# Assessing the Importance of Sampling Weights in Bayesian Small Area Models for Ordinal Data

Miguel Ángel Beltrán-Sánchez  Miguel Ángel Martínez-Beneito  Ana Corberán-Vallet

Department of Statistics and Operations Research, University of Valencia, Burjassot (Valencia), Spain

## Workshop “Explaining and Promoting Health Equity”

## Introduction

- Accounting for the sampling design is essential in survey analysis.
- Bayesian models for survey data should include design variables or survey weights.

## Objectives

- Describe two Bayesian models for spatial ordinal survey data:
  (i) Likelihood model including design variables [1].
  (ii) Pseudo-likelihood model incorporating sampling weights.
- Assess the impact of survey weights via simulation.

## Methodology

- $Y_i \in \{1, \dots, J\}$: ordinal response variable for individual $i$, $i = 1, \dots, N$.
- Stratified sampling design with $Z$ strata.
- Study region partitioned into $M$ small areas.
- $\boldsymbol{y} = (y_1, \dots, y_n)^T$: observed data from $n$ survey respondents.
- $\boldsymbol{x}_i = (z_i, m_i)^T$: covariate vector, where $z_i$ is the stratum and $m_i$ the area.

### Likelihood model

$$y_i | \boldsymbol{\pi}(\boldsymbol{x}_i) \sim \text{Categorical}(\boldsymbol{\pi}(\boldsymbol{x}_i)),$$
$$\text{logit}(\gamma_j(\boldsymbol{x}_i)) = \kappa_j + \alpha_{z_i} + \theta_{m_i}.$$

- $\boldsymbol{\pi}(\boldsymbol{x}_i) = (\pi_1(\boldsymbol{x}_i), \dots, \pi_J(\boldsymbol{x}_i))^T$, and $\pi_j(\boldsymbol{x}_i) = P(Y_i = j | \boldsymbol{x}_i)$.
- $\gamma_j(\boldsymbol{x}_i) = \sum_{r=1}^{j} \pi_r(\boldsymbol{x}_i)$, $j = 1, \dots, J-1$: cumulative probabilities.
- $\{\alpha_z\}_{z=1}^{Z}$ and $\{\theta_m\}_{m=1}^{M}$: stratum and location effects.
- As an extension, $\{\kappa_j\}_{j=1}^{J-1}$ could vary across strata.

### Pseudo-likelihood model

$$y_i | \boldsymbol{\pi}(\boldsymbol{x}_i) \sim \text{Categorical}(\boldsymbol{\pi}(\boldsymbol{x}_i))^{\tilde{w}_{z_i}},$$
$$\text{logit}(\gamma_j(\boldsymbol{x}_i)) = \kappa_j + \theta_{m_i}.$$

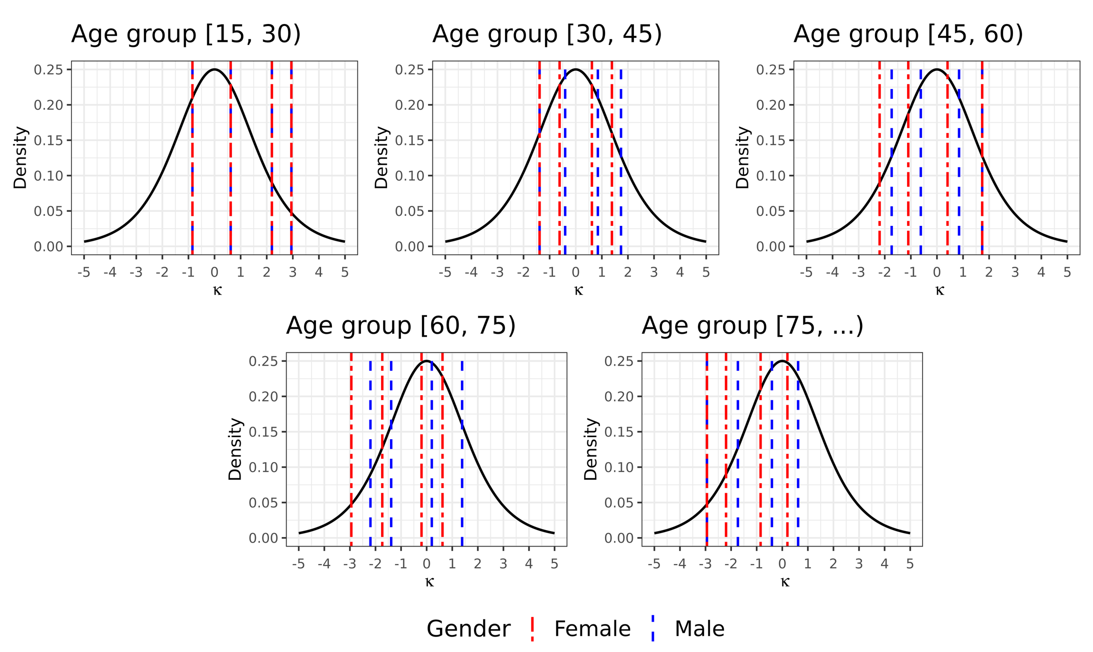
- $\tilde{w}_{z_i}$: normalized weight, $\tilde{w}_{z_i} = w_{z_i} / (\sum_{i=1}^{n} w_{z_i}/n)$.
- $w_{z_i}$: design weight (inverse of the inclusion probability).

