Multilevel Bivariate Areal Modelling for School Data

an application with R-INLA

Leonardo Cefalo
Università degli Studi di Bari Aldo Moro
Virgilio Gómez – Rubio,
Universidad de Castilla – La Mancha

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
GRINS FOUNDATION

Scope

- Study how the **Invalsi** scores at **the 2nd** year of **high schools** in **Italian and Mathematics** are driven by:
    1) The infrastructural state of **municipalities**
    2) Unobservable **spatial effects ➔ areal modelling**
- Observation period: school year **2022/23**

Covariates

Choice: forward selection

1. Share of high schools served by **urban public transport**
2. Share of high schools served by **ultra-broadband** connection
3. ISTAT **inner areas** municipality taxonomy:
   A – B: **Central** (infrastructural **poles**) ➔ model: `Central`
   C – D: **Intermediate** ➔ model: `1 - Central - Peripheral`
   E – F : **Peripheral** ➔ model: `Peripheral`

Spatial structure – random effects

- Data observed only in **874** municipalities over ab. 7900: **few links** between municipalities
- How to define the **adjacency structure**?
  - ➢ Spatial random effects at a **higher level**:
    a) **Provinces: 105** areas
    b) **Catchment areas** of **infrastructural poles** (ISTAT inner areas taxonomy): **206** areas

Model outline

- Generic score for j–th municipality of i–th province:

$$\begin{pmatrix} y_{1,i,j} \\ y_{2,i,j} \end{pmatrix} = x_{1,i,j}\begin{pmatrix} \beta_{11} \\ \beta_{12} \end{pmatrix} + \cdots + x_{p,i,j}\begin{pmatrix} \beta_{p1} \\ \beta_{p2} \end{pmatrix} + \begin{pmatrix} z_{1,i} \\ z_{2,i} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,i,j} \\ \varepsilon_{2,i,j} \end{pmatrix}$$

- y: Invalsi score; x: covariates; β: fixed effects; z: random effects; ε: error term
- Dependence is accounted for by the **random effect**
- Mathematics: Normal model; Italian: skew–Normal model

IMCAR model (Mardia, 1988)

$$\begin{cases} \boldsymbol{z}(s) \sim N(\boldsymbol{0}, [\boldsymbol{\Lambda} \otimes (\boldsymbol{D} - \boldsymbol{W})]^{-1}) \\ \boldsymbol{\Lambda}^{-1} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \end{cases}$$

- $\boldsymbol{W}$: binary **neighbourhood matrix** (1: neighbours; 0: not neighbours)
- $\boldsymbol{D}$: **diagonal** matrix of the **number of neighbours** of each area
- $\boldsymbol{\Lambda}$: global **precision** parameter:

  - **dense** $\boldsymbol{\Lambda}$ ➔ **correlated** z processes
  - **diagonal** $\boldsymbol{\Lambda}$ ➔ **independent** z processes

IMCAR model (Mardia, 1988)

- Multivariate **extension** of the popular **Besag** model (1974)
- **Singular precision** matrix ➔ **improper** model
  - ➢ **Sum-to-zero constraint** on all **connected components**
- Then, each connected component needs a **specific intercept**

Besag, J.: *Spatial Interaction and the Statistical Analysis of Lattice Systems*. J. R. Stat. Soc. Ser. B 36(2), 192–236 (1974)
Mardia, K.V.: *Multi-dimensional multivariate Gaussian Markov random fields with application to image processing*. JMA 24, 265-284 (1988)

PMCAR model (Gelfand, 2003)

$$\boldsymbol{z}(s) \sim N(\boldsymbol{0}, [\boldsymbol{\Lambda} \otimes (\boldsymbol{D} - \alpha \boldsymbol{W})]^{-1})$$

- **Proper** model: the **precision** matrix now has **full rank**
- Implies an **additional** hyperparameter $\alpha \in [0; 1]$
- The improper model can be seen as the limit case for $\alpha \rightarrow 1$

Gelfand A.E., Vounatsou P.: Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics 4(1), 11-25 (2003)

Restricted Regression (Reich et al. 2006)

- Spatial confounding: **linear dependence** between $\boldsymbol{x}$ and $\boldsymbol{z}$
  - Restriction on z
  - z is **projected** onto the **orthogonal** space to the column space of x
- Very **strong hypothesis** ➔ **shrinks** the posterior **variance** of **fixed effects** ➔ **questioned** in literature

Reich BJ, Hodges JS, Zadnik V.: Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics; 62(4):1197-1206 (2006)

Model implementation

- Multivariate CAR models can be easily implemented with R-INLA
- Specific **R** package: **INLAMSM**
- A number of models is supported, both as correlated and independent ones

Palmí-Perales, F., Gómez-Rubio, V., Martinez-Beneito, M. A.: Bayesian Multivariate Spatial Models for Lattice Data with INLA. J. Stat. Softw. 98(2), 1–29 (2021)

