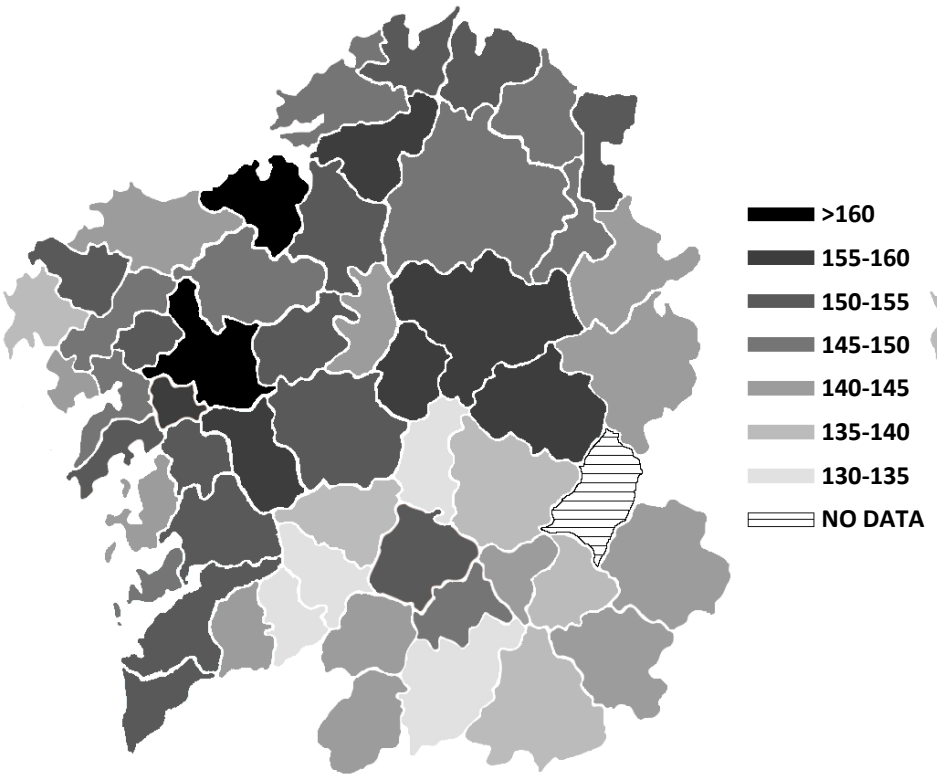
A study of the households and their socio-economical conditions in Galicia in 2005.

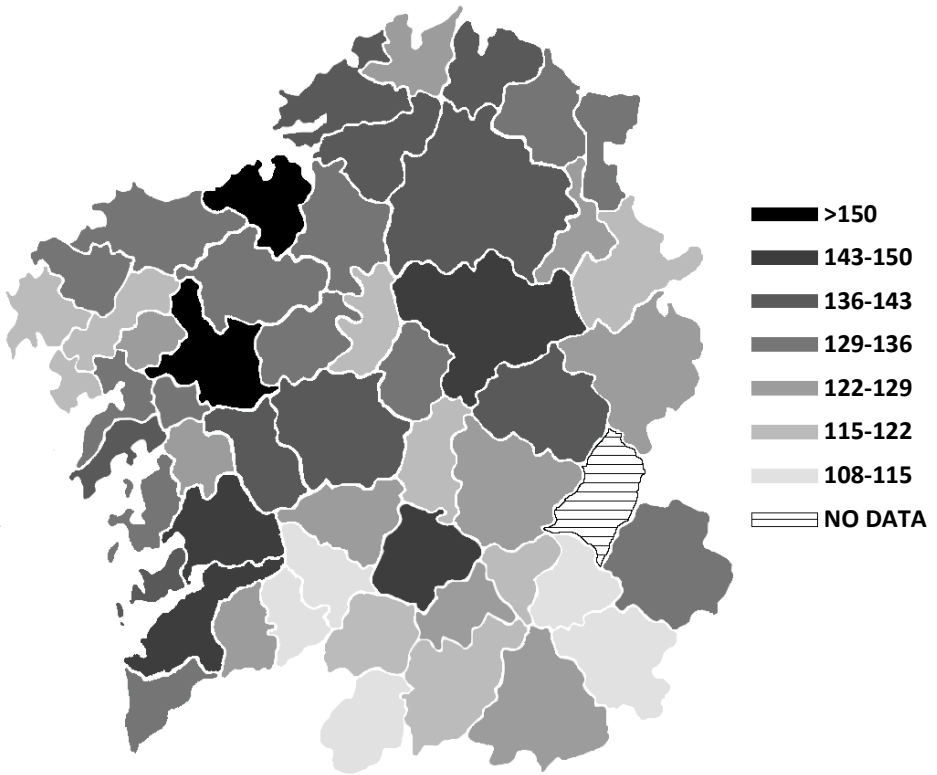
- $\boldsymbol{y}$ is a square root of total income of each household.
- Covariates: selected using xGAIC of Lombardía et al. (2017); $\boldsymbol{x}_1 - \boldsymbol{x}_{11}$ are related to households and $\boldsymbol{x}_{12} - \boldsymbol{x}_{18}$ are related to the person who earns the most in the household.
- Small areas: $D = 52$ comarcas in Galicia with $n_d = 18$ to $n_d = 1008$, $n = 9203$.



Modelling income in Galicia: SPI

SPI estimation of a square root of household income at the level of comarca in Galicia: (left) upper limit, (right) lower limit.





Thank you for the attention.

Outline

1. Ganesh, N. (2009). Simultaneous credible intervals for small area estimation problems. *Journal of Multivariate Analysis*, 100(8), 1610-1621.
2. Lombardía, M. J., López-Vizcaíno, E. & Rueda, C. (2017). Mixed generalized Akaike information criterion for small area models. *Journal of the Royal Statistical Society: Series A*, 180(4), 1229–1252.
3. Sun, J., & Loader, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics*, 22(3), 1328–1345.
4. Weyl, H. (1939). On the volume of tubes. *American Journal of Mathematics*, 61(2), 461-472.