### Prior distributions

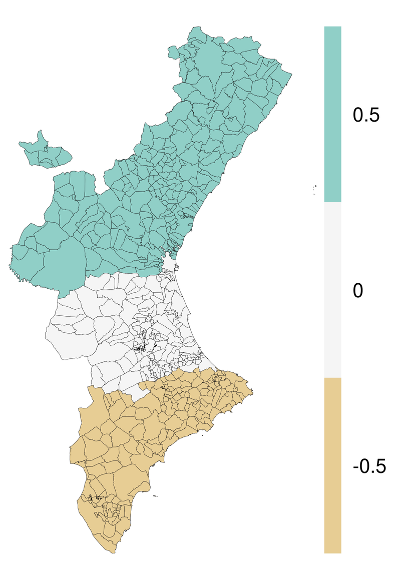
- Non-informative priors are specified for $\{\kappa_j\}_{j=1}^{J-1}$ and $\{\alpha_z\}_{z=1}^{Z}$.
- Conditional autoregressive (CAR) prior for $\{\theta_m\}_{m=1}^{M}$ [2].

## Simulation study

- Spatial risk surface over the Region of Valencia defined via horizontal risk bands.
- Small-area level: municipalities, $M = 542$.
- Individual-level responses generated according to gender, age and location.



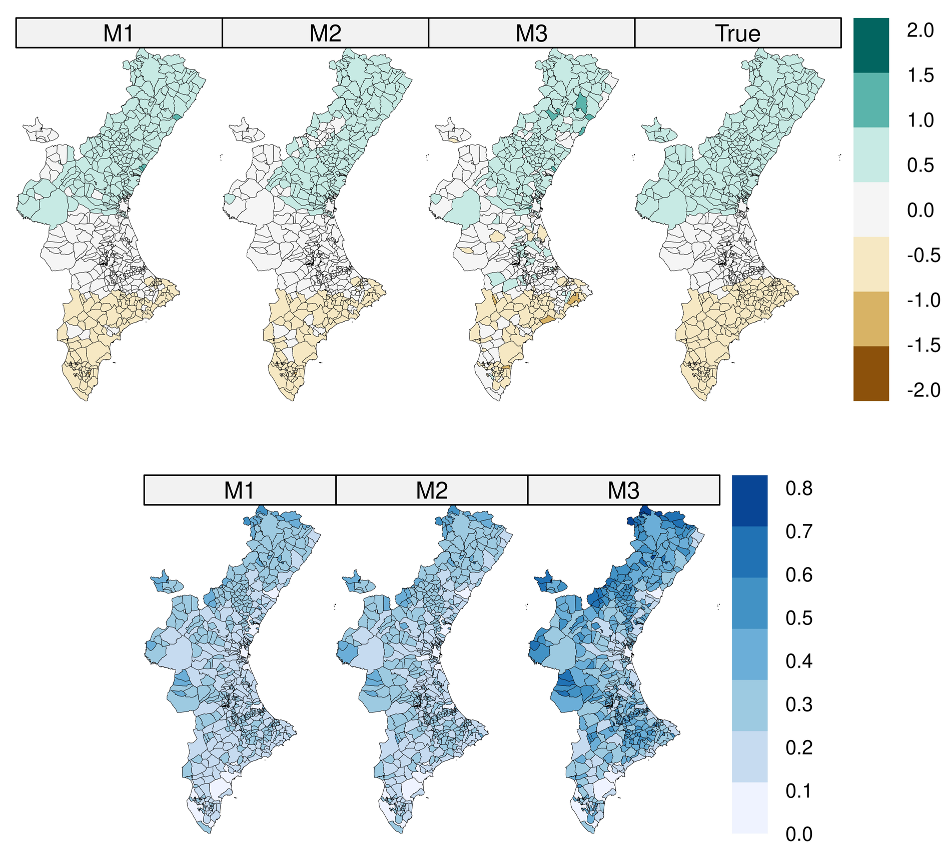

Intercepts by gender and age




Spatial risk surface

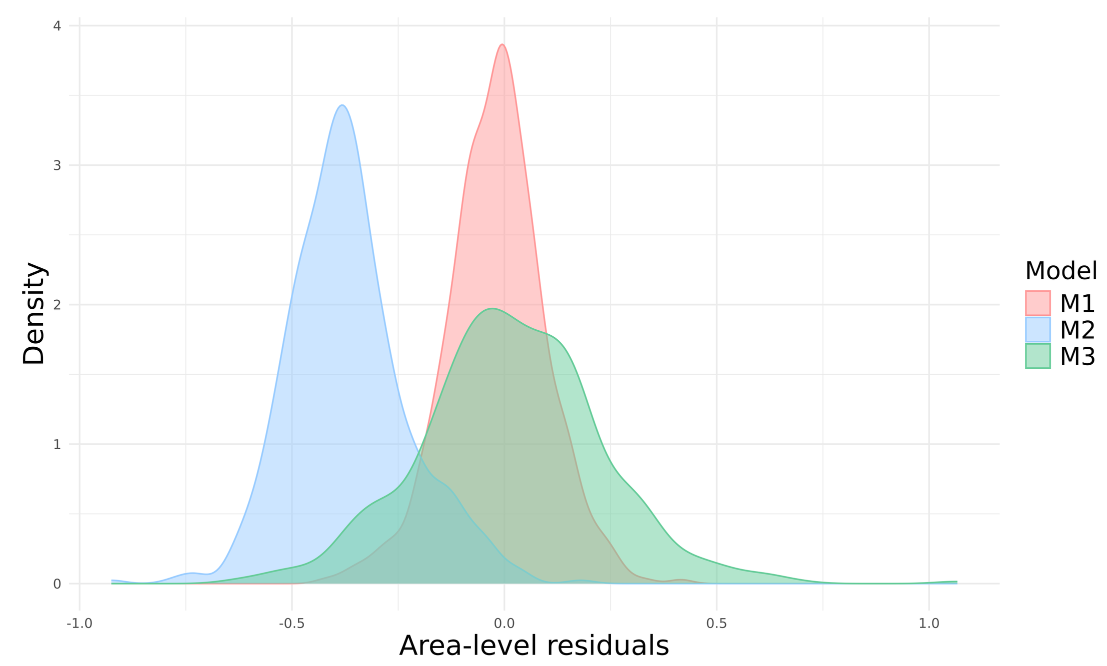
- Non-representative stratified sampling design by gender and age.
- **M1:** Likelihood model + all design variables + spatial random effect.
- **M2:** Likelihood model + only gender + spatial random effect.
- **M3:** Pseudo-likelihood model + normalized weights + spatial random effect.

## Results




Posterior mean (top) and SD (bottom) of the spatial random effect

- Posterior means similar across models.
- **M3** shows larger variability due to the inclusion of survey weights.




Density of area-level residuals: true population mean minus posterior predictive mean

- Densities for **M1** and **M3** are centered around zero, while **M2** exhibits a clear bias.
- Although **M2** captures a similar spatial pattern, its intercept estimates are biased due to incomplete adjustment for the sampling design.

## Conclusions

- Survey weights are unnecessary when all design variables are included.
- If the covariate vector does not fully account for the sampling design, ignoring survey weights may lead to biased population-level inference.

### Future lines of research

- Extension of the pseudo-likelihood model to include design variables.
- Application to the European Social Survey.

## References

[1] Miguel Angel Beltrán-Sánchez, Miguel-Angel Martinez-Beneito, and Ana Corberán-Vallet. Bayesian modeling of spatial ordinal data from health surveys. *Statistics in Medicine*, 43(21):4178–4193, 2024.

[2] Brian G Leroux, Xingye Lei, and Norman Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 2000.

## Funding

This paper is part of the project PID2022-136455NB-I00, funded by Ministerio de Ciencia, Innovación y Universidades of Spain (MCIN/AEI/10.13039/501100011033/ FEDER, UE) and the European Regional Development Fund.

## Correspondence

Miguel Ángel Beltrán-Sánchez, Department of Statistics and Operations Research, University of Valencia, Burjassot (Valencia), Spain. Email: angel.beltran@uv.es https://orcid.org/0000-0001-9450-2973

GitHub profile

VNIVERSITAT DE VALÈNCIA  VAlència BAyesian Research  NTNU Norwegian University of Science and Technology  MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES  UNIÓN EUROPEA FONDO EUROPEO DE DESARROLLO REGIONAL "Una manera de hacer Europa"  AGENCIA ESTATAL DE INVESTIGACIÓN