Model comparison

Model	z level	Restr.	Independent CPO	Independent DIC	Independent MSE	Dependent CPO	Dependent DIC	Dependent MSE
ICAR	Prov	Unr	6.720,84	13.439,36	239,03	6.688,50	13.376,42	235,09
ICAR	Prov	Restr	6.809,53	13.613,08	254,59	6.763,20	13.524,43	251,58
ICAR	Pole	Unr	6.729,47	13.456,59	237,90	6.693,44	13.386,56	232,08
ICAR	Pole	Restr	6.814,89	13.623,89	246,81	6.753,73	13.506,98	239,22
PCAR	Prov	Unr	6.721,18	13.439,41	238,51	6.688,52	13.376,46	233,77
PCAR	Prov	Restr	6.869,74	13.732,62	269,53	6.819,30	13.636,30	265,98
PCAR	Pole	Unr	6.730,24	13.458,00	237,35	6.694,18	13.388,09	230,39
PCAR	Pole	Restr	6.876,65	13.746,51	260,24	6.808,12	13.615,77	250,75
NULL	-	-	6.979,62	13.959.54	346,1			

- **Correlated** models outperform **independent** ones
- All **spatial** models outperform the **null** one
- **Unrestricted** models are preferable

Fixed effects summary

Subj	Covariate:	mean	q0.025	q0.975	sd	signif.
MAT	Intercept	-0,948	-8,073	6,208	3,546	
MAT	Peripheral	2,726	0,942	4,510	0,910	*
MAT	Central	-2,297	-4,270	-0,324	1,006	*
MAT	Broadband activ.	3,304	1,193	5,414	1,076	*
MAT	Urban public tpt	2,527	0,459	4,596	1,055	*
ITA	Intercept	-1,485	-7,450	4,483	2,965	
ITA	Peripheral	2,421	0,477	4,379	0,995	*
ITA	Central	-1,896	-3,949	0,162	1,048	
ITA	Broadband activ.	2,344	0,136	4,560	1,128	*
ITA	Urban public tpt	2,905	0,809	5,007	1,070	*

- Model: **unrestricted PCAR** with **province-level** correlated random effects
- **Flat prior** used for all random effects
- All covariates **range** $[\mathbf{0}, \mathbf{1}]$
- The ISTAT **inner areas taxonomy** seems to be the strongest driver

Hyperparameters summary

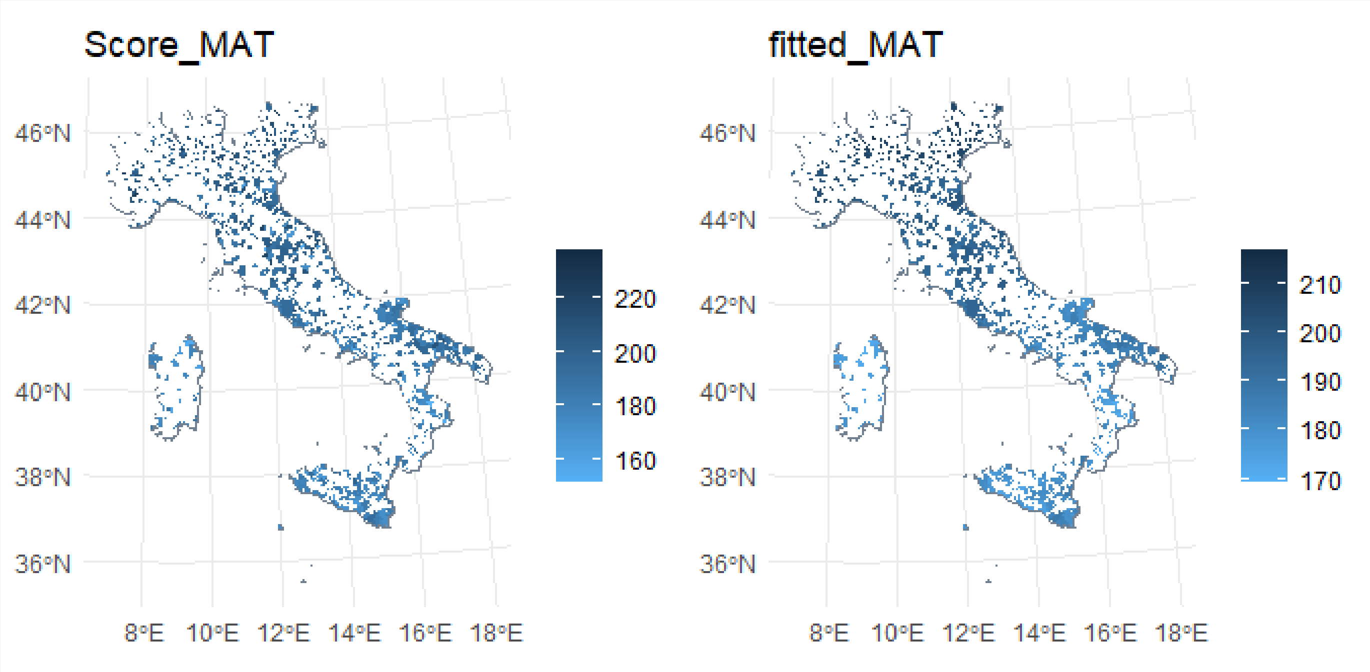
Subj	Hyperparameter	Median	q0.025	q0.975	sd
-	alpha	0,987792	0,954111	0,996503	0,011304
MAT	Random eff. variance	48,17235	29,7795	78,66275	12,51115
ITA	Random eff. variance	31,36284	18,62418	53,20078	8,857486
-	Random eff. correlation	0,967596	0,8789	0,99278	0,030281
MAT	Error variance	110,4913	100,4776	121,681	5,399837
ITA	Error scale param.	131,3615	118,9709	145,085	6,650431
ITA	Error skewness	-0,36348	-0,48862	-0,22397	0,067343

- Model: **unrestricted PCAR** with correlated **province-level** random effects
- **Wishart prior** on **random effects precision**
- **Flat prior** on the square roots of error scale parameters
- High **within – provinces variability** unexplained by covariates

Preliminary findingss

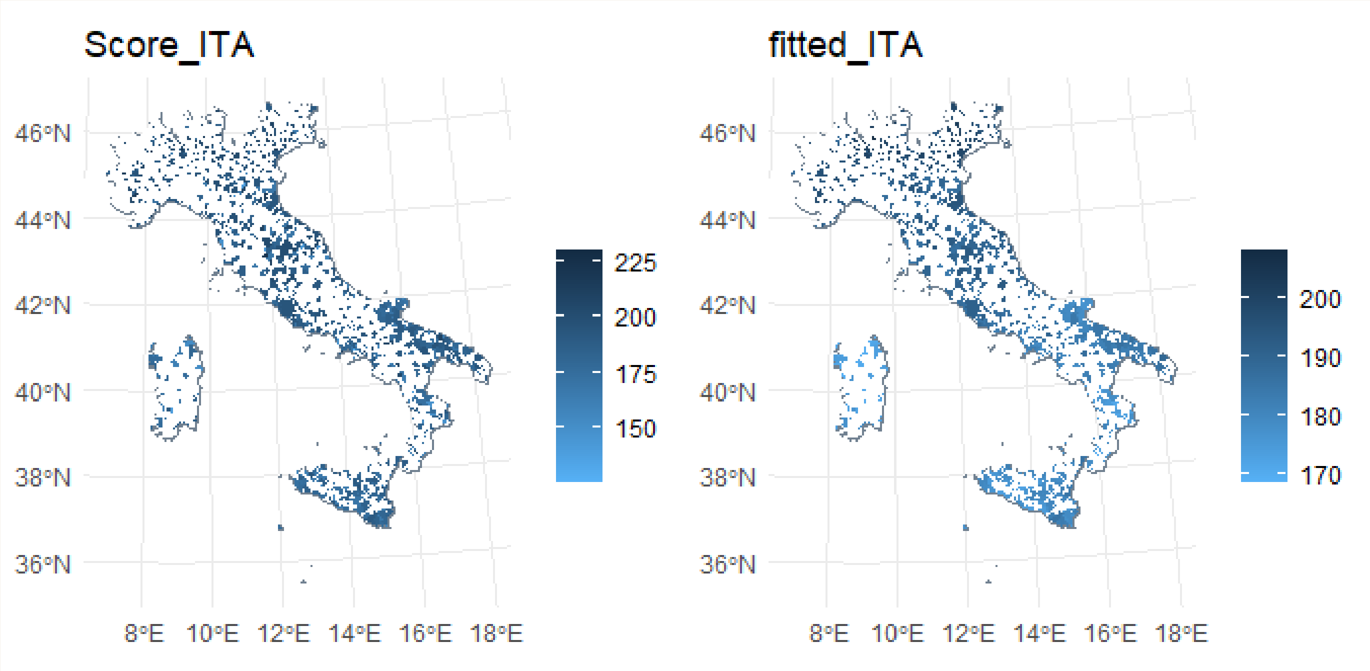
- Assuming that x and z are **independent** leads to **poor** model accuracy ➔ spatial confounding should **not** be removed via **restricted regression**
- **Both** the **infrastructural** datum (covariates) and the **territorial structure** are **necessary** to explain disparities in Invalsi scores
- **Skewness** in **Italian** scores cannot be explained by existing information and cannot be ignored

Fitted values – Mathematics



Model: **unrestricted** PCAR with correlated **province-level** random effects

Fitted values – Italian



Model: **unrestricted** **PCAR** with correlated **province-level** random effects

Possible future developments

- Implementing **scaled IMCAR** model
- Implementing more **recent de-confounding** methodologies
- Studying the **accuracy** of the **Laplace approximation** for **Skew-Normal** data
- Extending the analysis to **other school grades**

Thank you for your attention

Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani