Thesis draft

Leonardo Cefalo

April 24, 2025

Contents

Chapter 1

The INLA and Gaussian Markov Random Fields

1.1 Introduction

In this chapter, we present an overview of a computational methodology extensively used throughout this thesis, the Integrated Nested Laplace Approximation.

In the context of Bayesian inference, integrating out the posterior distribution of the parameters of interest only happens in a small number of fortunate cases, i.e. when among all possible prior distributions, a set of conjugate ones is assumed.

The most popular way to overcome non-conjugacy in the posteriors is the approximation of posteriors by simulation using Markov Chain Monte Carlo (MCMC hereinafter) methods. MCMC methods are a true institution in Bayesian statistics due to their rigorous background and the capability of balancing computational speed and accuracy.

Nevertheless, an alternative approach has emerged in the last seventeen years, the Integrated Nested Laplace Approximation [74, 84, INLA hereinafter]. The core idea behind INLA is to directly apply a deterministic approximation to the posterior, and dates back to the proposal by [80] to approximate the first and second moments of the parameters. The Laplace approximation leverages on the convergence of the full posterior to a Gaussian distribution when parameters are Gaussian *a priori*, the approximation error being in $\mathcal{O}(n^{-1})$ where n is the length of the target variable.

The most convenient area of application of INLA are the models for Gaussian Markov Random Fields [73], due to the computational properties we describe in Section 1.2. That being said, the remainder of this chapter will briefly summarise how the INLA works in Section 1.3 and the `R` software in which it is implemented in Section 1.4

1.2 Why using INLA? Gaussian Markov Random Fields

1.2.1 Conditional Independence in Gaussian Random Fields

An interesting property of the Normal distribution is that if two observations are mutually independent conditioned on all the other ones, then the marginal precision element corresponding to that couple is zero even though the marginal covariance is non-zero [73].

To see this, consider a generic n-dimensional Gaussian random field $\mathbf{X} = (X_1, X_2 ... X_n) \sim N(\mu, \mathbf{\Sigma})$. Then, a deductive prove is given for this property:

$$X_i \perp X_j \mid \mathbf{X_{-ij}} \iff q_{ij} = 0 \tag{1.1}$$

For any $i \neq j, \quad i, j \in [1, n]$; $q_{i,j}$ is the element in position (i, j) in the precision matrix $\mathbf{Q} := \mathbf{\Sigma^{-1}}$, assuming that $\mathbf{\Sigma}$ is nonsingular. Any set of indices $A \subset \mathbb{N}^n$ can be chosen in order to partition $\mathbf{X}$ into $\mathbf{X_A}$ and $\mathbf{X_B}$, being $B = \mathbb{N}^n / A$ with $card(A) := m < n$, $card(B) = n - m$, namely:

$$\mathbf{X} := \begin{pmatrix} \mathbf{X_A} \\ \mathbf{X_B} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_\mathbf{A} \\ \mu_\mathbf{B} \end{pmatrix} \text{ and } \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{AA}} & \mathbf{\Sigma_{AB}} \\ \mathbf{\Sigma_{BA}} & \mathbf{\Sigma_{BB}} \end{pmatrix}$$

Matrix $\mathbf{\Sigma_{AA}}$ has dimensions $m \times m$, $\mathbf{\Sigma_{BB}}$ is $(n - m) \times (n - m)$, $\mathbf{\Sigma_{AB}}$ is $n \times (n - m)$ and $\mathbf{\Sigma_{BA}} = \mathbf{\Sigma'_{AB}}$. Since $\mathbf{X}$ is Gaussian by definition, we know its density is

$$\pi(\mathbf{X}) = \frac{1}{(2\pi)^{n/2} \mid \mathbf{\Sigma} \mid^{-1}} \, e^{-\frac{d^2}{2}}, \quad \text{with} \quad d^2 := (\mathbf{X} - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{X} - \mu)$$

Being d^2 the squared Mahalanobis distance. From the law of compound probability we also know that

$$\pi(\mathbf{X}) = \pi(\mathbf{X_A} | \mathbf{X_B}) \, \pi(\mathbf{X_B}) \quad \text{or equivalently} \quad \ln \pi(\mathbf{X}) = \ln \pi(\mathbf{X_A} | \mathbf{X_B}) + \ln \pi(\mathbf{X_B})$$

$\pi(\mathbf{X}) \propto exp(-\frac{d^2}{2})$ implies that $d = d^2_{A|B} + d^2_B$, being $d^2_{A|B}$ and d^2_B the squared Mahalanobis distances of respectively the conditional distribution of $\mathbf{X_A}$ given $\mathbf{X_B}$ and the marginal distribution of $\mathbf{X_B}$.

To compute the precision matrix $\mathbf{Q} := \mathbf{\Sigma^{-1}}$, we keep treating $\mathbf{\Sigma}$ as a block matrix and decompose it into three matrices of the same size whose inversion is somewhat easier. Specifically we want to factorise $\mathbf{\Sigma}$ as $\mathbf{T_U} \times \mathbf{C} \times \mathbf{T_L}$, where $\mathbf{T_L}$ and $\mathbf{T_U}$ are respectively a lower triangular and an upper triangular matrix; $\mathbf{C}$ instead must have zeroes as the upper right and lower left blocks. Let us start by defining:

$$\mathbf{T_L} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Sigma_{BB}^{-1} \Sigma_{BA}} & \mathbf{I} \end{pmatrix}$$

which allows to factorise $\mathbf{\Sigma}$ as:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{AA}} & \mathbf{\Sigma_{AB}} \\ \mathbf{\Sigma_{BA}} & \mathbf{\Sigma_{BB}} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma_{AA}} - \mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}} & \mathbf{\Sigma_{AB}} \\ \mathbf{0} & \mathbf{\Sigma_{BB}} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}} & \mathbf{I} \end{pmatrix}$$

For the sake of simplicity, let us define the Schur complement of $\mathbf{\Sigma_{BB}}$ in $\mathbf{\Sigma}$ [43, paragraph 0.7.3] as $\mathbf{S_{BB}} := \mathbf{\Sigma_{AA}} - \mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}$; non-singularity is assumed for the Schur complement. Now, the following factorisation holds:

$$\begin{pmatrix} \mathbf{S_{BB}} & \mathbf{\Sigma_{AB}} \\ \mathbf{0} & \mathbf{\Sigma_{BB}} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{S_{BB}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{BB}} \end{pmatrix}$$

with

$$\mathbf{T_U} = \begin{pmatrix} \mathbf{I} & \mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} \mathbf{S_{BB}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{BB}} \end{pmatrix}$$

so that:

$$\mathbf{\Sigma} = \mathbf{T_U}\,\mathbf{C}\,\mathbf{T_L} = \begin{pmatrix} \mathbf{I} & \mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{S_{BB}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{BB}} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}} & \mathbf{I} \end{pmatrix}$$

It follows straightforwardly that:

$$\mathbf{T_U^{-1}} = \begin{pmatrix} \mathbf{I} & -\mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad \mathbf{T_L^{-1}} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}} & \mathbf{I} \end{pmatrix}$$

and

$$\mathbf{C}^{-1} = \begin{pmatrix} \mathbf{S_{BB}^{-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{BB}^{-1}} \end{pmatrix}$$

Thus we derive this general result, which holds for any distribution, provided that marginal covariance matrices are nonsingular:

$$\mathbf{Q} = \mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{S_{BB}^{-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{BB}^{-1}} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{1.2}$$

If we perform the multiplication we obtain

$$\mathbf{Q} = \begin{pmatrix} \mathbf{S_{BB}^{-1}} & \mathbf{S_{BB}^{-1}}\mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}} \\ \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{AB}}\mathbf{S_{BB}^{-1}} & \mathbf{\Sigma_{BB}^{-1}} + \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\mathbf{S_{BB}^{-1}}\mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}} \end{pmatrix}$$

As for $\mathbf{S_{BB}}$, we may define: $\mathbf{S_{AA}} := \mathbf{\Sigma_{BB}} - \mathbf{\Sigma_{BA}}\mathbf{\Sigma_{AA}^{-1}}\mathbf{\Sigma_{AB}}$ and notice that

$$\left(\mathbf{\Sigma_{BB}^{-1}} + \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\mathbf{S_{BB}^{-1}}\mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}}\right)\,\left(\mathbf{\Sigma_{BB}} - \mathbf{\Sigma_{BA}}\mathbf{\Sigma_{AA}^{-1}}\mathbf{\Sigma_{AB}}\right) =$$

$$= \mathbf{I} - \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\mathbf{\Sigma_{AA}^{-1}}\mathbf{\Sigma_{AB}} + \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\mathbf{S_{BB}^{-1}}\mathbf{\Sigma_{AB}} - \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\mathbf{S_{BB}^{-1}}\mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\mathbf{\Sigma_{AA}^{-1}}\mathbf{\Sigma_{AB}} =$$

$$= \mathbf{I} - \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\left(\mathbf{S_{BB}^{-1}} - \mathbf{S_{BB}^{-1}}\mathbf{\Sigma_{AB}}\mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\mathbf{\Sigma_{AA}^{-1}} - \mathbf{\Sigma_{AA}^{-1}}\right)\mathbf{\Sigma_{AB}} =$$

$$\mathbf{I} - \mathbf{\Sigma_{BB}^{-1}}\mathbf{\Sigma_{BA}}\left[\mathbf{S_{BB}^{-1}} - \mathbf{S_{BB}^{-1}}\left(\mathbf{\Sigma_{AA}} - \mathbf{S_{BB}}\right)\mathbf{\Sigma_{AA}^{-1}} - \mathbf{\Sigma_{AA}^{-1}}\right]\mathbf{\Sigma_{AB}} = \mathbf{I}$$

This enables us to write the precision matrix in compact form:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{S_{BB}^{-1}} & \mathbf{S_{BB}^{-1}\Sigma_{AB}\Sigma_{BB}^{-1}} \\ \mathbf{\Sigma_{BB}^{-1}\Sigma_{AB}S_{BB}^{-1}} & \mathbf{S_{AA}^{-1}} \end{pmatrix} \tag{1.3}$$

What matters is that the precision elements corresponding to the variables included in the subset A, namely $\mathbf{X_A}$ are all in the matrix $\mathbf{S_{BB}^{-1}}$.

Recalling equation 1.2, it is possible to show a fundamental property of the Normal distribution [55, result 4.6]. If $\mathbf{X}$ has been partitioned into sets A and B, the squared Mahalanobis distance can be written as:

$$\begin{aligned} d^2 = \begin{pmatrix} \mathbf{X_A} - \mu_A \\ \mathbf{X_B} - \mu_B \end{pmatrix}' \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{-\Sigma_{BB}^{-1}\Sigma_{BA}} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{S_{BB}^{-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{BB}^{-1}} \end{pmatrix} \times \\ \times \begin{pmatrix} \mathbf{I} & \mathbf{-\Sigma_{AB}\Sigma_{BB}^{-1}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{X_A} - \mu_A \\ \mathbf{X_B} - \mu_B \end{pmatrix} = \\ = \begin{pmatrix} \mathbf{X_A} - \mu_A + (\mathbf{X_B} - \mu_B)\mathbf{\Sigma_{BB}^{-1}\Sigma_{BA}} \\ \mathbf{X_B} - \mu_B \end{pmatrix}' \begin{pmatrix} \mathbf{S_{BB}^{-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{BB}^{-1}} \end{pmatrix} \times \\ \times \begin{pmatrix} \mathbf{X_A} - \mu_A + (\mathbf{X_B} - \mu_B)\mathbf{\Sigma_{BB}^{-1}\Sigma_{BA}} \\ \mathbf{X_B} - \mu_B \end{pmatrix} \end{aligned}$$

By setting $\mu_{A|B} := \mu_A + (\mathbf{X_B} - \mu_B)\mathbf{\Sigma_{BB}^{-1}\Sigma_{BA}}$ d^2 is decomposed into:

$$d^2 = \left(\mathbf{X_A} - \mu_{A|B}\right)' \mathbf{S_{BB}^{-1}} \left(\mathbf{X_A} - \mu_{A|B}\right) + \left(\mathbf{X_B} - \mu_B\right)' \mathbf{\Sigma_{BB}^{-1}} \left(\mathbf{X_B} - \mu_B\right)$$

The second term is d_B^2, namely the Mahalanobis distance for the partition $\mathbf{X_B}$. The first term is the Mahalanobis distance of a Normal distribution with mean $\mu_{A|B}$ and covariance $\mathbf{S_{BB}}$, but it must also be the Mahalanobis distance of $\mathbf{X_A}$ conditioned on $\mathbf{X_B}$, by the law of conditional probability. Therefore it is proved that, if $\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$, for any two partitons A and B such that $\mathbf{X_B}$ is the complement to $\mathbf{X_A}$ with respect to $\mathbf{X}$, then

$$\mathbf{X_A}|\mathbf{X_B} \sim N\left(\mu_A + (\mathbf{X_B} - \mu_B)\mathbf{\Sigma_{BB}^{-1}\Sigma_{BA}},\ \mathbf{\Sigma_{AA}} - \mathbf{\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}}\right) \tag{1.4}$$

Now we can prove equation 1.1. The partition $\mathbf{X_A}$ conditional to $\mathbf{X_B}$, has therefore precision equal to $\mathbf{S_{BB}^{-1}}$, according to equation (1.4). Yet also the element of the joint precision matrix corresponding to the subset A is equal to $\mathbf{S_{BB}^{-1}}$. This means that by choosing the index set A as any couple (i, j) with $i \neq j$, we have

$$\mathrm{COV}[X_i, X_j|\mathbf{X_{-ij}}] \equiv q_{ij} \tag{1.5}$$

And if these two variables are conditionally independent, then their precision element is zero, as stated in equation 1.1.

Equation 1.1 implies that modelling Gaussian processes with sparse precision matrix implies a gain in computation.

1.2.2 The Markov Property

Section 1.2.1 shows that conditional pairwise independence in Gaussian random fields implies the corresponding marginal precision element is zero.

It is thus convenient to identify a class of probabilistic models for Gaussian random fields for which conditional pairwise independence holds by definition. A sufficient condition to this aim is to satisfy the Markov property. Following [73, Section 2.2], a Markovian random field can be intuitively defined with respect to a graph; it is easy to notice that the graph structure is a common representation for data like time series or lattice (either regular or irregular, in any number of dimensions).

We start by assuming that each element of an n-dimensional Gaussian random field can thus be associated with a node in a graph denoted as $\mathcal{G} := (\mathcal{S}, \mathcal{E})$, where $\mathcal{S} = \{s_i\}_{i=1}^n$ is the set of nodes of size n and $\mathcal{E} := \{(s_i, s_j) : s_j \in \partial s_i\}_{i,j=1}^n$ is the set of edges; hereinafter ∂s_i denotes the set of neighbours of node s_i, or equivalently the set of nodes *connected with* node s_i. By convention, it is assumed that $s_i \notin \partial s_i$.

The relationship $X : \mathcal{S} \to \mathbb{R}^n$ is thus assumed to hold, and for brevity we will write X_i in place of $X(s_i)$. The *global* Markov property [39] means that:

$$\pi(X_i \mid X_{-i}) = \pi(X_i \mid \partial s_i) \tag{1.6}$$

For a generic probability density function π. This property is relevant to our aims because it implies that

$$X_i \perp X_j \mid X_{-i} \quad \forall j \notin \partial s_i$$

i.e. the probability distribution of X_i is uniquely specified by its neighbours, and X_i is independent on any non neighbouring site.

Putting together the Markov property (Equation 1.6) with the precision structure of Gaussian random fields when conditional pairwise independence holds (Equation 1.1) it becomes clear that random processes satisfying both the conditions, namely the *Gaussian Markov Random Fields* are characterised by a precision matrix, say $\mathbf{Q}$ which only has nonzero entries corresponding to neighbouring pairs, namely $\text{card}(\mathcal{E})$. For autoregressive time series and processes defined on regular lattices $\text{card}(\mathcal{E})$ is a multiple of n; for irregular lattices there is no such an exact relationship, still the number of edges is typically in $\Theta(n)$ in connected graphs.

1.3 Introduction to the INLA

1.3.1 Latent Gaussian model outline

The baseline of how the INLA applies to our aims is a generic hierarchical regression model:

$$\begin{cases} E[y \mid \eta, \Psi] = g^{-1}(\eta) \\ \eta = A\vartheta \end{cases}$$

where y denotes the response variable consisting of N observations; η is the linear predictor linked with a function g to the expected value of the response variable; ϑ is a generic vector of latent variables of interest; A is a known design matrix; Ψ is the array of hyperparameters, whose size is usually in $\mathcal{O}(10^1)$.

The linear predictor, to our aims, can be defined as:

$$\mathbf{A} := (X \quad \xi) \quad \text{and} \quad \vartheta := \begin{pmatrix} \beta \\ \mathbf{z} \end{pmatrix}$$

where β is the array of regression coefficients of length p, z is an additional array of latent effects of length n, X is a matrix of explanatory variables of size $N \times p$, and ξ is the design matrix of latent effects of size $N \times n$.

Labelling β as "fixed" and z as "random" effects is quite frequent in hierarchical regression applications, but due to the polysematic nature of these terms we will tend to avoid them. In a strictly probability perspective, there is no conceptual distinction between them: they are both unknown random variables entering the linear predictor through a known design matrix (either X or ξ), are typically assumed to be Gaussian *a priori* and their posterior distribution is the primary aim of statistical inference. We keep them separated due to how they are interpreted: β represents the association between a set of known variables (X) and y, while z is an unobserved process shaping the structure of y itself.

The first necessary assumption is that $(y_j \perp\!\!\!\perp y_k)|\vartheta, \Psi$ and $(y_j \perp\!\!\!\perp \eta_k)|\vartheta, \Psi$ for any $j, k \in [1, N]$ with $j \neq k$, hence each observation y_j depends on ϑ only through one value η_j. This also implies that the likelihood can be factorised as follows:

$$f(y|\vartheta, \Psi) = f(y|\eta, \Psi) = \prod_{j=1}^{N} f(y_j|\eta_j, \Psi)$$

Moreover, we assume that ϑ follows a Normal distribution conditioned on Ψ; in our case, we have

$$\vartheta|\Psi \sim N(\mathbf{0}, (\mathbf{Q}_\Psi)^{-1})$$

where $\mathbf{Q}_\Psi$ denotes the precision matrix. Assuming that ϑ satisfies the Markovian properties is not necessary but it ensures the computational gains implied by the sparsity of $\mathbf{Q}_\Psi$.

Posterior marginals, namely $\pi(\vartheta_i|y)$ and $\pi(\Psi_j|y)$ $\forall i \in [1, n+p]$ and $j \in [1, \text{card}(\Psi)]$, are obtained by solving the integrals:

$$\begin{aligned} \pi(\vartheta_i|y) &= \int_\Psi \pi(\vartheta_i|y, \Psi)\pi(\Psi|y)d\Psi \\ \pi(\Psi_j|y) &= \int_{\Psi_{-j}} \pi(\Psi|y)\, d\Psi_{-j} \end{aligned} \tag{1.7}$$

which is feasible provided that Ψ has a small size. However, how we will see in the next paragraph, this operation can hardly be completed in closed form.

1.3.2 Approximating Hyperparameters Posterior

Given Equation 1.7, the first task is computing

$$\pi(\Psi|y) = \frac{f(y|\eta,\Psi)\pi(\vartheta|\Psi)\pi(\Psi)}{\pi(\vartheta|y,\Psi)f(y)}$$

The numerator is known *a priori*; $f(y)$ does not depend on the parameters of interest and can be treated here as a normalising constant. The function which is hardly available in closed form, instead, is the full conditional $f(\vartheta|y,\Psi)$.

Thus, the idea behind the INLA it to replace it with its Gaussian approximation:

$$\pi(\vartheta|y,\psi) \approx \pi_G(\vartheta|y,\psi) = \frac{\pi(\vartheta,y|\Psi)}{\int \pi_G(\vartheta,y|\Psi)\,d\vartheta} \tag{1.8}$$

Where the subscript G denotes the Gaussian approximation. For brevity, define $g(\vartheta) := ln\pi(\vartheta,y|\Psi)$. Then, a Taylor approximation truncated at the second order is applied:

$$g(\vartheta) \approx g(\vartheta_0) + \nabla g(\vartheta_0)'(\vartheta-\vartheta_0) + \frac{1}{2}(\vartheta-\vartheta_0)'H_g(\vartheta_0)(\vartheta-\vartheta_0)$$

Where $H_f(x_0)$ denotes the Hessian matrix of a generic scalar-valued function $f(\cdot)$ evaluated at x_0
[1]

Then, the point ϑ_0 is set as the the mode of $g(\vartheta)$, denoted as $\vartheta_0(y,\Psi)$ to highlight its dependence on observed data and on hyperparameters. By doing so, the first-order term in the above formula (the gradient) is zero and $\pi_G(\vartheta|y,\Psi)$ equals indeed a Gaussian density:

$$\pi_G(\vartheta,y|\Psi) = \pi(\vartheta_0(y,\Psi),y|\Psi)\,e^{\frac{1}{2}[\vartheta-\vartheta_0(y,\Psi)]'Hg(\vartheta_0(y,\Psi))[\vartheta-\vartheta_0(y,\Psi)]}$$

Whose integral in ϑ is simply a Gaussian integral, providing the normalising constant in the denominator of equation 1.8. We thus have

$$\pi_G(\vartheta \mid y,\Psi) \propto e^{-\frac{1}{2}(\vartheta-\mu_0)Q_0(\vartheta-\mu_0)} \tag{1.9}$$

[1] Please notice the Hessian matrix can be further simplified. To see this, first consider that

$$H_g(\vartheta) = H_{\ln\pi(\vartheta|\Psi)}(\vartheta) + \sum_{j=1}^{N} H_{\ln f(y|\vartheta,\Psi)}(\vartheta)$$

. The first addendum is $-Q_\Psi\forall\vartheta$. Concerning the second term, notice that instead of ϑ inference can be made on the first N elements of the vector $\theta := ((\eta+\epsilon)^\top,\beta^\top,\vartheta^\top)^\top$, which is itself a GMRF [75]. The term ϵ is a Gaussian error with arbitrarily small variance employed to make the distribution of the augmented predictor $\eta+\epsilon$. We then have, $\forall i,j\in[1,N]$, that $\partial^2 \ln f(y_i \mid \theta_i,\Psi)/\partial\theta_r\partial\theta_c$ is different from zero only for $r=c=i$, hence under this parametrisation $H_f(\theta)$ is a diagonal matrix.

where $\mu_0 = \vartheta_0(y, \Psi)$ and $Q_0 = -Hg(\vartheta_0(y, \Psi))$. Posterior marginals for hyperparameters can thus be computed applying the Laplace approximation

$$\pi(\Psi|y) \approx \pi_{LA}(\Psi|y) \propto \left. \frac{f(y|\eta, \Psi)\pi(\vartheta|\Psi)\pi(\Psi)}{\pi_G(\vartheta|y, \Psi)} \right|_{\vartheta=\vartheta_0(\Psi)}$$

Given $\pi_{LA}(\Psi|y)$, hyperparameter marginal posterirors in 1.7 are integrated numerically by moving along a multidimensional grid starting from the posterior mode of Ψ.

1.3.3 Approximating Parameters Posterior

To approximate $\pi(\vartheta_i|y, \Psi)$, a rough solution would be using the Gaussian approximation in equation 1.8. Even though this is a computationally cheap operation, it may suffer low accuracy. For this reason [74] proposed two alternatives.

The first one consists in reiterating the Laplace approximation for each element of ϑ, i.e. marginalising ϑ_i out from $\pi(\vartheta|y, \Psi)$ by using a Gaussian approximation to $\pi(\vartheta_{-i}|\vartheta_i, y, \Psi)$ and setting ϑ_{-i} equal to the mode of $\pi(\vartheta_{-i}|\vartheta_i, \Psi)$. This is the most rigorous choice, but is computationally demanding.

The other approach is known as "simplified Laplace" approximation, representing a compromise between the former two approaches in terms of accuracy and computational cost; it basically consists in truncating the Taylor approximation of $g(\vartheta)$ at the third order term, while still locating the approximation at $\vartheta_0(y, \Psi)$. This allows to fit a Skew-Normal density to $g(\vartheta)$.

These three approaches complete the original INLA framework. More details on how the parameters vector is defined, see [75].

A recent paper introduced an additional alternative strategy based on a Variational Bayes correction to the posterior mean of ϑ [83]. This latter approach consists in correcting μ_0 (as in eq. 1.9) by an additive vector, say λ, whose entries are nonzero only for a subset I drawn from the total set of indices of ϑ, namely $I \subset \{1, 2, ...n\}$:

$$\pi_{VB}(\vartheta|y, \Psi) \propto e^{-\frac{1}{2}(\vartheta - \mu_0 - \Sigma_I\lambda)'Q_0(\vartheta - \mu_0 - \Sigma_I\lambda)}$$

where Σ_I is a projection matrix determined from a subset of the columns of Q_0^{-1}. Now, λ is determined as the minimum to the following objective function:

$$\begin{aligned} \lambda := \arg\min_{\lambda} \Bigg\{ & KLD\left(f_{VB}(\vartheta|y, \Psi) \;||\; f(\vartheta|\Psi)\right) + \\ & - \int_{\vartheta} \sum_{i=1}^{N} f(y_i|\eta_i, \Psi) f_{VB}(\vartheta \mid y, \Psi)\, d\vartheta \Bigg\} \end{aligned} \tag{1.10}$$

where $KLD(f(x) \;||\; g(x)) := \int f(x) ln \frac{f(x)}{g(x)} dx$ denotes the Kullback-Leibler divergence between a proposed model $f(x)$ and a baseline model $g(x)$. Given the relatively small cardinality of λ, this methodology is referred to as low-rank correction.

1.4 The R-INLA package

The INLA is implemented into a comprehensive and self-sufficient R environment, the INLA R package [38, 86]. It will be referred to as R-INLA hereinafter.

A notable feature of this software is operational flexibility. Firstly, it is worth noticing the user-friendliness of regression models syntax, being it analogous to the glm() environment in R. Aditionally, a high number of likelihood functions has been implemented so far, covering not only most of the exponential family but also distributions such as the Skew-Normal [5] or the Skew-t. While many prior distributions for latent effects are ready-made as well (the list can be consulted with function inla.list.models()), the system also allows users to define custom models through the function inla.rgeneric.define().

Due to the size of the complex .dll libraries on which this software relies for model computation, it is not available on CRAN, but only in the dedicated repository https://inla.r-inla-download.org/R/. Unless otherwise stated, throughout this thesis the 2024.10.13 testing version of the software is employed.

The four approaches described in Section 1.3.3 to approximate $\pi(\vartheta_i \mid y)$ are available within the software, which can run either in "classic" (old) or "compact" (new) mode. The former supports the Gaussian, Simplified Laplace and Full Laplace approximations, and needs to be activated with the command: inla(... , inla.mode = "classic", ...)). The latter mode, supporting the VB mean correction to the Gaussian approximation, is implemented by default (or can be equivalently set with the command inla(..., inla.mode = "compact", ...)). Unless differently stated, we rely on the VB approximation.

Chapter 2

Elements of areal data modelling

2.1 Introduction

In this chapter, we provide an overview on the statistical methodology followed throughout the last two chapters. In particular, this chapter is intended as a minimal toolbox in areal data modelling, with specific application in the next chapters.

2.2 The intrinsic CAR model

Consider a p-variate random variable $z_i = (z_{i,1}, \ldots, z_{i,p})'$ taking values over the i-th node of a graph $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ with $\mathcal{S} \subset \mathbb{N}^n$ which is described by:

- The neighbourhood matrix W, with $w_{ij} = 1 \iff s_i \in \partial s_j$ and w_{ij} otherwise; notice that by convention, $w_{ii} = 0$ (cfr Chapter 1.2)
- The diagonal degree matrix D, whose elements correspond to the number of edges of each node, hence $d_i = \sum_{j=1}^{n} w_{ij} = \text{card}(\partial s_i)$
- The Laplacian matrix $R = D - W$, whose rank deficiency equals the number of connected components in $\mathcal{G}$

Following [61], see also chapter 10 of [7], we suppose that z_i follows this prior conditional distribution:

$$z_i | z_{-i}, \Lambda \sim N\left(\sum_{j=1}^{n} \frac{w_{ij}}{d_i} z_j, \frac{1}{d_i}\Lambda^{-1}\right)$$

Where Λ is a matrix-valued precision parameter. Please notice this formulation corresponds to a special case of the CAR model described by [61], who proposed

the more general form $z_i|z_{-i}, \Lambda \sim N\left(\sum_{j=1}^{n} B_{i,j} z_i, \frac{1}{d_i}\Lambda^{-1}\right)$ and presented $B_{ij} = \frac{w_{ij}}{d_i} I_p$ as a simplifying assumption (see Corollary 2 to Theorem 2.1).

To compute the joint distribution $\pi(y|\sigma^2)$ we rely on the Brook's Lemma [39, 11]:

$$\pi(z) = \prod_{i=1}^{n} \frac{\pi\left(z_i \mid \{x_j\}_{j<i} \cap \{z_k\}_{k>i}\right)}{\pi\left(x_i \mid \{x_j\}_{j<i} \cap \{z_k\}_{k>i}\right)} \pi(x) \tag{2.1}$$

Where x is a set of known variables satisfying the positivity condition. Also, notice $\pi(x)$ is a constant term we can treat a a normalising constant. Then, consider:

$$\ln\left(z_i|\{x_j\}_{j<i} \cap \{z_k\}_{k>i}, \Lambda\right) = \tag{2.2}$$

$$= C_0 - \frac{d_i}{2}\left(z_i - \sum_{j=1}^{i-1} \frac{w_{ij}}{d_i} x_j - \sum_{k=i+1}^{n} \frac{w_i k}{d_i} z_k\right)^{\top} \Lambda \left(z_i - \sum_{j=1}^{i-1} \frac{w_{ij}}{d_i} x_j - \sum_{k=i+1}^{n} \frac{w_i k}{d_i} z_k\right) \tag{2.3}$$

Where C_0 is an additive constant. Now, with no loss of generality, let us assume $x_i = 0 \,\forall\, i \in [1, n]$, so that numerator elements in 2.1

$$\ln \pi\left(z_i|\{x_j = 0\}_{j<i} \cap \{z_k\}_{k>i}, \Lambda\right) = C_0 - \frac{d_i}{2}\left(z_i - \sum_{k=i+1}^{n} \frac{w_{ij}}{d_i} z_k\right)^{\top} \Lambda \left(z_i - \sum_{k=i+1}^{n} \frac{w_{ij}}{d_i} z_k\right) = \tag{2.4}$$

$$= C_0 - \frac{d_i}{2}\left[z_i^{\top}\Lambda z_i - 2\left(\sum_{k=i+1}^{n} \frac{w_{ik}}{d_i} z_k\right)^{\top} \Lambda z_i\right] - \frac{d_i}{2}\left(\sum_{k=i+1}^{n} \frac{w_{ik}}{d_i} z_k\right)^{\top} \Lambda \left(\sum_{k=i+1}^{n} \frac{w_{ik}}{d_i} z_k\right) \tag{2.5}$$

And denominator elements take the form:

$$\ln \pi\left(x_i|\{x_j = 0\}_{j<i} \cap \{z_k\}_{k>i}, \Lambda\right)|_{x_i=0} = C_0 - \frac{d_i}{2}\left(\sum_{k=i+1}^{n} \frac{w_{ik}}{d_i} z_k\right)^{\top} \Lambda \left(\sum_{k=i+1}^{n} \frac{w_{ik}}{d_i} z_k\right) \tag{2.6}$$

Then, with a little bit of standard algebra we get:

$$\ln \pi(z \mid \Lambda) = C_1 + \sum_{i=1}^{n} d_i z_i^{\top} \Lambda z_i - \sum_{i=1}^{n}\sum_{k=1}^{n} w_{ik} z_i^{\top} \Lambda z_k \tag{2.7}$$

Where C_1 is an additive constant independent on z. Now, let us introduce the row-wise vectorising operator vec_r, which stacks rowwise the elements of a $n \times p$

matrix into a $np \times 1$ vector. First consider:

$$\sum_{i=1}^{n} d_i z_i^\top \Lambda z_i = \begin{pmatrix} z_{11} \\ \vdots \\ z_{1p} \\ z_{21} \\ \vdots \\ z_{2p} \\ \vdots \\ z_{n1} \\ \vdots \\ z_{np} \end{pmatrix}^\top \begin{pmatrix} d_1\Lambda & 0 & \dots & 0 \\ 0 & d_2\Lambda & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n\Lambda \end{pmatrix} \begin{pmatrix} z_{11} \\ \vdots \\ z_{1p} \\ z_{21} \\ \vdots \\ z_{2p} \\ \vdots \\ z_{n1} \\ \vdots \\ z_{np} \end{pmatrix} = \text{vec}_r(z)^\top (W \otimes \Lambda) \text{vec}_r(z)$$

Then, consider:

$$\sum_{i=1}^{n}\sum_{k=1}^{n} w_{ik} z_i^\top \Lambda z_k = \begin{pmatrix} z_{11} \\ \vdots \\ z_{1p} \\ z_{21} \\ \vdots \\ z_{2p} \\ \vdots \\ z_{n1} \\ \vdots \\ z_{np} \end{pmatrix}^\top \begin{pmatrix} 0 & w_{12}\Lambda & \dots & w_{1n}\Lambda \\ w_{21}\Lambda & 0 & \dots & w_{2n}\Lambda \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1}\Lambda & w_{n2}\Lambda & \dots & 0 \end{pmatrix} \begin{pmatrix} z_{11} \\ \vdots \\ z_{1p} \\ z_{21} \\ \vdots \\ z_{2p} \\ \vdots \\ z_{n1} \\ \vdots \\ z_{np} \end{pmatrix} = \text{vec}_r(z)^\top (D \otimes \Lambda) \text{vec}_r(z)$$

This proves the joint distribution of z is [61, 7, 32, this latter case regarding the PCAR]:

$$\pi(z|\Lambda) \propto -\frac{1}{2} \text{vec}_r(z)^\top R \otimes \Lambda \text{vec}_r(z) \tag{2.8}$$

For practical reasons, if we want to fit such a model in `R-INLA`, it is convenient to rearrange z columnwise through the vec_c operator [69], such that $\text{vec}(z) = (z_{11}, \dots, z_{n1}, z_{12}, \dots, z_{n2}, \dots, z_{1p}, \dots, z_{np})^\top$. To do so, we are actually using an orthogonal permutation matrix $\mathcal{P}$, such that

$$\text{vec}_c(z) = \mathcal{P} \times \text{vec}_r(z)$$

Here we give a definition of $\mathcal{P}$. To do so, consider the i-th basis vector of $\mathbb{R}^p$, e_i, the i-th row of I_n, whose i-th element is 1 and all other elements are zero, $\forall i \in [1, n]$. We then have:

$$\mathcal{P} := \begin{pmatrix} I_n \otimes e_1^\top \\ I_n \otimes e_2^\top \\ \vdots \\ I_n \otimes e_p^\top \end{pmatrix}$$

Considering thatVAR$[\text{vec}_c(z)] = \mathcal{P} \times \text{VAR}[\text{vec}_r(z)] \times \mathcal{P}^\top$, hence $Prec[\text{vec}_c(z)] = \mathcal{P} \times Prec[\text{vec}_r(z)] \times \mathcal{P}$, it holds that:

$$\mathcal{P}\left(R \otimes \Lambda\right)\mathcal{P}^\top = \Lambda \otimes R$$

Which can be proved with some algebra. Hence, in total equivalence with equation 2.8, we know that the joint distribution of a multivariate ICAR process is

$$z|\Lambda \sim N\left(0, (\Lambda \otimes R)^+\right) \tag{2.9}$$

The $^+$ superscript here denotes the Moore-Penrose pseudoinverse. This variance definition is implied by the rank deficiency of R, which equals the number of connected components in the underlying graph $\mathcal{G}$. $\pi(z|\Lambda)$ is therefore an improper distribution [see e.g. 12, 41].

2.2.1 Precision scaling

As it can be seen in the previous equations, the ICAR precision is the product of two terms, a global parameter and a structure matrix. This hinders parameter interpretation at the global level; to start discussing it, we first consider the case of a connected graph. Denoting with $\mathbf{S} := \mathbf{R}^+$, the marginal variance of a generic i-th realisation of the latent effect for the h-th variable is:

$$\text{VAR}[z_{ih} \mid T] = \sigma_h^2 s_{ii}$$

Where σ_h^2 is the marginal variance of the h-th variable, corresponding to the h-th diagonal element of of Λ^{-1}; in other words, s_{ii} would be the marginal variance of z_i if the relevant global precision was equal to 1.

Hence, by definition, ICAR variance incorporates a factor determined by the neighbourhood structure. Taking this consideration to the global level, it implies that process variability is not described by the scale parameter (or by its reciprocal, the precision) alone, since a deterministic scale factor concurs as well. The intuitive solution to this issue is to scale the ICAR model, as shown by [79]. Following their operational proposal, we take the geometric mean of the diagonal of $\mathbf{S}$ as scaling factor - or, in other words, the geometric mean of marginal variances if all global precision parameters (diagonal entries of T) were equal to 1, namely

$$\bar{\sigma}^2 = \prod_{i=1}^{n} s_{ii}^{1/n}$$

this scaling factor is sometimes referred to as reference variance (or generalised variance, or typical marginal variance). We are thus able to separate the parameter Λ from the effect on precision induced by the graph structure. To do so, the scaled model does not employ $\mathbf{R}$ as the structure matrix, but the scaled matrix $\mathbf{R}_{scaled} = \bar{\sigma}^2\mathbf{R}$. On the other hand, the precision parameter becomes $\dfrac{1}{\bar{\sigma}^2}T$, and this is what actually expresses the precision of the process independently of how the graph is structured.

When the graph has $G > 1$ connected components, its Laplacian matrix is (a permutation of) the direct sum of the component Laplacians, as mentioned in the previous paragraph. The pseudo-inverses of these matrices may clearly have different typical marginal variances. In this case, it is necessary to scale the component-specific precisions separately, as shown in [28]:

$$\mathbf{R}_{scaled} = \mathcal{P}\left(\bar{\sigma}_1^2 \mathbf{R_1} \oplus \bar{\sigma}_2^2 \mathbf{R_2} \oplus ... \oplus \bar{\sigma}_G^2 \mathbf{R_G}\right) \mathcal{P}^\top$$

Where $\mathcal{G}$ is an appropriate permutation matrix, $\mathbf{R_i}$ is the Laplacian of the i-th component of the graph, and $\bar{\sigma}_i^2$ is the relevant typical variance.

In `R-INLA`, precision scaling is implemented automatically for intrinsic models, like the univariate ICAR, through the option `scale.model` within the `inla()` function call; otherwise the scaled structure matrix can be computed as a standalone object with `inla.scale.model()`.

In the multivariate case, `INLAMSM` provides readily-defined models for which the user is required to provide the neighbourhood matrix $\mathbf{W}$ instead of the Laplacian matrix (as different models with the same neighbourhood matrix have different structure matrices). Hence, to scale a multivariate ICAR model we derive $\mathbf{W}$ from the scaled Laplacian (function `ScaleQ()` generalises `inla.scale.model()` when the graph is disconnected):

2.3 Spatial confounding

When a spatially structured latent random variable is included in a regression model, it may happen to be correlated with some covariates. This issue can be traced to the more general problem of spatial confounding. Several approaches have been developed in almost two decades of literature ([82], [25]), starting from the intuitive solution of constraining random effects to be linearly independent on covariates ([72], [42]), which goes under the name of restricted spatial regression (RSR); this is done by projecting random effects to the orthogonal subspace the covariates matrix, i.e. setting $X(X^\top X)^{-1}X^\top \xi z = 0$.

To illustrate the issue, we recall the hierarchical regression model in Section 1.3.1 and suppose y is Gaussian. The linear predictor includes a Gaussian random effect of size n with prior mean 0 and non-negative defined structure matrix Q.

$$\begin{aligned} y &= X\beta + \xi z + \varepsilon \\ z &\sim N(0, \tau^{-1}Q^{-1}) \\ \beta &\sim N(0, M) \\ \varepsilon &\sim N(0, \sigma_\varepsilon^2 I) = N(0, \tau_\varepsilon^{-1} I_n) \end{aligned} \tag{2.10}$$

The covariate matrix X has size $n \times p$; ξ is the model matrix for the random effects (e.g. in many applications it is the $n \times n$ identity matrix).

This model can be parametrised in a more general way [see e.g. 78]:

$$\begin{aligned} y &\sim N(A_1\vartheta_1, \tau_\varepsilon^{-1} I_n) \\ \vartheta_1 &\sim N(A_2\vartheta_2, \tau_\varepsilon^{-1} C) \end{aligned} \tag{2.11}$$

Where $A_1 = (X\,\xi)$, and $\vartheta_1 = (\beta^\top z^\top)^\top$ is the vector of parameters. The term $A_2\vartheta_2$ has by construction both mean and variance zero.
Here we define

$$C^{-1} = \frac{\tau}{\tau_\varepsilon}\begin{pmatrix} \frac{1}{\tau M} I_p & 0_{p\times n} \\ 0_{n\times p} & Q \end{pmatrix}$$

Since the error term ε and the random effects z have different precision parameters, the matrix C^{-1} cannot be considered known Note: This notation is actually an artifact to proceed with further explanation. To see this, consider the prior precision of the random effects defined as τQ; $\tau_\varepsilon C^{-1}$ must thus be structured as $\begin{pmatrix} \frac{1}{M} I_p & 0_{p\times n} \\ 0_{n\times p} & \tau Q \end{pmatrix}$ where M is the known variance parameter of β, but τ is unknown and isolating τ to express $\tau_\varepsilon C^{-1} = \tau \begin{pmatrix} \frac{1}{\tau M} I_p & 0_{p\times n} \\ 0_{n\times p} & Q \end{pmatrix}$ would leave it dependent on an unknown parameter.

Now, it is possible to show [60] that the posterior mean of ϑ_1, conditioned on the model hyperparameters, is

$$E[\vartheta_1|y,\tau,\tau_\varepsilon] = \left(A_1^\top A_1 + C^{-1}\right)^{-1} A_1^\top y \tag{2.12}$$

For brevity, we refer to the matrix $A_1\left(A_1^\top A_1 + C^{-1}\right)^{-1} A_1$ as the hat matrix or H. Based on how we defined A_1 and C, notice that

$$\left(A_1^\top A_1 + C^{-1}\right)^{-1} = \begin{pmatrix} X^\top X + r_\varepsilon I_p & X^\top \xi \\ \xi^\top X & \xi^\top\xi + \bar{r}Q \end{pmatrix}^{-1} =$$

$$= \begin{pmatrix} G^{-1} + G^{-1}X^\top\xi(\xi^\top\xi + \bar{r}Q - \xi^\top XG^{-1}X^\top\xi)^{-1}\xi^\top XG^{-1} & -G^{-1}X^\top\xi(\xi^\top\xi + \bar{r}Q - \xi^\top XG^{-1}X^\top\xi)^{-1} \\ -(\xi^\top\xi + \bar{r}Q - \xi^\top XG^{-1}X^\top\xi)^{-1}\xi^\top XG^{-1} & (\xi^\top\xi + \bar{r}Q - \xi^\top XG^{-1}X^\top\xi)^{-1} \end{pmatrix}$$

With $\bar{r} := \dfrac{\tau}{\tau_\varepsilon}$ and $G := (X^\top X + r_\varepsilon I_p)$, where $r_\varepsilon := \dfrac{1}{\tau_\varepsilon M} = \dfrac{\sigma_\varepsilon^2}{M}$. It can thus be shown that:

$$E[\beta|y,\tau_\varepsilon,\bar{r}] = G^{-1}\{X^\top + X^\top\xi\left[\xi^\top\xi + \bar{r}Q - \xi^\top XG^{-1}X^\top\xi\right]^{-1}\xi^\top XG^{-1}X^\top + \\ -X^\top\xi\left[\xi^\top\xi + \bar{r}Q - \xi^\top XG^{-1}X^\top\xi\right]^{-1}\xi^\top\}\,y \tag{2.13}$$

And

$$E[z|y,\tau_\varepsilon,\bar{r}] = \left[\xi^\top\xi + \bar{r}Q - \xi^\top XG^{-1}X^\top\xi\right]^{-1}(\xi^\top - \xi^\top XG^{-1}X^\top)\,y \tag{2.14}$$

Equivalently, if for brevity we define $\Gamma = \xi^\top\xi + \bar{r}Q - \xi^\top XG^{-1}X^\top\xi$ this means that

$$H_{\tau_\varepsilon,\bar{r}} = XG^{-1}X^\top + \xi\Gamma^{-1}\xi^\top - XG^{-1}X^\top\xi\Gamma^{-1}\xi^\top - \xi\Gamma^{-1}\xi^\top XG^{-1}X^\top + XG^{-1}X^\top\xi\Gamma^{-1}\xi^\top XG^{-1}X^\top \tag{2.15}$$

If a flat prior is defined on β, we have $M \longrightarrow \infty$, with $G \longrightarrow X^\top X$. In this case, it is possible to show that the limits for equations (4) and (5) are the

results stated by [72] at page 1199; in that case, $\xi = I_n$. Moreover, the hat matrix does not depend on τ_ε but only on $\bar{r}$.

We now consider the projection matrix of y onto its least squares fitted values, and its orthongonal complement, respectively:

$$P := X(X^\top X)^{-1}X^\top \qquad \text{and} \qquad P_\perp := I - X(X^\top X)X^\top = I - P$$

The crucial step of all the forthcoming analysis is the eigendecomposition of $P_\perp$. Firstly, consider that its first $n-p$ eigenvalues are ones, and the latter p are zeroes.

To see that the eigenvalues of $P_\perp$ are either zeros or ones, first consider that $P_\perp = P_\perp P_\perp$. Then for a generic eigenvalue-eigenvector pair λ, v it holds that $P_\perp v = \lambda v$. With no loss of generality it is also true that $P_\perp P_\perp v = \lambda P_\perp v$, which equals to the identity $P_\perp v = \lambda^2$, hence $\lambda = \lambda^2$. The number of nonzero eigenvalues is the rank of $P_\perp$, which is in turn $n - rk(P)$, hence $n-p$.

We label the eigenvector matrix of $P_\perp$ as (LK); L is the matrix whose columns are the first $n-p$ nonzero eigenvectors, while K belongs to the null space of $P_\perp$. Thus the decomposition of $P_\perp$ is:

$$P_\perp = \begin{pmatrix} L & K \end{pmatrix} \begin{pmatrix} I_{n-p} & 0_{(n-p)\times p} \\ 0_{p\times(n-p)} & 0_{p\times p} \end{pmatrix} \begin{pmatrix} L^\top \\ K^\top \end{pmatrix} = LL^\top$$

We further notice that:

$$\begin{pmatrix} L^\top \\ K^\top \end{pmatrix} (L\,K) = (L\,K) \begin{pmatrix} L^\top \\ K^\top \end{pmatrix} = I_n\,, \quad KK^\top = P, \quad K^\top L = 0$$

Without loss of generality, it holds that

$$p(z|\tau) \propto exp\{-\frac{\tau}{2} z^\top Q z\} = exp\{-\frac{\tau}{2} z^\top (L\,K) \begin{pmatrix} L^\top \\ K^\top \end{pmatrix} Q (L\,K) \begin{pmatrix} L^\top \\ K^\top \end{pmatrix} z\}$$

We now define $\zeta = \begin{pmatrix} L^\top \\ K^\top \end{pmatrix} z$ such that $z = (L\,K)\zeta$ and decompose the newly defined random effect ζ as

$$\zeta = \begin{pmatrix} \zeta_1 \\ \zeta_0 \end{pmatrix} \quad \text{where} \quad \zeta_1 := L^\top z \quad \text{and} \quad \zeta_0 := K^\top z$$

The distribution of the new random effect is then:

$$\zeta = \begin{pmatrix} \zeta_1 \\ \zeta_0 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tau^{-1} \begin{pmatrix} L^\top QL & L^\top QK \\ K^\top QL & K^\top QK \end{pmatrix}^{-1} \right)$$

Spatial confounding occurs when the random effect z displays multicollinearity with the covariates X; when $p = 1$ this can be checked by comparing X and the last eigenvector of Q outside its null space.

Restricted regression means to project z onto the subspace orthogonal to X, which implies to fit a different model:

$$\text{RSR}: \quad y = X\beta + P_\perp z + \varepsilon = X\beta + L\zeta + \varepsilon \tag{2.16}$$

The rationale behind deconfounding z is constraining it to be totally independent of X, which is the same as imposing that z shall lie on the null space of P, hence the constraint is

$$\zeta_0 = K^\top z = 0$$

. Under the unrestricted model, it holds that $prec[\zeta_1|\tau] = (L^\top(Q)^{-1}L)^{-1}$; here, instead, we have $prec[\zeta_1|\tau, K^\top z = 0] = L^\top QL$, as in [eq. 16-17 at p. 1201 72]. To see this, recall:

$$lnp(\zeta|\tau; K^\top z = 0) \propto -\frac{\tau}{2}(\zeta_1^\top \, 0_p^\top) \begin{pmatrix} L^\top \\ K^\top \end{pmatrix} Q(L\,K) \begin{pmatrix} \zeta_1 \\ 0_p \end{pmatrix} = -\frac{\tau}{2}\zeta_1^\top \, L^\top QL\,\zeta_1$$

With respect to equation 2.16, if we consider $\xi = L$ and $\vartheta_1^\top = (\beta^\top, (L^\top z)^\top)$, it is straightforward to see that $\mathbb{E}[\beta|y, \tau_\varepsilon, \bar{r}] = G^{-1}X^\top y$; if $p(\beta)$ is flat, its posterior expectation reduces to $X^\top(X^\top X)^{-1}y$, which is the ordinary least squares estimator.

As a side effect of deconfounding z, it may be worthy noticing that the the size of the new precision matrix is $(n-p) \times (n-p)$ and its rank is $\leq \min\{rk(L), rk(Q)\} = \min\{(n-p), rk(Q)\}$.

Hence, if the rank deficiency of Q is less than p, the new random effect has a proper distribution even if z is an ICAR process. If the underlying graph is disconnected and $p(z)$ can be factorised in a factor for each connected component (CC) then the rank deficiency of Q is equal to the number of CCs, but it is necessary to impose a sum to zero constraint for each CC, hence the same number of intercepts would be required; thus, also in this case $n-p$ is smaller than $rk(Q)$ and $L^\top QL$ is still proper. However, there are still some possible situations in which this inconvenience can arise, e.g. in multilevel models; see further.

Referring to the flat β case, consider that under the restricted model:

$$\mathrm{E}[\beta_{RSR}|y, z, \tau_\varepsilon, \bar{r}] = \beta_{OLS} - (X^\top X)^{-1}X^\top(I - X(X^\top X)^{-1}X^\top)z = \beta_{OLS}$$

Where β_{RSR} denotes the vector of covariate effects under restricted spatial regression and β_{OLS} are covariate effects estimated via the ordinary least squares. This implies that the fixed effects are estimated independently of any random effect, and their expectation equals that of the nonspatial model.

Here we keep assuming a flat prior on β; referring to equation (6), we see that under the unrestricted model, with $\xi = I_n$, $\Gamma = \bar{r}Q + P_\perp$

$$H_{\tau_\varepsilon,\bar{r}} = P + (\bar{r}Q + P_\perp)^{-1} - P(\bar{r}Q + P_\perp)^{-1} - (\bar{r}Q + P_\perp)^{-1}P + P(\bar{r}Q + P_\perp)^{-1}P = \\ = P + P_\perp(\bar{r}Q + P_\perp)^{-1}P_\perp$$

Under the restricted model, i.e. using the random effect $\zeta = L^\top z \sim N(0, L^\top QL)$ we have $\Gamma = \bar{r}L^\top QL + I_{n-p}$, hence

$$H_{\tau_\varepsilon,\bar{r}} = P + L(\bar{r}L^\top QL + I_{n-p})^{-1}L^\top$$

Roughly speaking, the idea behind restricted regression is to rule out the bias in fixed effects estimation implied by spatial confounding. RSR can be extended to the multilevel case [65], in which e.g. random effects are defined at a higher scale than observations, as in Chapter 2.

Now, as it has been seen, under RSR the posterior means of covariate effects will approximate those of a nonspatial model. [25] argue that setting such a constraint would indeed yield a bias in $\mathbb{E}[\beta|y]$ if confounding occurs (intuitively, the expectation would be approximately the same of a model which ignores the existence of the spatial component). In terms of interpretation, this would mean that adding a spatial effect to the model would not significantly alter the estimated regression coefficients.

A different approach to deal with spatial confounding is based on adjusting the covariates rather than constraining random effects [26], by representing the former as the sum of a spatial and a nonspatial component; it is known in the literature as Spatial+. We employ a multilevel version of the variant of Spatial+ developed by Urdangarin et al., [81], which has the advantage of not requiring an explicit spatial model on X. In our multilevel framework, the value of the k-th covariate $X_{\cdot;k}$ observed in municipality j belonging to macro-area i is firstly decomposed as

$$x_{ij;k} = \bar{x}_{i;k} + \Delta x_{ij;k}$$

Being $\bar{x}_{i;k}$ the average value of the covariate observed in the i-th macro-area, i.e. at the same level of aggregation of the spatial effect, and is the (i,k)-th element of the matrix $\bar{X} := (\xi^\top \xi)^{-1}\xi^\top X$; $\Delta x_{ij;k}$ can be considered as the municipality-level noise.
In matrix form, this decomposition would be: $X = \xi\bar{X} + \Delta X$. Now, let us consider the eigendecomposition of the Laplacian $\mathbf{R}$:

$$\mathbf{R} := VLV^\top$$

where V is the matrix of eigenvalues and L is the diagonal matrix of eigenvectors. We know that L has $n - G$ non-null entries, corresponding to the first $n - G$ columns of V, while the last G columns of E are the null space of $\mathbf{D} - \mathbf{W}$.

With no lack of generality, let us consider

$$\bar{X} := Vb$$

where b is a appropriately chosen $n \times p$ matrix. Once we express $\bar{X}$ as a linear combination of the eigenvectors of the structure matrix of the spatial random effect, we find that the spatial component of $\bar{X}$ is determined by the last columns of V among those which do not correspond to null eigenvalues. We(can decompose $\bar{X}$ in

$$\bar{X} = \bar{X}^{(NS)} + \bar{X}^{(S)} + \bar{X}^{(0)}$$

Where $\bar{X}^{(NS)}$ is the nonspatial component given by the linear combination of the first $n - G - t$ eigenvectors Come posso rappresentare un set di indici in forma intervallare? (i.e. $\bar{X}^{(NS)} = (V)_{\cdot[1,n-G-t]}(b)_{[1,n-G-t]}$) and represents the nonspatial component of the covariates matrix, $\bar{X}^{(s)}$ is the linear combination

of the eigenvectors associated with the last t nonzero eigenvalues and represents the spatial component of the covariates matrix, and $\bar{X}^{(0)}$ is built with the G eigenvectors in the null space of the structure matrix and is constant within connected components. To remove spatial confounding, in the regression model we only take $\xi\left(\bar{X}^{(NS)} + \bar{X}^{(0)}\right) + \Delta X$ as covariates matrix. By doing so, covariates should not be spatially structured anymore. In [81] the term we label as $\bar{X}^{(0)}$ was actually removed; since in the cases covered therein G was equal to one, the eigenvector associated with the null eigenvalue was proportional to a vector of ones and its inclusion in the deconfounded covariates matrix would only scale it by an additive constant.

Chapter 3

The SchoolDataIT R Package

3.1 Introduction

The proper management of the public education system requires a full understanding of the territorial endowment in school infrastructure and the quality of education. Infrastructure endowment, in particular, is a direct area of policy intervention at various administrative levels. The depth of the link between the endowment in the material infrastructure and the quality of education is a matter of common knowledge and encompasses numerous dimensions of the education system, as highlighted in [9]. The evidence gathered therein across different countries sheds light on the relevance of several material factors on student achievements and education equity.

The first infrastructural dimension to be taken into account is the accessibility of schools and learning spaces, also in terms of school size, since less crowded schools both enforce the bond between students and the learning environment and allow for a more dense distribution of schools over the territory, which reduces the average travel distance from households. A closely related issue is classroom size, which is typically shown in the literature to negatively affect education quality [9].

Another dimension drawing attention from the literature is safety in school buildings, which can be assessed both with respect to outdoor hazards like pollution or natural events such as earthquakes, and in terms of indoor environmental quality, which can be summarized by factors such as illumination, indoor air quality (the main threat being the concentration of CO_2, which also can undermine student attention), air temperature and acoustic quality, which however may strongly depend on outside acoustic disturbances. Another element to be taken into account is the impact of health hazards on school attendance, which is also relevant in developed countries, mainly on the side of respiratory diseases. Lastly, it is worth remarking on the importance of adequate physical

extra-classroom spaces, such as IT laboratories, and recreational spaces like gymnasia or canteens, which intuitively allow for full-time schooling, which in turn is interpreted as a gain in school years attended by pupils.

The Italian school system offers a self-evident case for the significance of territorial disparities in education quality, both in terms of infrastructure endowment and student outcomes. Regarding the first case, [30] and [18] provide a detailed analysis of the distribution of school infrastructure on the national territory; Importantly, the northern regions show an advantage in terms of recreational spaces, learning spaces, safety certifications and school accessibility. In addition, [18] show that such infrastructural characteristics have an impact on the results of the students. Regarding student outcomes, it is worth noting that the North-South divide is widely acknowledged to shape dramatically the distribution of student performances, e.g., as shown in [2]. In particular, this disparity increases along the schooling process, implicitly suggesting that educational gaps tend to accumulate over time [62]. In addition to the North-South gap, evidence for spatial patterns in student outcome results can also be detected within the Northern and Southern macroregions and territorial clusters in both cases [6, 24, respectively]. Overall evidence suggests therefore the need for policy actions directed at improving the material conditions of schools in the most vulnerable areas.

Thus, allocating adequate resources is a sensitive challenge for policymakers, also considering the heterogeneous funding system of school buildings and the uneven spending capacities between regions [as in the case of Northern special statute regions, see 18].

Motivated by the previous considerations, we believe that a structured set of multidimensional data about the Italian school system gathered from several institutional sources, along with georeferenced information, would be a valuable tool to detect the main areas of vulnerability and to plan appropriate development policies across the country. To this aim, we have developed `SchoolDataIT`, a software written in the `R` programming language [71] which retrieves and harmonizes some relevant institutional databases at the territorial level of either municipalities (LAU hereinafter) and provinces [NUTS-3 henceforth, 27]. The `SchoolDataIT` package is intended as a contribution to a broader repository, namely the AMELIA platform (`https://grins.it/progetto/piattaforma-amelia`), an open-data platform designed to produce and harmonize high-quality statistical data and analyses, managed by the *Growing Resilient, Inclusive, and Sustainable* (GRINS) Foundation, a multidisciplinary initiative funded by the NextGenerationEU (NGEU) Recovery Plan [22, 23]. An example of package usage to build some simple data sets to be uploaded into AMELIA can be found in this GitHub repository.

Data providers are the Italian Ministry of Education (formerly MIUR, Ministry of Education, University and Research) [51], the Institute for the Evaluation of the Education System (hereinafter Invalsi) [45], the Italian National Institute of Statistics [ISTAT, 46, 48, 47], and the in-house company Infratel SPA on behalf of the Italian Ministry of Enterprises and Made in Italy [MIMIT, 44], which is responsible of implementing and managing the ultra-broad band

strategic plan. Since all of the data we take as input are open and publicly accessible, we retrieve them via web scraping, allowing for real-time updated inputs while requiring no storage space in the local machine of the user.

The `SchoolDataIT` software is currently available under version 0.2.4, released on March 28th 2025 on the Comprehensive R Archive Network (CRAN). To ensure constant package maintenance, experimental version 0.2.5 is hosted on GitHub.

The remainder of this chapter is structured as follows. In Section 3.2, we offer a concise yet comprehensive overview of the infrastructural state of Italian schools in light of the scientific literature on the national case and on official documents from the Ministry. In Section 3.3, we describe in detail the structure of the library and the most relevant functions made available for the users. In Section 3.4, we describe the datasets that can be accessed through the package, while including some relevant examples and potentiality. Finally, in Section 3.5, an empirical exercise involving the implementation of Bayesian spatial regression models is presented to investigate the student outcomes across the Italian territory.

3.2 School infrastructure in Italy

In this section, we briefly assess the current state of public school infrastructure in Italy using data provided by the Ministry of Education and processed through the `SchoolDataIT` package. For the sake of brevity, throughout the chapter we only comment on the main findings that can be inferred from the original data, while more detailed information is resumed in the tables reported in Appendix .1.

The first dimension we take into consideration is school size. According to [18], in the Italian context, Northern regions leverage on a marked advantage in terms of school surface per student, particularly for kindergarten and primary schools. This result is particularly interesting if we consider how Northern schools are more crowded than Southern ones and have a lower teachers/students ratio (in this regard, see also Section 3.4.3). On the one side, the number of municipalities hosting a primary or a middle school is relatively high. Indeed, according to the National School Registry 3.4.1, for school year 2021/2022, roughly 6748 (85.38% of the national total) and 5258 municipalities (66.52% of the total) host at least one primary and one middle school respectively. Conversely, high schools are located in only 1473 municipalities (18.64% of the total), thus having a more sparse distribution, especially in the peripheral inland. However, [18] showed that only in 139 municipalities (1.76% of the total) the travel time to the nearest school exceeds the threshold of 30 minutes. This finding is consistent with the smaller size of schools in such territories, which allows for a relatively widespread distribution of school buildings. If we move our focus to access to full-time schooling in primary schools, the North-South divide becomes an obvious cross-regional phenomenon. As reported in Table 4 in Appendix .1, among the 18 regions for which data are available, 8 out of the 9 regions with

the lowest values are located in the South (except Umbria), while 8 out of the 9 regions with the highest values are in the Center-North (except Basilicata).

Another fundamental factor in school accessibility is the availability of public transport. As declared by [51], see also Section 3.4.2, interurban and railway transport is considered available if the nearest hub is located within 500 meters from the school, while urban transport is considered available if the hub lies within a range of 250 meters. As documented in [30], Northern regions generally outperform Southern ones in terms of urban and interurban public transport availability, though significant differences are observed within macro-regions. For instance, within the Southern regions, Abruzzo owns the percentage of schools served by public transport systematically exceeding the national average, while Campania and Calabria appear to display the most vulnerable profile. The availability of urban, interurban, and disabled-people-specific transport at the regional level is shown in Table 5 in Appendix .1.

Regarding school building safety, one can consider at least two kinds of hazards. The first is pollution exposure. In particular, three main risk factors are explicitly monitored by the Ministry of Education, namely the proximity to either hazardous industries, pollutant waters, or sources of air pollution. These specific issues occur in a relatively small number of localized cases and would deserve a more dedicated analysis due to their severity. A general finding to be considered is that air pollution poses an important threat in terms both of health and physical well-being and education quality indeed, as recent evidence [10] shows that not only does the presence of particulate matter ($PM_{2.5}$) impact student outcomes, but the significance of this impact increases as the socio-economic status of students decreases.

Another serious hazard affecting the whole Italian territory is the unpredictable occurrence of an earthquake. Based on 2023 data, almost half of the school buildings are located in high or medium-high seismic risk areas[1]. An organic framework to assess the seismic risk of school buildings, integrating several extant methodological approaches is described in [19]. Tables 2 and 3 in Appendix .1 show the distribution of school buildings by the seismic risk of the relevant municipality and the number of schools located in high seismicity areas. The status of regions such as Basilicata, Molise, or Calabria appears particularly critical, especially in the latter case, with more than half of the buildings in high-risk areas.

Lastly, both [30] and [18] stress the importance of the endowment in learning and recreational spaces. Southern regions have a general disadvantage in the availability of both canteens and gymnasia, especially in the case of Calabria, Sicily, and Campania. The North-South divide becomes less distinct in the case of learning spaces. Indeed, for what concerns technical and IT rooms, this trend

[1] In Italy, the seismic risk of a given area is classified based on the relative peak ground acceleration (PGA). High seismicity areas: $\geq 0.25g$; medium-high seismicity areas: $[0.15g, 0.25g[$; medium-low seismicity areas: $[0.05g; 0.15g[$; low seismicity areas: $< 0.05g$, where g is the gravitational acceleration on Earth. For more details, see e.g. `https://rischi.protezionecivile.it/en/seismic/activities/emergency-planning-and-damage-scenarios/seismic-classification/`.

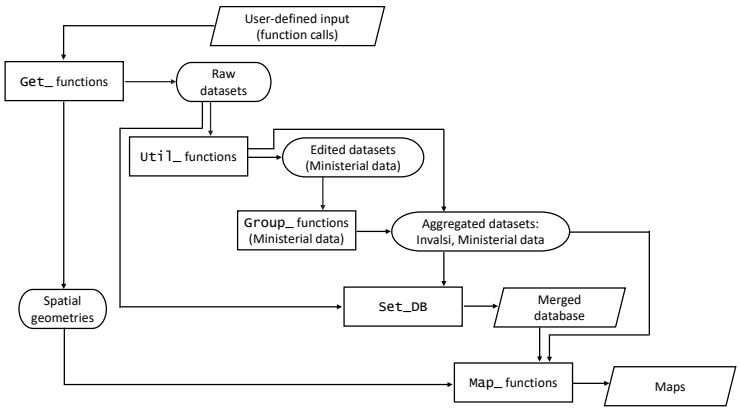

Figure 3.1: Flowchart of the package. Rhomboids denote inputs and `R` output objects; rectangles denote functions and rounded rectangles denote intermediate `R` objects.

is only observable in primary and middle schools, as we show in table 6.

Based on the information provided by the Ministry of Education, one can observe that the overall distribution of school infrastructure endowment is affected indeed by patterns of territorial vulnerability. Henceforth, one could then reasonably expect that these territorial disparities are reflected in terms of student outcomes, provided the role of school infrastructure in learning processes. Such assertion is confirmed by [16], who show that infrastructural variables can contribute to explaining part of the North-South divide in Programme for International Student Assessment (PISA) test scores [67].

3.3 Package workflow

The `SchoolDataIT` package is organised according to a chained sequence of steps. Except for the mapping functions, all outputs are `data.frame` objects, specifically structured as `tibbles` [64], thus fully compatible with the `Tidyverse` [89]. Figure 3.1 presents a flowchart illustrating the skeleton of the package.

The first step involves retrieving school system data from institutional sources through the `Get_` functions. The user specifies some key requests in function calls, such as the school year of interest for school buildings or student counts data. The software can thus navigate to the webpage of the data provider, inspect its HTML structure and identify the static links to the raw data to be downloaded. Raw data are then converted to `.csv` and eventually into `R` objects. In doing so, it is needful to recall that no storage space is required on the local machine of the user.

The main retrieval functions are the following; reported data availability is assessed on April 3rd 2025.

- `Get_DB_MIUR` for school infrastructure data, available for school years 2015/16, 2017/18, 2018/19, 2020/21, 2021/22 and 2022/23;
- `Get_Invalsi_IS` for Invalsi census data, available for school years from 2012/13 to 2023/24 except for 2019/2020;
- `Get_nstud` for student counts, available for school years 2015/16, 2017/18, 2018/19, 2020/21, 2021/22, 2022/23 and 2023/24;
- `Get_BroadBand` for the activation status of the ultra-broadband connection across single schools at a user-specified date;
- `Get_nteachers_prov` for teachers counts by province, available for the same school years as the school buildings data;
- `Get_Registry` for the National Schools Registry, available for school years from 2015/16 to 2024/25;
- `Get_School2mun` to map each school to the relevant administrative unit codes, available for the same years as the school buildings data.

The resulting objects are faithful to the original data published by providers, since at this stage data are not edited yet besides some manual corrections to municipality and province names needed for harmonising and mapping.

Indeed, data manipulation is reserved for the subsequent step, namely the `Util_` functions. One aim of these auxiliary functions is to transform input data into objects that can be handled with the next group of functions. The other aim is to perform data quality checks or editing. This group includes `Util_Check_nstud_availability` to check how many schools have available the student counts, `Util_DB_MIUR_num` to structure Boolean and numeric fields in the school buildings database or remove either observations with missing fields or fields with a given amount of missing observations, `Util_Invalsi_filter` to filter the Invalsi survey for the school year, grade and subject, and `Util_nstud_wide` to reshape the student counts dataset for it to have one school per row, compute average classroom size for each school grade, and filter out schools for which the classroom size is considered an outlier. Additional details on these functions are provided in Section 3.4.

Ministerial data are provided at the school level. `Group_` functions allow users to bring them to the same level of detail, namely at the LAU or NUTS-3 level. The function `Set_DB` merges one or more datasets from any previous step into a unique, aggregated database which can be considered as the final data output of the package workflow.

Lastly, the `Map_` functions render aggregated data with static or dynamic choropleth maps. Notice that the former employs the `ggplot2` [90] environment for graphical representation, allowing a simplified export. Interactive maps, obtained through the `leaflet` [20] and `mapview` [4] libraries, preserve all the

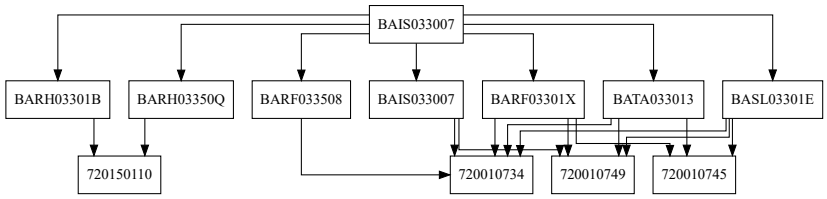

Figure 3.2: Example of the codes of all the schools (middle nodes) and the buildings in which they are located (bottom nodes), pertaining to the same reference institute (top node)

information of the dataset to be rendered. Spatial geometries used for mapping are provided by the Italian Institute of Statistics through specific shape files [47].

3.4 The Italian Education System Data

This section describes the main datasets currently retrieved and handled by the package. For the school infrastructure system evaluation, the most valuable public source of information is the Unique School Data Portal [51], an open data portal managed by the Italian Ministry of Education according to Law 107/2015 [54]. The National Schools Registry (Section 3.4.1), the school buildings database (3.4.2) and the counts of students and teachers (3.4.3) are all provided through this website. Another relevant aspect of school infrastructure assessment is the implementation of ultra-broadband connection, whose timeline, available through the ultra-broadband activation dashboard, provided by [44] has been included in the package as well. Lastly, Invalsi censuary survey data [45] have been included to assess education quality (3.4.5).

Italian schools are officially identified by a 10-digit alphanumeric code. In Fig. 3.2 we show an example of the identifiers (ID) hierarchy. The top node is the reference institute ID; intermediate nodes are the school institutes IDs; whereas bottom nodes are the school buildings IDs. For instance, under the same reference institute, two schools are located in the municipality of Casamassima (BA, code 72015) while five other ones are distributed across three buildings in the municipality of Acquaviva delle Fonti (BA, code 72001).

3.4.1 National Schools Registry

The National Schools Registry includes the list of all public and private schools on the national territory. Due to the completeness of the records, this dataset is used as the baseline to harmonise other objects defined at the school level. The function `Get_Registry` downloads the dataset. Notice that the relevant municipality of each school is not identified by its official administrative code but only by the cadastral code. To fill this gap, the function `Get_School2mun`

associates each school listed in this registry with the relevant administrative (LAU and NUTS-3) codes [48].

3.4.2 School Buildings

This database covers several infrastructural dimensions, accounting for a total of about 90 variables in the last available year:

- Environmental context of school buildings
- Accessibility through private or public transport, namely whether a building lies within a given range (e.g. 250 or 500 meters) from a transport hub
- Environmental or administrative restrictions
- School area surface and building volume
- Intended use of learning and recreational spaces
- Overcoming architectural barriers (e.g. the presence of external ramps or stairlifts)
- Building and adaptation period
- Various information regarding heating systems
- Measures and devices to reduce energy consumption
- Acoustic insulation
- Static testing certification and seismic design

Observations are detailed at the level of school buildings. For this reason, the database embeds a standalone registry different from the National Schools Registry mentioned in the previous paragraph.

The input dataset downloaded with `Get_DB_MIUR` includes about $60,000$ observational units. Most variables are binary (Y/N), denoting whether a given feature occurs in a school building or not, and encoded as strings.

The function `Util_DB_MIUR_num` converts strings to Boolean or numeric values when necessary. For some variables, there is a high number of missing values. For example, in school year 2022/23, the field denoting whether a school is reached by a bicycle lane is missing for 38.7% of high schools, 44.1% of primary and 45.9% of middle schools. The user may choose to remove either the fields with a given number of missing records ($20,000$ by default) or the units with at least one missing variable (not active by default).

Observations can be aggregated with the function `Group_DB_MIUR`. Numeric and Boolean variables are summarized by their mean and qualitative variables by their mode. Since territorial averages provide no information about missing

values, by default the function returns two additional data frames providing the number of missing observations of each variable per area.

Finally, for better insight into the general infrastructural state, we add the Inner Areas taxonomy, published by the Italian Institute of Statistics (ISTAT) and updated every six years [46]. It divides Italian municipalities into six classes: A, B and C are considered central areas, while D, E, and F classes are labeled as "inner" (i.e. peripheral) areas. Class A identifies standalone pole municipalities, characterized by a comprehensive and self-sufficient combination of school, health, and transport infrastructure [46]; class B identifies inter-municipality poles, i.e. clusters of neighbouring municipalities which, taken together, fulfill the requirements of pole municipalities. The remaining classes are defined based on increasing road travel time to the closest pole: Class C: $0' - 27'42''$; Class D: $27'42'' - 40'54''$; Class E: $40'54'' - 1h\,6'54''$; Class F: $> 1h\,6'54''$.

In Figure 3.3 we show the percentage of schools served by public transport in 2022/23 at the province and municipality level, in this latter case only for the Apulia region, which is the region with the highest share of municipalities hosting at least one high school (124 over 257). As mentioned in Section 3.4, though northern and central regions have a higher proportion of schools served by urban public transport, regions like Abruzzo in the South or Veneto and Emilia-Romagna in the North are in contrast the general trend. In the provinces of Aosta, Trieste, La Spezia (North), Massa (Center) and Chieti (South) all schools are reached by public transport, while this percentage is higher than 95% in the provinces of Pavia, Bergamo (North), Pesaro-Urbino, Pisa, Lucca and Latina (Center). On the other hand, in the province of Salerno in Southern Italy only 0.07% of schools is served by public transport; this percentage is lower than 40% in the provinces of Ferrara and Pordenone in the North and Crotone, Foggia and Naples in the South.

The code to download the raw input dataset and display these maps and all the following ones is in the Supplementary Material.

3.4.3 Number of students and teachers

The Ministry of Education also publishes the counts of students per school grade for every school on the Italian territory and the counts of teachers for every Italian province. These datasets have the same temporal dimension as the school buildings database. classroom size is indeed useful information in the assessment of education quality, which is typically acknowledged to improve as classroom size decreases [15, 17]. In the case of Italy, however, caution is needed when studying the relationship between classroom size and student outcomes at the aggregate level [3]. In our view, an important factor to consider is how classroom size reflects the degree of centrality of municipalities. As it can be seen in Figure 3.4 as an example for the last year of middle schools, peripheral areas, usually characterized by lower student outcomes, have less crowded classrooms. We will have a deeper look at the association between classroom size and education quality in Section 3.5.

The function `Util_nstud_wide` rearranges the input dataset into a wide

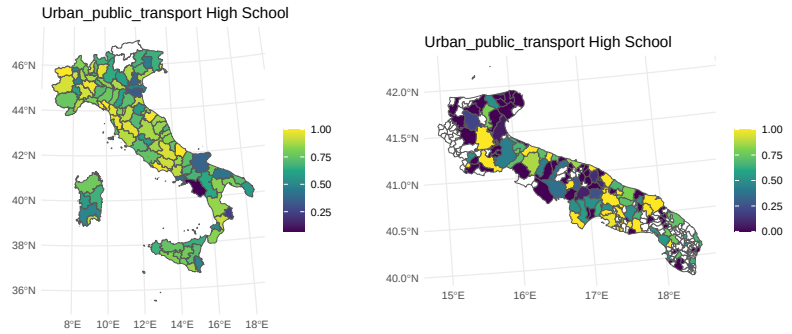

Figure 3.3: Province-level and municipality-level percentage of high schools served by public transport in 2022/23, on the whole national territory and in the Apulia region respectively. Data of the Trentino-Alto Adige region are not provided by the Ministry.

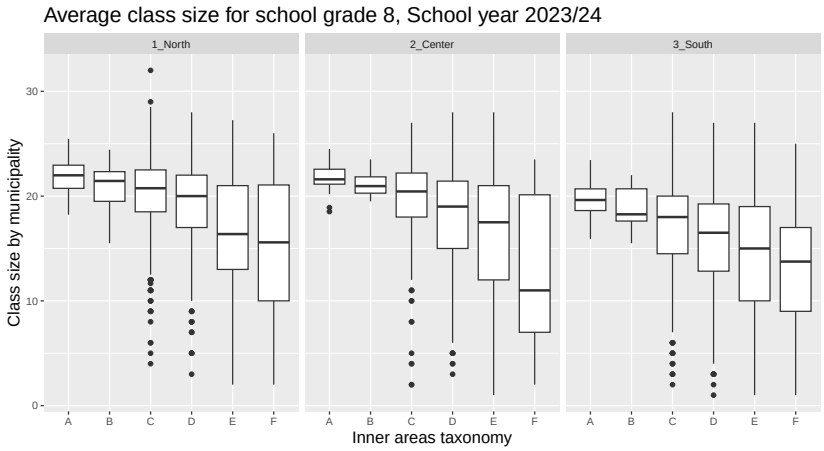

Figure 3.4: Average municipality-level classroom size by Inner Areas taxonomy in 2023/24, last grade of middle schools. Inner area taxonomy follows a descending order of centrality: A - standalone infrastructural poles; B - inter-municipality infrastructural poles; C - belt municipalities; D - intermediate areas; E - peripheral; F - ultra-peripheral.

format in which each row corresponds to a school and computes the average classroom size per school for each educational grade. National regulation sets upper classroom size limits of 25, 26 or 27 students in primary, middle and high schools respectively [50, Art. 5] other than lower limits of 15 students (8 for multi-year classes) in primary schools and 18 students in middle schools through the Decree n.90/2023 of the Ministry of Education [53, Artt 10, 11]. A framework of waivers is established by the Ministry Decree n.90/2023 [53], regarding cases of low Economic, Social and Cultural Status (ESCS) scores, high school withdrawal rate or high depopulation. However, the range of observed classroom sizes is often wider than the general rule, especially in high schools. In the latter case, taking the school year 2022/23 as an example, the number of schools with classroom size ≥ 40 students was equal to 9, 8, 8, 18 and 1 for the five high school grades respectively, over a total of 6455 schools. To remove values considered extreme, the user can set an upper and a lower boundary of acceptance in terms of classroom size either at the level of whole schools or single school grades. In the former case, only schools whose average classroom size (computed across all grades, e.g. for middle schools the average of 6th, 7th and 8th grades) exceeds the acceptance boundary are removed from the dataset, while in the latter case removal applies to all schools where classroom size exceeds the boundary in any grade. For what concerns primary schools, it is also possible to download student counts by type of schooling time, namely distinguishing between full-time and half-time (only morning) schooling.

To monitor statistical data quality, the function `Util_nstud_check` computes, for all municipalities and provinces, the percentage of schools listed in the National Registry for which the count of students is available.

The function to aggregate school-level data is `Group_nstud`.

Teacher counts, instead, are only available at the province level. The average number of teachers per student and per class can be computed with the function `Group_nteachers4stud`. In Figure 3.5 we render the average classroom size in the 2nd year of high school and the average number of teachers by student in the year 2022/2023. classroom size is higher in densely populated areas, such as the Po Valley and the surroundings of Rome and Naples, while it is smaller in most of the South, especially in the Apennines and in Sardinia. The teacher/student ratio follows a similar distribution.

3.4.4 Ultra - Broadband connection in schools

This dataset consists of the list of schools of the National Ultra-Broadband Plan, approved by the Ministry of Economic Development with the decree of 07/07/2020 [49]. The Plan aims at providing 32.164 schools with internet connection with a maximum speed of 1 gigabit/second and a symmetric minimum guaranteed speed of 100 megabits/second until the peering is reached. Data are updated monthly [44]. In Table 7 in the .1 we show the number of schools in which the ultra-broadband was activated in different years for all regions (one school in Trentino Alto Adige had a broadband connection before 2020). The function to download this dataset is `Get_BroadBand`; the `Date` argument speci-

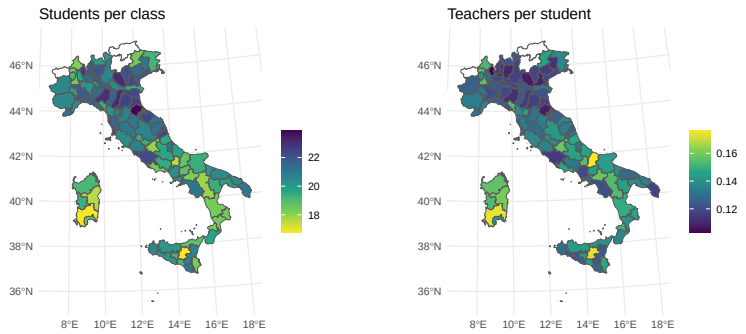

Figure 3.5: Average province-level classroom size in the 2nd year of high school and province-level teachers/students ratio in high schools in 2022/23. Data are not available for the Trentino-Alto Adige and Aosta Valley regions. Additionally, schools with average classroom size of less than 10 or more than 40 students in any grade have been filtered out.

fies the reference date for checking whether the ultra-broadband connection was activated or not in each school.

3.4.5 Invalsi census survey

To develop a spatially homogeneous indicator of education quality, Italian law No. 176/2007 [52] mandates the Italian Institute for the Evaluation of the Education System (INValSi) to assess the skills of students through a specific test. The test currently covers four subjects, Italian, Mathematics, English reading, and English listening, and is carried out yearly in the 2nd, 5th, 8th, 10th, and 13th school grades. The Invalsi Institute publishes several open datasets [45], the widest class being that of sample surveys, which also includes anonymized microdata regarding single students. The other class of datasets consists of census surveys, detailed at either municipalities or provinces. Regarding municipality data, for privacy reasons only the municipalities with at least two schools of the same order are included in the survey; otherwise identifying average Invalsi scores of single schools would be easily possible. In this package, we focus on the census dataset since it provides more spatial information (sample datasets providing no territorial information other than the region) and is, in our judgment, more suitable for spatial analysis.

Consistently with OECD standards [66] the score is expressed through the weighted likelihood estimator (WLE) of student ability defined by a Rasch psychometric model, whose basic idea is that the probability that a generic student i provides the correct answer to a generic item (test question) j depends on two variables, namely the student ability b_i and the item difficulty d_j. The

relationship can be expressed as

$$\text{Prob}\{\text{student } i \text{ answers correctly item } j\} = \frac{e^{b_i - d_j}}{1 + e^{b_i - d_j}}$$

For interpretational reasons, the estimator of b_j is scaled to a global mean of 200 points and a global between-students standard deviation of 40 points. The advantage of this model is isolating the ability of students from the intrinsic difficulty of items. For primary schools only, the percentage of sufficient tests is also reported. Scores are already corrected from the effect of cheating, which would otherwise hinder their meaning, other than shrinking their variance. The functions to download and filter the Invalsi database per school year, grade and subject are respectively `Get_Invalsi_IS` and `Util_Invalsi_filter`. No data quality checks are deemed necessary as this dataset is already carefully processed by the Invalsi Institute.

A case-study with more details on the Invalsi census survey is provided in Chapter 4.

3.5 Student outcomes in Mathematics and classroom size: an example using the SchoolDataIT package

Here we provide an example of spatial statistical application to the data covered by the `SchoolDataIT` package. Following Section 3.4.3, suppose the user is interested in studying to what extent classroom size is associated with student outcomes. We can first map the two variables. Concerning the last year of middle school in 2023/2024, we regress the Invalsi score in Mathematics on the average classroom size at the municipality level. To ease model results interpretation, classroom size is scaled to zero mean and unit variance. Additionally, schools with an average class size of less than 10 or more than 40 students have been removed from the dataset. In this case, observational length is equal to $n = 780$ municipalities, i.e. those for which Invalsi scores and classroom size were both available at 2025/04/03.

OLS regression would result in an expected effect of classroom size equal to 4.153, with standard error 0.407. If no additional information is taken into account, classroom size would then appear to have a positive and statistically significant relationship with Invalsi scores.

A closely linked result, namely the significantly positive association of Invalsi scores with the students/teacher ratio, was noticed by [8], always in the context of middle schools, and attributed to the impact of schools reputation on their attractiveness. Interestingly, when an instrumental variable relating to teachers' mobility was taken into account in their regression model, the estimated effect of the student-to-teacher ratio was not significant anymore.

Given the greater amount of information at our disposal, it would be naïve to limit the analysis to this amount of information. First, considering what we

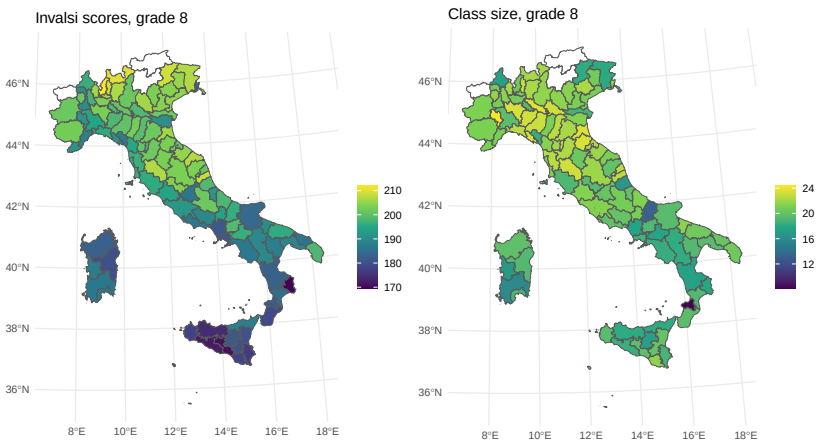

Figure 3.6: Invalsi score in Mathematics and average classroom size, last year of middle school, school year 2023/24. Trentino-Alto Adige and Aosta Valley are not included due to lack of classroom size data.

have seen in 3.4, the user may be interested in adding the inner areas taxonomy as an explanatory variable in the simplest possible way, namely classifying municipalities among central (A, B, C) and inner (C, D, E) areas. Moreover, the territorial structure of the dataset suggests to include spatial terms in the regression model. Considering the number of municipalities for which data are available (780 over a national total of 7901), the spatial structure is rather sparse and we would need to define some neighbouring rules in alternative to shared borders. To overcome this issue, we choose to treat the municipality-level average Invalsi score as a point-referenced process, thus assuming that the data-generating process is defined on a continuous spatial domain. Specifically, we postulate that the locations at which this process is observed are the centroids of municipalities as defined on January 1st, 2023. Spatial information is taken into account through a linear spatial trend and a spatially structured Gaussian process $u(s)$, with $s \in [1, 780]$, whose autocorrelation decays as distance increases. The model becomes thus:

$$y(s) = \beta_0 + \beta_{\text{nstud}} X_{\text{nstud}}(s) + \beta_I X_I(s) + \beta_\ell \ell(s) + \beta_\phi \phi(s) + u(s) + \varepsilon(s) \quad (3.1)$$

where β_0 is the intercept, X_{nstud} is classroom size, X_I is the dummy for the inner areas taxonomy (1: inner area, 0: central area), ℓ is the longitude of municipality centroids, ϕ is the latitude; β terms are covariates effects; ε is a Gaussian IID error such that $\varepsilon \sim N(0, \sigma^2_\varepsilon I_n)$. u follows *a priori* a Normal distribution with mean zero and covariance matrix whose elements depend on the distance between the two corresponding points, but not on the direction of their link (isotropy); second-order stationarity is additionally assumed for u [7, Section 2.1]. Based on these two assumptions, we assign u the Matérn covariance function, which depends on a global variance σ^2 and a range parameter r. In

detail, the Matérn covariance function between two generic sites i and j is given by:

$$\sigma_{ij} = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} (\kappa d_{ij})^{\nu} K_{\nu}(\kappa d_{ij}) \tag{3.2}$$

where d_{ij} is the Euclidean distance between the i-th and j-th location, σ^2 is the global (common across locations) variance, κ is a scale parameter and ν controls smoothness. These parameters are linked to the range r since in this model $r = \frac{\sqrt{8\nu}}{\kappa}$. In this example, we keep fixed $\nu = 1$. $K_{\nu}(x)$ denotes the modified Bessel function of the second kind [1, Section 9.6]:

$$K_{\nu}(x) = \frac{\pi \left(I_{-\nu}(x) - I_{\nu}(x)\right)}{2sin(\nu\pi)}$$

Where $I_{\nu}(x)$ denotes the modified Bessel function of the first kind, defined in turn as a solution to the equation $I_{\nu}(x) = f : x^2 + \frac{d^2 f(x)}{dx^2} + x\frac{df(x)}{dx} - (x^2 + \nu^2)f(x) = 0$:

$$I_{\nu}(x) = \sum_{k=0}^{\infty} \left(\frac{x}{2}\right)^{2k+\nu} \frac{1}{k!\Gamma(k+\nu+1)}$$

Now, [88] shown that a stochastic process $u(s)$ with this kind of covariance function is a solution to the SPDE (we limit our illustration to the bidimensional case):

$$(\kappa^2 - \Delta)^{\frac{\nu+1}{2}} \tau u(s) = \epsilon(s) \tag{3.3}$$

Where Δ is the Laplacian operator and $\epsilon(s)$ denotes a spatial Gaussian white noise process; this equation does describe indeed a Gaussian autoregressive process on an infinite lattice.

Parameters ν and with κ are linked to the global variance σ^2 through the relationship:

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu+1)4\pi\kappa^{2\nu}\tau^2}$$

Fitting spatial models such as the one in equation 3.1 implies the inversion of the covariance matrix of u, which is typically dense as it can be seen from equation 3.2. In geostatistical literature, this issue is typically referred to as the big-n problem [56] as the inversion operation has a computational cost in $\mathcal{O}(n^3)$.

A potential solution is offered by the Stochastic Partial Differential Equation (SPDE) approach for Gaussian fields. [59] show, indeed, that processes with Matérn covariance function can be represented as Gaussian Markov Random Fields defined on a discrete spatial domain, which is determined via Delaunay triangulation over a manifold technically known as mesh.

The point-referenced field $u(s)$ is replaced by Az where A is an $n \times N$ matrix to project observation locations onto the mesh, and z is a Gaussian Markov

	mean	s.d.	$Q_{0.025}$	$Q_{0.975}$
β_0 (Intercept)	195.103	1.080	192.918	197.211
β_{nstud} 0.192	0.317	-0.429	0.813	
β_I	-2.329	0.719	-3.743	-0.920
β_ϕ	8.997	1.044	6.928	11.061
β_ℓ	1.865	1.038	-0.174	3.934

Table 3.1: Estimated effects of classroom size, inner area dummy, latitude and longitude on Invalsi scores in Italian, last year of middle school, under model 3.1

random field of length N defined on the mesh which, in our case, has $N = 1384$ nodes. Markov properties, together with the Normal distribution, ensure the precision matrix of z has a number of nonzero entries in the order of N [73], as seen in Section 1.2.

Mapping the Matérn field onto a GMRF shrinks the computational cost to $\vartheta(N^{3/2})$. TBD: AGGIUNGERE FONTE O DIMOSTRARE

For model fitting, we firstly need to define the mesh and index all its nodes; in our case the mesh is chosen to have more nodes than the observation points. This procedure can be entirely handled internally to `R-INLA`.

The hierarchical model in equation 3.1 also requires prior assumptions on the distribution of its hyperparameters. Error precision σ_ε^{-2} is assumed to follow a Gamma distribution with shape parameter 1 and rate parameter $5 \cdot 10^{-5}$. Moreover, we define a penalized complexity (PC) prior [77] on the range r and the global standard deviation σ of the latent spatial field. The behaviour of these distributions is described in [29]. We select two fixed values, σ_0 and r_0 such that, for two fixed values p_σ and p_r, $prob\,(r < r_0) = p_r$ and $prob\,(\sigma > \sigma_0) = p_\sigma$.

Based on prior knowledge and ignoring the information available from exploratory data analysis, we assume that the range, namely the distance at which the correlation of the random fields is shrunk under a 0.10 threshold, is smaller than 300 kilometers with 5% probability, and the standard deviation of the random field is higher than 4 points with 5% probability. Again, for the sake of model results interpretation, classroom size, latitude and longitude are scaled to mean 0 and variance 1.

The summaries of covariate effects are reported in Table 3.1. Employing this amount of information, the effect of classroom size no longer appears to be significant, while belonging to an inner area still implies an expected disadvantage of 2.329 points in Invalsi scores compared to central areas (either infrastructural poles or municipalities close to them). The evidence for the North-South divide is very strong, as the current model suggests that being located one standard deviation of the northing distribution (≈ 291.25 km) further north than a reference location implies an expected advantage of 8.997 Invalsi points. To visualize the extent to which the spatial structure influences Invalsi scores, we plot the expected value of the linear trend $(\beta_\ell \ell + \beta_\phi \phi)$ and the latent Gaussian process (u) in Figure 3.7.

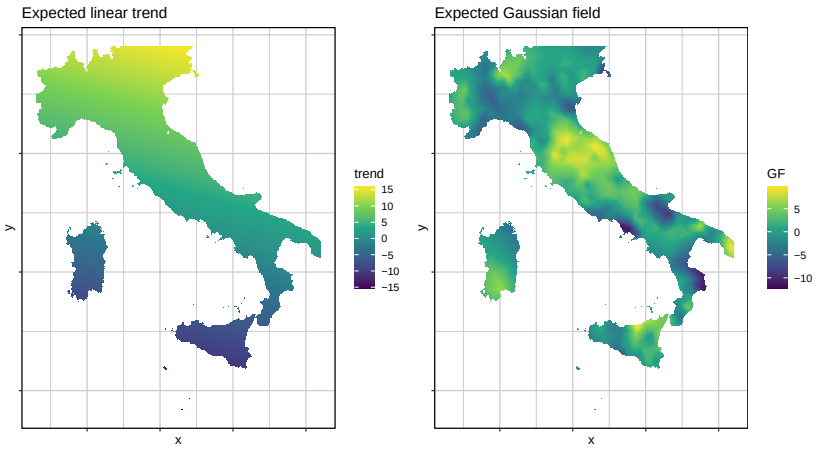

Figure 3.7: Posterior expectation of the linear trend on the left and of the spatial latent variable on the right, the latter being added to the model to explain spatial variability in Invalsi scores after controlling for the linear trend and the explanatory variables.

The darker zones in the right panel (lower values of $\mathbf{E}[u|y]$) can be interpreted as areas of educational vulnerability net of classroom size, general infrastructural conditions, and net of the North-South trend as well. Most critical areas include the urban area of Naples and the upper Ionian coast in Calabria. Conversely, brighter areas represent relatively advantaged territories, such as much of Central inland and the Salento Peninsula.

This simple example highlights that assessing the impact of classroom size on student outcomes is somewhat less simple than it may initially appear and warrants a multidimensional analysis. Even though only aggregated data are taken into account, auxiliary information such as the degree of centrality of a municipality or the spatial location becomes thus crucial.

3.6 Concluding remarks

The package `SchoolDataIT` allows `R` users to automatically construct an organic database by combining from different sources those open data we deem to be the most informative about the Italian education system and the state of school infrastructure in Italy, covering some of the Ministerial open data sets (e.g., the Schools Registry, the school buildings database, students and teachers counts), the Invalsi census survey, and the registry of Ultra-Broadband activation status.

One key feature we emphasize is the relevance of the spatial context, hence territorial units of aggregated data can be easily associated either with their boundaries or centroids. With this library, we aim to provide an insight into the state of the Italian education system as clearly as possible, to ease statis-

tical analysis on its main aspects, also employing a spatial framework, and to identify areas of education vulnerability following either a unidimensional or a multidimensional approach. With this in mind, exploratory analysis should be facilitated by mapping functions. While this paper presents a cross-sectional data usage example, panel analysis is possible as well, considering a current observation length of six years.

Although the present version of the package covers a given amount of data, implementing any additional function to retrieve, edit and structure further data sets would not imply a significant increase in file weight for the package. Therefore, a possible line of development of this package would be to plug in additional data sets.

Lastly, the package appeals to generic `R` users, from whom the various functions are designed to require as little effort as possible. Despite our efforts for maintaining a user-friendly perspective, the final output of this package is data objects defined coherently with the `Tidyverse` environment to ensure object portability and ease of use of the present data also outside `R`.

.1 Tables

	Primary		Middle		High	
Seismicity	buildings	% tot.	buildings	% tot.	buildings	% tot.
High	1375	7.53%	877	8.45%	662	6.1%
Mid-High	6999	38.33%	3938	37.96%	4267	39.31%
Mid-Low	7197	39.42%	3970	38.27%	4402	40.55%
Low	2687	14.72%	1589	15.32%	1524	14.04%

Table 2: Seismic risk classification of municipalities hosting school buildings

	Primary		Middle		High	
Region	buildings	% tot.	buildings	% tot.	buildings	% tot.
Abruzzo	87	19.46%	62	22.46%	40	17.24%
Basilicata	99	39.76%	73	40.33%	61	33.7%
Calabria	608	60.5%	399	60.73%	234	52.58%
Campania	191	10.43%	140	13.17%	143	12.25%
Emilia - Romagna	1	0.09%	1	0.16%	0	0%
Friuli - Venezia Giulia	42	8.47%	24	10.04%	16	6.3%
Lazio	67	4.87%	34	4.32%	28	3.16%
Liguria	0	0%	0	0%	0	0%
Lombardia	1	0.04%	0	0%	0	0%
Marche	2	0.38%	2	0.67%	0	0%
Molise	40	30.3%	23	25.27%	21	22.11%
Piedmont	0	0%	0	0%	0	0%
Apulia	16	1.65%	14	2.34%	11	1.19%
Sardinia	1	0.16%	2	0.41%	0	0%
Sicily	131	7.78%	60	6.47%	48	4.75%
Tuscany	0	0%	0	0%	0	0%
Umbria	52	14.44%	25	12.69%	36	20.11%
Aosta Valley	0	0%	0	0%	0	0%
Veneto	37	2.26%	18	2.05%	24	2.97%

Table 3: School buildings located in high seismicity municipalities, both in absolute numbers and as a proportion of the regional total

Region	HT Students	FT Students	% Full Time
Abruzzo	37956	11796	23.71%
Basilicata	9660	10275	51.54%
Calabria	56018	19730	26.05%
Campania	181001	47326	20.73%
Emilia - Romagna	78588	95936	54.97%
Friuli - Venezia Giulia	23497	19326	45.13%
Lazio	86366	131766	60.41%
Liguria	22811	27457	54.62%
Lombardia	172638	219666	55.99%
Marche	39288	19837	33.55%
Molise	9281	1049	10.15%
Piedmong	70881	88229	55.45%
Apulia	126995	29939	19.08%
Sardinia	32316	21912	40.41%
Sicily	176058	23811	11.91%
Tuscany	57331	77632	57.52%
Umbria	23206	10463	31.08%
Veneto	111004	78964	41.57%

Table 4: Number of primary school students attending either full time (FT) or half time (HT) schooling and proportion of the former over the total

Region	Urban			Interurban			Disabled people		
	High	Middle	Primary	High	Middle	Primary	High	Middle	Primary
Abruzzo	89.18%	60.44%	63.12%	83.98%	67.03%	64.48%	49.78%	65.57%	68.55%
Basilicata	69.83%	64.25%	67.34%	81.56%	56.42%	53.63%	45.81%	67.04%	72.58%
Calabria	74.61%	40.28%	37.06%	68.54%	34.92%	30.97%	19.78%	51.45%	50.45%
Campania	47.07%	45.07%	45.13%	46.98%	34.26%	30.84%	19.23%	43.92%	42.41%
Emilia - Romagna	68.17%	50.56%	52.71%	66.39%	54.55%	49.01%	24.32%	60.13%	58.31%
Friuli - Venezia Giulia	72.4%	56.9%	50.71%	72%	51.88%	47.88%	44.4%	63.6%	65.05%
Lazio	80.91%	66.54%	67.23%	49.09%	41.86%	38.98%	34.89%	48.47%	49.64%
Liguria	84.38%	80.82%	82.24%	53.52%	53.47%	44.96%	29.3%	80.82%	74.78%
Lombardia	80.76%	53.94%	52.84%	76.27%	50.1%	47.61%	40.15%	48.35%	48.53%
Marche	86.14%	63.97%	73.24%	69.88%	59.26%	51.42%	47.89%	68.69%	71.73%
Molise	76.84%	40.66%	38.64%	41.05%	39.56%	43.18%	67.37%	45.05%	49.24%
Piedmont	86.58%	50.35%	51.78%	77.09%	61.41%	53.72%	33.87%	63.21%	60.16%
Apulia	55.53%	50.85%	55.71%	57.13%	33.84%	33.61%	31.13%	69.39%	73.19%
Sardinia	69.57%	42%	46.09%	62.01%	49.69%	50.47%	52.17%	47.19%	53.91%
Sicily	72.74%	54.43%	57.71%	52.84%	32.83%	30.97%	34.33%	44.17%	45.5%
Tuscany	87.7%	71.08%	68.64%	63.8%	60.67%	54.93%	36.07%	56.97%	53.99%
Umbria	79.21%	56.41%	65.17%	43.26%	41.54%	40.73%	32.02%	57.44%	62.36%
Aosta Valley	100%	72.73%	78.05%	96.88%	77.27%	63.41%	100%	100%	80.49%
Veneto	65.43%	47.54%	46.18%	71.13%	51.54%	45.08%	23.17%	36.46%	36.45%

Table 5: Proportion of schools served by either urban or interurban public transport or by specific transport dedicated to disabled people

	IT classrooms			Technical classrooms		
Region	High	Middle	Primary	High	Middle	Primary
Abruzzo	68.33%	72.64%	59.77%	70.14%	51.89%	38.51%
Basilicata	76.73%	75%	64.02%	65.41%	62.9%	37.8%
Calabria	93.92%	65.87%	56.53%	91.71%	41.98%	28.15%
Campania	64.96%	75.54%	62.03%	67.43%	58.99%	42.12%
Emilia - Romagna	69.61%	73.68%	66.48%	67.31%	67.63%	53.19%
Friuli - Venezia Giulia	73.53%	69.35%	57.92%	75.21%	66.67%	38.18%
Lazio	75.62%	82.25%	60.09%	83.12%	71.86%	44.13%
Liguria	89.45%	82.46%	68.64%	84.77%	56.87%	41.13%
Lombardia	72.37%	77.15%	72.92%	74.46%	70.67%	47.27%
Marche	74.7%	72.04%	65.06%	75.3%	64.16%	49.6%
Molise	60%	72.6%	58.72%	65.71%	53.42%	35.78%
Piedmont	72.19%	74.08%	66.92%	73.18%	61.76%	36.48%
Apulia	77.03%	79.18%	65.35%	70.95%	66.59%	38.28%
Sardinia	73.63%	67.94%	57.52%	78.14%	54.7%	40.05%
Sicily	78.27%	68.35%	51.14%	74%	51.69%	33.07%
Tuscany	73.44%	72.01%	62.24%	71.6%	66.67%	41.31%
Umbria	70.55%	63.09%	56.08%	78.08%	53.02%	34.12%
Aosta Valley	81.82%	100%	71.01%	75.76%	94.44%	47.83%
Veneto	70.28%	68.09%	63.73%	72.67%	59.87%	34.56%

Table 6: Proportion of schools endowed with IT and technical classrooms

Region	Schools	N.A.	% N.A.	A. 2020/before	A. 2021	A. 2022	A. 2023
Abruzzo	971	269	0.28	0	379	242	81
Basilicata	524	174	0.33	0	87	186	77
Calabria	1532	992	0.65	0	67	357	116
Campania	2922	1195	0.41	0	696	788	243
Emilia - Romagna	1759	942	0.54	75	360	201	181
Friuli - Venezia Giulia	926	433	0.47	315	32	40	106
Lazio	2444	843	0.34	0	542	778	281
Liguria	878	345	0.39	0	194	250	89
Lombardia	3989	935	0.23	0	784	1489	781
Marche	1113	541	0.49	0	268	254	50
Molise	305	129	0.42	0	30	100	46
Piedmont	2295	837	0.36	0	465	675	318
Apulia	1869	327	0.17	0	866	560	116
Sardinia	1350	939	0.70	0	20	211	180
Sicily	3259	936	0.29	0	889	1137	297
Tuscany	2119	631	0.30	0	531	658	299
Trentino-Alto Adige/Südtirol	303	37	0.12	1	210	24	31
Umbria	584	323	0.55	0	28	60	173
Aosta Valley	199	38	0.19	0	89	30	42
Veneto	2609	601	0.23	0	727	858	423
TOT	31950	11467	0.36	391	7264	8898	3930

Table 7: Ultra-broadband activation progress; N.A. = "not activated" (by the end of 2023); last 4 columns report the number of schools in which the ultra-broadband was activated in different years for different regions.

Chapter 4

Analysis of High Schools Invalsi Scores: a Spatial Approach

4.1 Introduction

As it has been stressed out in 3.6, territorial disparities in the Italian public education system are a severe and widely recognised issue. Exploratory analysis of school infrastructure endowment highlights a North-South divide encompassing several infrastructural dimensions.

However, comparing student outcomes across a country's complex geography is not a trivial question. To this aim, a framework to define standardised and spatially homogeneous indicators has been developed by the OECD throughout the Programme for International Students Assessment [PISA, 66]. In Italy, this task is attributed by law [52] to the Institute for the Evaluation of the Education System.

Indeed, territorial gaps in Invalsi scores are immediately evident and have been noticed to expand throughout the schooling process [62], the gap in high school scores being a matter of particular concern. Considering analyses carried out at the individual (student) level for both PISA and Invalsi scores, a significant effect is often associated with North - Centre - South dummy variables [as in e.g. 36, 16, 2, 37] unless more explanatory variables regarding the labour market and socio-demographic dynamics are taken into account in relatively complex econometric models [16]. Additionally, [2] partition the data set of Invalsi scores (last year of middle school) among Northern, Central and Southern Italy, running three different regression models, in which the intercepts range almost 11 points apart in absence of explanatory variables and as far as almost 14 points apart when some explanatory variables are introduced (at the time, Invalsi scores were expressed in a $[0 - 100]$ points range). A slightly different

approach employs regression models with region-specific intercepts, allowing a higher amount of geographical information [63] (in this case working with PISA scores); estimated intercepts display a clear territorial pattern, with all Northern regions exceeding the nationwide average and all Southern regions except for Apulia and Basilicata below it.

In this chapter, we propose a spatial modelling framework for the average Invalsi scores for Italian municipalities. We focus on the second year of high school (10-th school grade), being it the last year of the compulsory education cycle for which Invalsi tests are designed (the last year of high school is beyond the compulsory education cycle). The subjects for which the test is designed for the school grade in scope are Italian and Mathematics.

We explore the association of Invalsi scores with the infrastructural state of municipalities in terms of their centrality degree expressed with the inner areas taxonomy [46], availability of ultra-broadband internet connection in schools, and school accessibility using urban public transport. Geographical information is taken into account introducing a spatially structured latent effect in the regression model, defined at a higher aggregation level than municipalities, either provinces or catchment areas of infrastructural poles. Based on the prior belief that, besides the effect of explanatory variables, Invalsi scores tend to be closer in value across nearby areas than among ones far apart [11], we assume an Intrinsic Conditional Autoregressive structure [hereinafter ICAR, 13]. Since the scores in two subjects are available, a bivariate ICAR [61] spatial effect is modelled.

To ensure that covariates and spatial effects do not compete in explaining the target variable, we employ the variant of the Spatial+ approach proposed by [81] allowing to overcome the need to define a spatial model on covariates by leveraging on the spectral properties of the neighbouring structure of the data.

The analysis proposed here follows a Bayesian paradigm and the main object of inference are therefore the marginal posterior distributions of both covariates and latent spatial effects. Due to the complexity of deriving the posterior marginals of interest, we resort to INLA INLA, VBINLA, see also Chapter 1. In particular, multivariate spatial modelling of areal data is implemented in the `INLAMSM` package [69, 68], available on GitHub.

The remainder of this chapter is structured as follows. In Section **??** the data employed and the spatial structure referred to are described. In Section 4.3 we outline the regression model used and the method followed to deal with spatial confounding. In Section **??** we summarise the application of the INLA and compare some possible model formulations. In Section 4.4 we discuss the results of the models implemented.

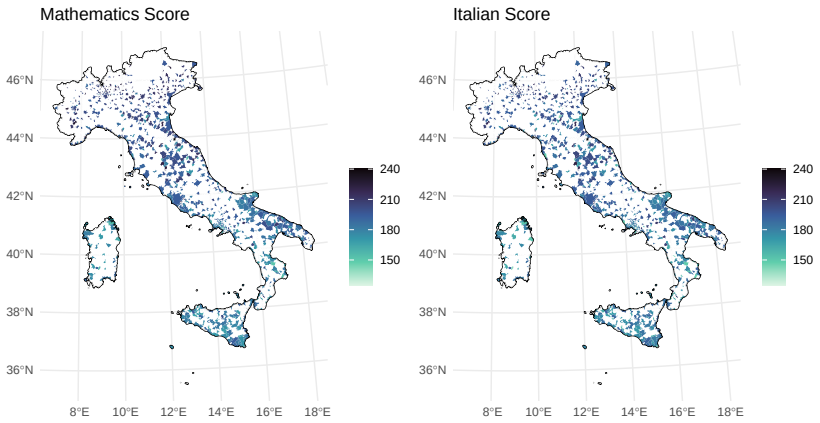

Figure 4.1: Invalsi scores in 2022/2023, 2nd year of high school. Trentino-Alto Adige is absent due to the lack of availability of auxiliary variables.

4.2 Student outcome data and infrastructural endowement

In this Section, the data on high school student outcomes together with some auxiliary data on the state of the infrastructure of Italian municipalities are described. Notice that Italian municipalities for which all relevant data are available amount to 873 over 7904 for the school year 2022/2023, the last school year for which all data are available on October 29, 2024. Data are obtained through the `SchoolDataIT` package, described in Chapter 3.6.

4.2.1 Invalsi scores

Municipality-level Invalsi scores in Mathematics and Italian at the 2nd year of high school in the school year 2022/2023 are displayed in Figure 4.1. The sparse structure of data is mostly due to the concentration of high schools across a limited number of municipalities, and in smaller part to the selection criteria of the Invalsi institute. Nevertheless, the N-S territorial pattern appears clear, especially for Mathematics scores.

4.2.2 Auxiliary variables

Auxiliary information considered herein has been selected to synthesize the general infrastructural state of municipalities and the accessibility to schools. To the state of our findings, the most informative variables are the following:

Urban public transport The municipality-level percentage of high schools located within 250 meters from a public urban transport hub, as reported in the School Buildings Section of the Unique School Data Portal [51]. Data are available for each public School building in Italy, except for the Trentino - Alto Adige region.

Ultra-Broadband activation status The municipality-level percentage of high schools where ultra-broadband connection had been implemented before September 1st, 2022. Ultra-broadband is defined as an internet connection with a maximum speed of 1 gigabit per second and a minimum guaranteed speed of 100 megabits/second until the peering, and open data regarding the implementation status are provided by [44]. Since the implementation plan does not regard all schools in Italy, the activation status is imputed to zero (not implemented) for all schools not listed in the Plan.

Inner Areas The taxonomy of inner areas is published by the Italian National Institute of Statistics [46] and includes six classes, defined as in Chapter 3.6. Municipalities in classes A and B, namely the ones serving as destination poles, are labelled as central. Municipalities in classes C-D and E-F are labelled as intermediate and peripheral respectively. Dummy variables "Central" and "Peripheral" are defined according to this distinction.

4.2.3 Spatial structure

Considering 873 municipalities over 7904 leads to a sparse pattern of observational units. For the forthcoming analysis, we opt to to define a less sparse spatial structure at the higher spatial aggregation level of macro-areas (see below). Say the total observational length is $N = 873$ municipalities, the number of macro-areas is n, and the number of municipalities within the i-th macro-area is N_i, then $N = \sum_{i=1}^{n} N_i$.

Two alternative definitions of the macro-areas are used, corresponding to two spatial models. The first level are provinces (NUTS-3 units), amounting to $n = 105$ macro-areas. Alternatively, infrastructural catchment areas are considered, defined as the ensemble of an infrastructural pole and all the intermediate and peripheral municipalities for which that pole is the destination pole. Infrastructural catchment areas amount to $n = 206$ units.

Macro-areas can be treated as the nodes of a graph $\mathcal{G}$ with $G = 3$ connected components, namely the continent and the islands of Sicily and Sardinia. The neighbourhood structure of $\mathcal{G}$ is described by the binary proximity matrix $\mathbf{W}$.

4.2.4 Spatial exploratory analysis of explanatory variables

In this Section, the spatial structure of explanatory variables at the province level is briefly explored. In Figure 4.2 auxiliary variables are mapped from municipalities to provinces by unweighted averages, i.e. the proportion of central and peripheral municipalities per province, and the unweighted averages of

municipality-level proportions of schools served by ultra-broadband and urban transport are computed. Large-scale spatial variation is particularly evident in the first two variables, showing a higher concentration of infrastructural poles in the north and, vice-versa, a higher concentration of peripheral municipalities in the South. Mainly in the North, we also notice that some provinces have no peripheral municipalities at all, i.e. all non-central municipalities have a road travel time shorter than 41 minutes [46] from the closest pole.

Concerning the ultra-broadband activation status, it is possible to observe a slight disadvantage in the mountainous inland regions and a strong disadvantage in the Sardinia region. The availability of public transport hubs shows a weak advantage for Central and Northwestern Italy. In Table 4.1 the Moran's I values is computed across provinces for the covariates. The standardised index I_{std} is obtained assuming the values $-1/104$ and 0.00459 for the mean and the variance under the null hypothesis of no spatial autocorrelation [21]. For the first three variables the values of I_{std}, suggesting a strong spatial autocorrelation, while the evidence of autocorrelation is weaker for the percentage of schools served by urban public transport. The values of I and I_{std} are computed with the `spdep` `R` package [14].

Variable	I	I_{std}
Central	0.2705	4.1338
Peripheral	0.4236	6.3447
Broadband avail.	0.1908	2.9234
Urban transport	0.0662	1.1072

Table 4.1: Moran's I and standardised I values for province-level averages of auxiliary variables.

When averaging covariates across infrastructural catchment areas, the values of I_{std} are generally higher, and we find evidence for spatial autocorrelation also for the proportion of schools served by urban public transport.

Variable	I	I_{std}
Central	0.7100	15.4990
Peripheral	0.2085	4.6337
Broadband avail.	0.3089	6.8224
Urban transport	0.3051	6.7238

Table 4.2: Moran's I values for infrastructural catchment area-level averages of auxiliary variables

4.3 A bivariate spatial model for student scores

In this Section, the general features of the bivariate spatial model for the Invalsi scores in Mathematics and Italian are outlined and general notation is intro-

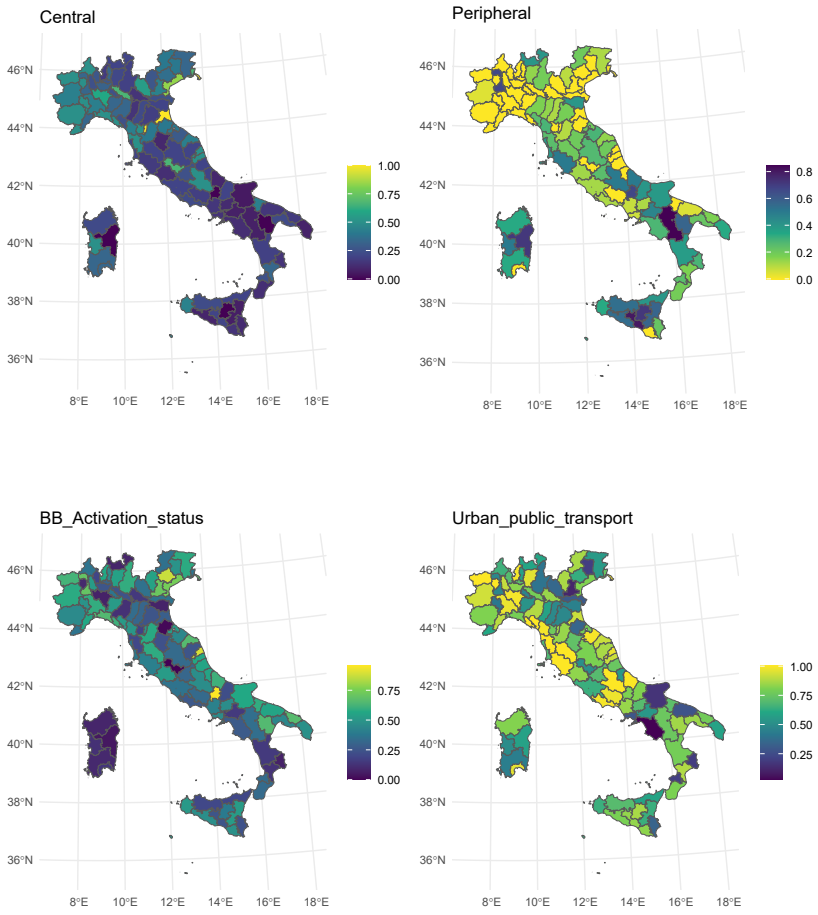

Figure 4.2: Upper panel: proportion of central (left) and peripheral (right) municipalities per province. Lower panel: ultra-broadband availability (left) and urban transport accessibility (right) per province.

duced. Assume the Invalsi scores in Mathematics and Italian $y = (y_1^\top, y_2^\top)^\top$ are defined as a vector of length $2N$ and modelled as follows:

$$y = \tilde{\mathbf{X}}\,\beta + \tilde{\xi}\,\tilde{\mathbf{C}}\,\beta_C + \tilde{\xi}\,z + \varepsilon \tag{4.1}$$

where $\tilde{\mathbf{X}} := I_2 \otimes \mathbf{X}$, $\tilde{\xi} := I_2 \otimes \xi$ and $\tilde{\mathbf{C}} := I_2 \otimes \mathbf{C}$. $\mathbf{X}$ is the $N \times 4$ matrix of auxiliary variables (see Section 4.2.2), and β is a vector of fixed effects of length 8. The $N \times n$ matrix ξ is binary and maps the n macro-areas onto N municipalities. The $n \times 3$ matrix $\mathbf{C}$ is also binary and denotes which connected component (continent, Sicilia, Sardinia) each macro-area belongs to, β_C is the vector of component-specific intercepts of length 6. The bivariate latent spatial field $z = (z_1^\top, z_2^\top)^\top$ is defined at the macro-area level and accounts for both the spatial variation and the correlation between Mathematics and Italian scores. Finally $\varepsilon = (\varepsilon_1^\top, \varepsilon_2^\top)^\top$ is the matrix-valued random error, following the distribution:

$$\begin{cases} \varepsilon_1 \mid \omega_1 \overset{\text{iid}}{\sim} N(\mathbf{0}, \omega_1) \\ \varepsilon_2 \mid \omega_2, \alpha \overset{\text{iid}}{\sim} SN(m_{\omega_2,\alpha}, s_{\omega_2,\alpha}, \alpha) \end{cases} \tag{4.2}$$

$SN(\cdot)$ in 4.2 denotes the Skew-Normal distribution [5] with location and scale parameters m_{α,ω_2} and s_{α,ω_2} defined to ensure that $\mathbb{E}[\varepsilon_2|\alpha] = 0$ and $VAR[\varepsilon_2|\alpha] = \omega_2$, and α is the shape parameter. This choice is due to the negative skewness in municipality-level Italian scores that neither auxiliary variables or spatial effects can explain (see Fig. 4.7 below). For interpretation reasons, in the remainder of this paper, we consider a transformation of α, namely the skewness parameter γ_1, that has the property of lying approximately in the interval $]-1, 1[$:

$$\gamma_1 := \frac{4-\pi}{2}\left(\frac{2\alpha^2}{\pi(1+\alpha^2)}\right)^{\frac{3}{2}}\left(1 - \frac{2\alpha^2}{\pi(1+\alpha^2)}\right)^{-\frac{3}{2}}$$

Following [85], α is assigned a Penalised Complexity prior [77] with a given rate parameter λ. $\lambda = 4$ is chosen based on empirical considerations, i.e. balancing model complexity and fit. However, the posterior distribution of the skewness parameter does not appear to be sensitive to the choice of λ [85].

Covariate effects β have $N(0, 10^3)$ non-informative priors, while priors for intercepts in β_C are set as $N(180, 10^3)$, according to the expected global mean of Invalsi ability scores (Section 4.2.1). Precision parameters for the error terms, namely ω_1 and ω_2, have independent Gamma vague priors with shape 10^{-3} and rate 10^{-3}.

4.3.1 Modelling the spatial component

Considering the neighbourhood structure outlined in Section 4.2.3, z is modelled as a bivariate ICAR defined on the graph $\mathcal{G}$, following this conditional prior

distribution at the spatial unit level for each i-th node, with $i \in [1, n]$:

$$z_i | z_{-i}, \Lambda \sim N\left(\sum_{j \sim i} \frac{w_{ij}}{d_i} z_j,\ \frac{1}{d_i}\Lambda^{-1}\right) \tag{4.3}$$

where $d_i := \sum_{s=1}^{n} w_{is}$ is the number of neighbours of node i and Λ is the precision parameter. This representation is a special case of the model developed by [61, theorem 2.1, corollary 2] and implies a joint Normal prior on z with zero mean and precision $\Lambda \otimes \mathbf{R}$,

$$\pi\left(z \mid \Lambda\right) = \left(\frac{1}{2\pi}\right)^{n-3} \sqrt{\mid \Lambda \otimes \mathbf{R} \mid_{+}}\ e^{-\frac{1}{2}\mathrm{vec}(z)'(\Lambda \otimes \mathbf{R})\mathrm{vec}(z)} \tag{4.4}$$

where $\mathbf{R} := \mathbf{D} - \mathbf{W}$ is the Laplacian matrix of the graph $\mathcal{G}$ with 3 connected components and $\mathbf{D} = \mathrm{diag}(d_1, d_2 \dots d_n)$ is the degree matrix of $\mathcal{G}$. Since $\mathbf{R}$ is singular with rank deficiency 3 and $\pi(z|\Lambda)$ is therefore improper [41], it is necessary to constrain z to sum to zero within each connected graph component [12]. This is the reason for adopting component-specific intercepts β_C in equation 4.1.

To ease the interpretation of Λ as the precision parameter of z, it is possible to cleanse it from the effect of the neighbourhood structure by defining a scaled version of $\mathbf{R}$ [79] and reparametrising the precision of z accordingly. Since $\mathcal{G}$ is disconnected, each component-specific block of $\mathbf{R}$ is multiplied by the relevant typical variance, namely the geometric mean of the diagonal of the corresponding block of its pseudoinverse, following the methodology proposed by [28]. It is therefore possible to define a precision parameter Λ_{scaled} which is not confounded with graph-induced effects. The scaled precision is assigned a Wishart prior [34] with $2k+1$ degrees of freedom and scale parameter equal to the identity matrix, i.e. $\Lambda_{\text{scaled}} \sim \text{Wishart}_k(I_k, 2k+1)$, with $k = 2$ [69].

$$\mathbf{R}_{\text{scaled}} := \mathcal{P}\left(\bar{\sigma}^2_{C_1}\mathbf{R}_{\mathbf{C_1}} \oplus \bar{\sigma}^2_{C_2}\mathbf{R}_{\mathbf{C_2}} \oplus \bar{\sigma}^2_{C_3}\mathbf{R}_{\mathbf{C_3}}\right)\mathcal{P}'$$

where $\mathcal{P}$ is an appropriate permutation matrix, $\mathbf{R}_{\mathbf{C_i}}$ is the Laplacian of the i-th connected component of the graph and $\bar{\sigma}^2_{C_i}$ is the relevant typical variance.

In `R-INLA`, precision scaling is implemented automatically for intrinsic models, like the univariate ICAR, through the option `scale.model` within the `inla()` function call; otherwise the scaled structure matrix for each component can be computed as a standalone object with `inla.scale.model()`. In the multivariate case, `INLAMSM` provides readily-defined models for which the user is required to provide the neighbourhood matrix $\mathbf{W}$ instead of the Laplacian (as different models with the same neighbourhood matrix have different structure matrices). Hence, to scale a multivariate ICAR model we derive $\mathbf{W}$ from the scaled Laplacian.

4.3.2 Spatial confounding

In our multilevel framework, the value of the m-th covariate $\mathbf{X_{\cdot m}}$ observed in municipality h belonging to macro-area i can be decomposed as

$$x_{ih;m} = \bar{x}_{i;m} + \Delta x_{ih;m}$$

being $\bar{x}_{i;m}$ the unweighted average value of the covariate within the i-th macro-area; the term $\Delta x_{ih;m}$ represents municipality-level noise. In matrix form, this decomposition is: $\mathbf{X} = \xi\bar{\mathbf{X}} + \boldsymbol{\Delta}\mathbf{X}$, where $\bar{\mathbf{X}} = (\xi^\top\xi)^{-1}\xi^\top\mathbf{X}$.

Consider the eigendecomposition of the Laplacian matrix [81]:

$$\mathbf{R} = \mathbf{VLV}^\top$$

where the eigenvalues in $\mathbf{L}$ are in decreasing order and the eigenvectors in $\mathbf{V}$ have a decreasing number of oscillations.

Within a generic component of the connected graph, the eigenvector associated with the lowest non-null eigenvalue follows a linear spatial trend (i.e. the Fiedler vector of the relevant subgraph), the one related to the second non-null eigenvalue follows a quadratic trend (one oscillation), and so on. For an appropriately chosen $n \times 4$ matrix $\mathbf{b}$, $\bar{\mathbf{X}}$ can be expressed as a linear combination of $\mathbf{V}$:

$$\bar{\mathbf{X}} = \mathbf{Vb}$$

Intuitively, the spatial component of $\bar{\mathbf{X}}$ is determined by the last columns of $\mathbf{V}$ [81]: without loss of generality, $\bar{\mathbf{X}}$ is decomposed into:

$$\bar{\mathbf{X}} = \bar{\mathbf{X}}^{(\mathbf{NS})} + \bar{\mathbf{X}}^{(\mathbf{S})} + \bar{\mathbf{X}}^{(\mathbf{0})}$$

where $\bar{\mathbf{X}}^{(\mathbf{NS})}$ is the nonspatial component, given by the linear combination of the first $n - G - K$ eigenvectors (with $G = 3$ connected graph components), $\bar{\mathbf{X}}^{(\mathbf{s})}$ is the combination of the eigenvectors associated with the last K nonzero eigenvalues and represents the spatial component, and $\bar{\mathbf{X}}^{(\mathbf{0})}$ is the combination of the $G = 3$ eigenvectors in the null space of the Laplacian matrix, constant within each connected component. To remove spatial confounding, it is sufficient to consider $\xi\left(\bar{\mathbf{X}}^{(\mathbf{NS})}\right) + \boldsymbol{\Delta}\mathbf{X}$ as the covariate matrix in the regression model.

In Figure 4.3 the eigenvectors corresponding to the last two non-zero eigenvalues of $\mathbf{R}$ at the macro-area level of provinces for the continent graph component are plotted. It is possible to see that the second-last eigenvector follows a quadratic trend with one oscillation, whereas the last eigenvector follows a linear North-South trend.

Identification of the spatial variation in the covariates

The Spatial+ procedure requires the choice of the number K of eigenvectors to define $\bar{\mathbf{X}}^{(\mathbf{NS})}$. A documented approach [81, 58] is to choose the value of K minimizing the Watanabe-Akaike Information Criterion [WAIC, 87]. Based on this method, our first strategy is searching for the optimal number of eigenvectors to be removed for each explanatory variable, subject to the following constraints.

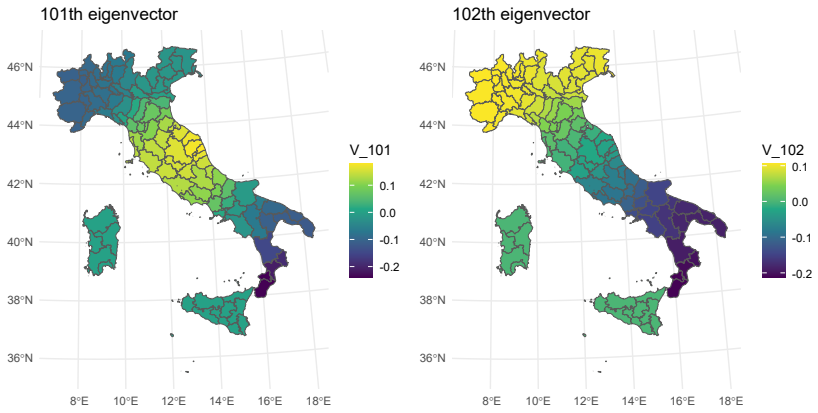

Figure 4.3: Lowest frequency eigenvectors, provinces

When z is defined at the province level (91 areas in the continent, 9 in Sicily, 5 in Sardinia), we remove a maximum of 9 eigenvectors in the continent, and 1 for each of the two islands, whereas when z is defined at the level of infrastructural catchment areas (186 macro-areas in the continent, 13 in Sicily, 7 in Sardinia) we remove up to 18 eigenvectors from the continent, 2 from Sicily and 1 from Sardinia.

As an alternative strategy, we removed the smallest number of eigenvectors for $\bar{\mathbf{X}}^{(\mathbf{NS})}$ not to display evidence of autocorrelation according to Moran's I index. For the first three variables at the province level, removing the last 4 eigenvectors leads to small standardized I values (0.79697, 1.3854, 1.2397), suggesting that spatial structure in these variables is driven by a linear trend over the continent, as seen in Figure 4.2. For infrastructural catchment areas, doing the same thing with the proportion of central and peripheral municipalities would lead to standardized I values of 1.2522 and 0.4598 respectively; again, this can be interpreted as the presence of a linear trend. For the last two covariates, instead, it is necessary to remove a higher amount of eigenvectors, which suggests the presence of spatial variation on a smaller scale before accepting the hypothesis of no spatial autocorrelation.

Details on the removal patterns are shown. The continent has 91 provinces and 186 infrastructural catchment areas; Sicily has 9 provinces and 13 infrastructural catchment areas; Sardinia has 5 provinces and 7 infrastructural catchment areas. The last eigenvector is constant within each component. Patterns S+(1) and S+(3) are the simplest ones allowing to shrink the value of the standardised Moran's index below the 95-th percentile of the Standard Normal distribution. At the province level, it is sufficient to remove the last 4 eigenvectors; at the level of infrastructural catchment areas this is only sufficient for the first two

covariates, while more eigenvectors, corresponding to higher order trends, need to be removed for what concerns the other two variables. Pattern S+(2) allows, to the best of our findings, for the best ICAR fitting at the province level; S+(4) does the same at the level of infrastructural catchment areas; S+(5) and S+(6) serve the same purposes but for the PCAR model. In Figure 4.4 the first two columns of $\bar{\mathbf{X}}^{(\mathbf{NS})}$ are shown, i.e. the province-level proportion of central and peripheral municipalities, using the S+(2) correction. Original values of these variables are in the upper panel of Figure 4.2.

Pattern	level	Component	Central	Peripheral	BB Activation	Urban transport
		Continent	2	2	2	0
S+(1)	Prov	Sicily	1	1	1	0
		Sardinia	1	1	1	0
		Continent	5	4	5	0
S+(2)	Prov	Sicily	1	1	1	0
		Sardinia	1	1	1	0
		Continent	2	2	10*	13
S+(3)	Pole	Sicily	1	1	1	1
		Sardinia	1	1	1	1
		Continent	8	8	9	10
S+(4)	Pole	Sicily	2	1	1	1
		Sardinia	1	1	1	1
		Continent	5	4	6	0
S+(5)	Prov	Sicily	1	1	1	0
		Sardinia	1	1	1	0
		Continent	9	8	7	13
S+(6)	Pole	Sicily	2	1	1	2
		Sardinia	1	1	1	1

Table 4.3: Eigenvectors removal patterns for each explanatory variable.* In this case, eigenvectors removed are the 172th and 178-186th ones

4.3.3 Model assessment

In this Section, some alternative model formulations are compared using a set of selection criteria internally computed by `R-INLA`: the negative Log Pseudo Marginal Likelihood [LPML, 31, 32], i.e. minus the logarithmic sum of the Conditional Predictive Ordinates [70, 40], and the Watanabe-Akaike Information criterion [WAIC, 87], following the formulation of [35], the Deviance Information Criterion [DIC, 78], alongside with the Mean Squared Error of posterior predictive response averages.

The Conditional Predictive Ordinate is a leave-one-out cross-validatory diagnostic given for a generic j-th observation from the h-th variable, with $j \in [1, N]$ and $h \in [1, 2]$, by:

$$\text{CPO}_{j,h} := \pi(y_{j,h} \mid y_{-(j,h)})$$

Low CPO values denote "surprising" observations and, hence, possible outliers. Details on how the CPO is computed in `R-INLA` are provided by [40]. Here, we

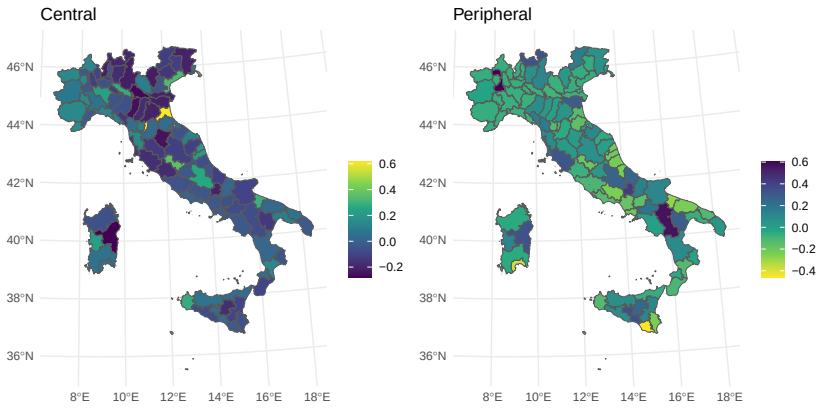

Figure 4.4: Proportion of central (left) and peripheral (right) municipalities per province once the spatial structure is removed by applying the S+(2) correction (see Table 4.3).

compare models through the logarithmic sum of CPOs changed of sign [31].

The Deviance Information Criterion (DIC, [78]) is given by:

$$\text{DIC} := P_D + \mathbb{E}[D(\vartheta, \Psi) \mid y]$$

Where

$$P_D = \mathbb{E}[D(\vartheta, \Psi)] - D(\hat{\vartheta}, \hat{\Psi})$$

denotes the effective number of parameters or unconstrained parameters and measures model complexity, $D(\vartheta, \Psi) = -2\ln\pi(y \mid \vartheta, \Psi)$ being the model deviance, whose expectation measures the goodness of fit. Following [74] we define $\hat{\vartheta}$ as $\mathbb{E}[\vartheta|y]$ and $\hat{\Psi}$ as the posterior mode of Ψ, this latter choice being due to the skewness in $\pi(\Psi|y)$. We show both the effective number of parameters and the expected deviance alongside the DIC.

The Watanabe-Akaike Information Criterion [87], also known as Widely Applicable Information Criterion, is also useful in balancing model complexity and goodness of fit, the former being measured as $\text{Var}[\ln\pi(y|\vartheta, \Psi)]$. An interesting feature of the WAIC is given by its averaging the likelihood over ϑ and Φ rather than using point estimates ([35]). We use the formulation provided by [35].

$$WAIC = 2\sum_{j=1}^{N} VAR[\ln\pi(y_j|\vartheta, \Psi)] - 2\sum_{j=1}^{N} \ln\mathbb{E}[\pi(y_j|\vartheta, \Psi)]$$

Models defined with ICAR random effects are compared in Table 4.4. "Base" and "RSR" denote the model with no correction for spatial confounding and

the RSR model respectively. S+(1) and S+(2) are Spatial+2.0 models with province-level latent effects with two different eigenvector removal patterns: the former is the most conservative one for which no evidence for autocorrelation in the covariates is found (see Section 4.2.4), the latter is the one with smallest WAIC. S+(3) and S+(4) are the Spatial+2.0 models developed with the same strategy but with latent effects defined at the level of infrastructural catchment areas. Detailed eigenvector removal patterns are shown in Appendix **??**.

Removing spatial autocorrelation from covariates based on Moran's test allows for some barely noticeable improvements in inference. Using finer support for the latent effects improves the fitting but this gain is outweighed by increased complexity (overall, the DIC increases).The model with province-level spatial effects is overall preferable based on all three metrics WAIC, DIC and LPML. RSR appears to perform poorly in both cases if compared to the base model. Furthermore, posterior means of β obtained by RSR result quite close to those of the nonspatial model, while credible intervals are narrower, which is consistent with the lesser coverage of Type-S errors in RSR models [57].

In Appendix .1 the results of a broader set of models are shown, including models with no random effects, with unstructured random effects and with two independent ICAR random effects; a focus on the estimates of β under the nonspatial model and under RSR is in Appendix .1.1.

Lastly, in Appendix **??** the results of a different model, the proper CAR [33], are summarised. The core feature of this formulation is introducing an additional parameter to account for the strength of spatial association.

z level	Model	-LPML	WAIC	DIC	MSE
Prov	Base	6689.917	13379.149	13379.808	235.551
Prov	RSR	6764.094	13526.761	13526.476	251.886
Prov	S+(1)	6689.672	13378.651	13379.311	235.326
Prov	S+(2)	**6689.500**	**13378.303**	**13378.909**	235.189
Pole	Base	6694.435	13387.776	13388.726	232.320
Pole	RSR	6754.037	13505.502	13507.315	239.313
Pole	S+(3)	6694.443	13387.747	13388.711	**231.811**
Pole	S+(4)	6694.055	13387.018	13387.948	231.838

Table 4.4: Model diagnostics for 8 ICAR model formulations: spatial aggregation level of z, spatial confounding treatment, negative Log Pseudo Marginal Likelihood, Watanabe-Akaike Information criterion, Deviance Information Criterion, Mean Squared Error of posterior predictive response averages.

4.4 Results

In Table 4.5 the estimated effects of covariates for the province-level ICAR are resumed, under both the base formulation and the S+(2) modification. Boundaries of credible intervals correspond to the 5-th and 95-th percentiles. Covariates in the deconfounded model have been scaled to keep the same variance as in the base model.

		Base model				S+(2)			
	Subj	mean	sd	LB	UB	mean	sd	LB	UB
Continent	MAT	191.399	0.961	189.515	193.285	193.332	0.855	191.656	195.009
Continent	ITA	187.107	0.999	185.137	189.056	188.585	0.863	186.886	190.270
Sicily	MAT	177.764	1.496	174.829	180.698	178.386	1.369	175.702	181.070
Sicily	ITA	176.884	1.550	173.835	179.915	177.273	1.417	174.489	180.046
Sardinia	MAT	174.325	2.197	170.017	178.634	174.561	2.159	170.327	178.795
Sardinia	ITA	171.914	2.267	167.460	176.351	172.126	2.208	167.790	176.450
Central	MAT	2.706	0.910	0.922	4.490	2.527	0.890	0.781	4.273
Central	ITA	2.379	0.996	0.433	4.338	2.460	0.979	0.547	4.386
Peripheral	MAT	-2.200	1.005	-4.171	-0.228	-2.018	0.958	-3.897	-0.139
Peripheral	ITA	-1.845	1.049	-3.901	0.215	-1.793	1.000	-3.753	0.170
BB Activation	MAT	3.331	1.074	1.226	5.437	3.262	1.049	1.205	5.319
BB Activation	ITA	2.296	1.126	0.090	4.509	2.130	1.100	-0.024	4.291
Urban transport	MAT	2.466	1.043	0.420	4.513	2.501	1.044	0.453	4.549
Urban transport	ITA	2.841	1.060	0.765	4.924	2.838	1.060	0.762	4.919

Table 4.5: Posterior summaries of intercepts and covariate effects when z is defined as a province-level ICAR, under the base model and the Spatial+2.0 model (optimal combination of eigenvector removal under the WAIC metric)

Modelling Italian scores appears to be generally subject to higher uncertainty. For both subjects, differences between the continent and the islands are strong: almost 15 Invalsi points on average between the continent and Sicily, more than 15 points between the continent and Sardinia, with non-overlapping credible intervals. Central municipalities have an expected advantage of 2.706 points over intermediate municipalities in Mathematics test, while this expectation slightly falls to 2.527 points once the share of infrastructural poles in each province is corrected with S+(2). The relative effect of central municipalities on Italian scores is comparable and slightly lower. The difference between intermediate and peripheral municipalities is lower in expected value and not even significant for Italian scores. A municipality in which all schools are provided with ultra-broadband connection has an expected advantage of more than 3 Invalsi points in the Mathematics score over one in which the connection is completely lacking, the effect being weaker on Italian scores (and possibly not significant once the spatial structure is removed from the covariate). Lastly, the availability of urban public transport is associated with a significant advantage in Invalsi scores, since a municipality where all schools are reachable has an expected advantage of almost 2.5 points in Mathematics scores and about 2.8 points in Italian scores.

In Figure 4.5 the expected spatial effect $\mathbb{E}[z|y]$ under the S+(2) model is plotted. Territorial gaps in Invalsi scores are severe, as one can argue from

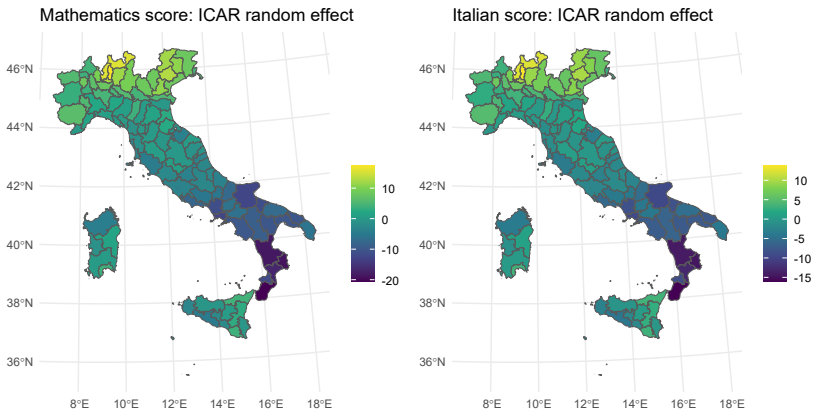

Figure 4.5: Expected values of z modelled as a province-level ICAR and applying S+(2) correction

the range of $\mathbb{E}[z|y]$. Focusing on the continent, the Calabria region appears particularly vulnerable, while Lombardia turns out to be the most advantaged region.

In Figure 4.6 the predicted values of y are shown, using the same model. The model captures the spatial trend, but still leaves a high municipality-level noise unexplained, as the high error variances suggest (ω_1 and ω_2 in Table 4.6).

The highest scores are estimated in the municipalities of Lecco (Lombardia), with expected scores of 216.072 points in Mathematics (observed score of 214.042 points) and 207.637 points in Italian (observed 204.990 points) and Merate (province of Lecco), with expected scores of 216.133 points in Mathematics (observed score 229.352 points) and 207.274 points in Italian (observed value 216.045 points).

Lowest scores in Mathematics are estimated in the municipalities of La Maddalena (province of Sassari, Sardinia) at 169.351 points and Oppido Mamertina (province of Reggio Calabria) at 170.301 points (observed 165.814 and 170.105 points respectively). Lowest scores in Italian are estimated in the municipalities of La Maddalena at 170.016 points and Bosa (province of Oristano, Sardinia) at 169.787 points (observed 168.071 and 168.890 points respectively).

Finally, in Table 4.6 the posterior summaries for hyperparameters Ψ are displayed. Please notice that the precision of z has been scaled, hence variances σ_1^2 and σ_2^2 do not depend on the graph-induced effect. The variance of spatial effects is higher in Mathematics scores (σ_1^2), while Italian scores have a higher amount of unexplained noise (ω_2). Correlation between the two scores is taken into account through the correlation between the two ICAR fields ρ, which turns out to be high, consistent with Figure 4.5. Lastly, the choice of modelling Italian

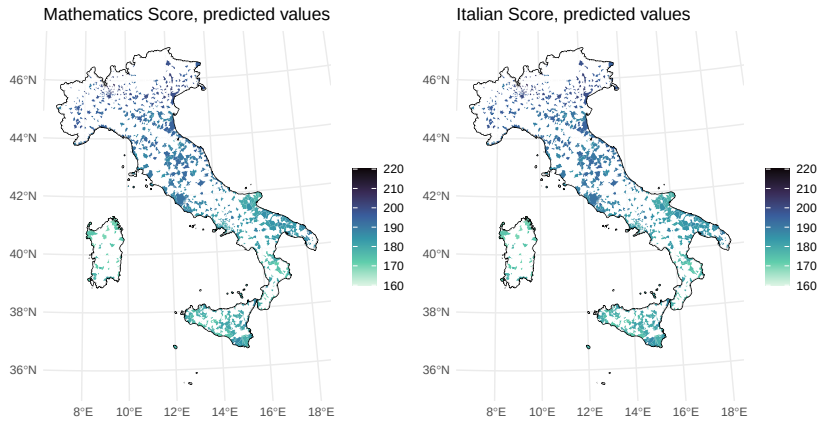

Figure 4.6: Predicted values of Invalsi scores using as a province-level ICAR latent effect and applying the S+(2) correction

scores as a Skew-Normal variable is corroborated by the posterior distribution of γ_1, whose credible interval ranges far from zero. Kernel density estimation of residuals in Italian scores is shown in Figure 4.7. Negative skewness is easily noticeable. Density is estimated by the Gaussian kernel, using the Silverman's thumb rule to define the bandwidth [76, Section 3.4.2]

		Base model				S+(2)			
	Subj	LB	Median	UB	sd	LB	Median	UB	sd
σ_1^2	MAT	16.621	27.414	46.208	7.578	17.288	28.722	46.823	7.562
σ_2^2	ITA	10.277	18.156	33.042	5.843	10.672	18.832	32.760	5.663
ρ		0.894	0.975	0.995	0.027	0.884	0.976	0.994	0.030
ω_1	MAT	100.830	110.924	122.138	5.426	100.712	110.790	121.982	5.417
ω_2	ITA	119.507	132.104	145.888	6.718	119.368	131.795	145.646	6.692
γ_1	ITA	-0.493	-0.371	-0.232	0.067	-0.493	-0.369	-0.232	0.066

Table 4.6: Posterior summaries of hyperparameters when z is defined as a province-level ICAR, under the base model and the Spatial+2.0 model (optimal combination of eigenvector removal under the WAIC metric)

4.5 Concluding remarks

While a good body of literature studies student ability scores at the individual level, we have turned the analysis to a different framework, attempting to explain the geographical distribution of Invalsi scores based on infrastructural

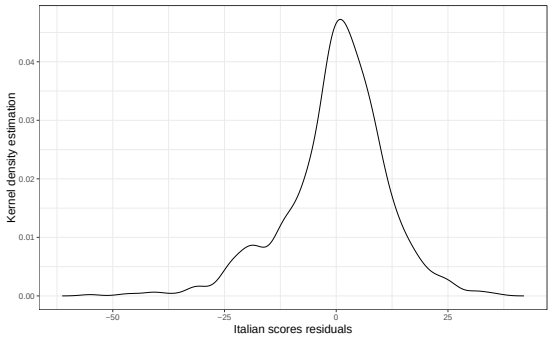

Figure 4.7: Kernel density estimation of the residuals of Italian scores under the S+(2) model

variables and using a multivariate spatial regression model. For a better understanding of the spatial variation, the precision of the spatial effect was scaled to account for discontinuities implied by the presence of two islands. The typical skew distribution in the Italian Invalsi scores was accounted for by relaxing the Normality assumption and fitting a Skew-Normal likelihood, with evidence for negative skewness. Lastly, to avoid explanatory variables being confounded with the spatial effect, we cleansed them from low-frequency spatial trends in estimating their effect on Invalsi scores.

The infrastructural state of schools and municipalities results in a significant impact on student performances, also when their effect is separated from spatial information. The classification of Italian municipalities into central, intermediate, and peripheral allows to report a noticeable advantage of the first to the second, while the difference between intermediate and peripheral municipalities is weaker. Our results do highlight the overall vulnerability of inner areas under the educational dimension.

Our analysis of Invalsi scores includes a spatially structured latent field defined at the level of macro-areas, either provinces or infrastructural catchment areas. We find that the between-macro-areas spatial effect is indeed a strong driver of Invalsi scores, in addition to between-municipalities factors. This confirms the strength of the territorial divides shaping many aspects of Italian society, especially the North-South gap.

Still keeping our focus on infrastructural explanatory variables, this analysis can undergo some possible developments: first and foremost it can be extended to different school grades and different years. The choice of statistical models, both in terms of likelihood and prior assumptions, could be extended as well, including more tailored models. The PCAR model, for instance, may represent a worthwhile extension, though questions like improving the interpretation of precision parameters remain open, further research being needful to this aim.

Overall, the mapping of infrastructure access and social vulnerability requires adequate statistical computational methods, and `R-INLA` appears a very

flexible tool to this aim.

.1 Extensive model comparison

In Table 7 all the models run throughout this analysis are compared. Two alternative strategies to approximate the full posterior of ϑ have been tested, namely the Variational Bayes and the Simplified Laplace approximations (respectively VB and SL, hereinafter). The former is implemented in the latest `R-INLA` framework and has been used to estimate the models whose results are summarised in Sections 4.3.3 and 4.4. The latter allows to preserve information about skewness in the full conditional and is implemented in an older software framework. Due to the difficulties in locating the mode of Ψ, y required to be centered at zero in the models approximated with the SL method.

Some additional model formulations are also compared: "NULL" denotes the model with no spatial effect (component-specific intercepts are still used); "IID" denotes the model with IID macroarea-level effects; "Ind. ICAR" denotes the model with two independent ICAR priors. The random intercept and the ICAR effect in the IID and independent ICAR models have an improper Uniform prior on the standard deviation. Models S+(1), S+(2), S+(3), and S+(4) are defined as in Appendix **??**. Alongside the selection criteria mentioned in Section 4.3.3, the two components of the DIC are shown, namely the expected deviance (Exp. Dev) and the effective number of parameters (P_D). The computational time of each model is shown as well, expressed in seconds.

This comparison highlights that the ICAR is more adequate to study Invalsi scores than both the null and IID models. In this latter case, point estimates are accurate (low MSE), but the high complexity suggests this may be due to overfitting. The correlated ICAR outperforms the independent one. The VB approximation also appears to be generally preferable over the SL.

The value of Conditional Predictive Ordinates (CPOs) computed by `R-INLA` is reliable only under some regularity conditions [40], which are always met except for the model with IID latent effects defined for infrastructural catchment areas and computed using the SL approximation; in this case, the observation for which the CPO computation is not reliable corresponds to the municipality of Melzo (MI), having the record highest Italian score (230 points). The CPOs for that model have then been recompiled (function `INLA::inla.cpo()`). Though the results are quite similar (values of all selection criteria are slightly lower under the VB approximation), the SL approximation required to center y at zero, other than being less computationally efficient. This encourages to employ the latest `R-INLA` version supporting the VB approximation even if we have a skewed likelihood for one of the two target variables.

z level	Approx	Model	-LPML	WAIC	DIC	Exp. Dev.	P_D	MSE	time
Null	VB	Null	6979.502	13959.022	13959.264	13942.292	**16.972**	346.158	2.038
Prov	VB	IID	6765.015	13524.273	13523.594	13368.734	154.860	233.906	6.204
Prov	VB	Ind. ICAR	6722.306	13443.342	13443.291	13355.484	87.807	239.622	7.389
Prov	VB	Base ICAR	6689.917	13379.149	13379.808	13307.056	72.752	235.551	11.501
Prov	VB	ICAR RSR	6764.094	13526.761	13526.476	13440.811	85.665	251.886	10.197
Prov	VB	ICAR S+(1)	6689.672	13378.651	13379.311	13306.071	73.239	235.326	11.111
Prov	VB	ICAR S+(2)	**6689.500**	**13378.303**	**13378.909**	13305.406	73.502	235.189	11.613
Pole	VB	IID	6840.664	13669.514	13666.691	13452.671	214.020	236.693	5.613
Pole	VB	Ind. ICAR	6755.812	13508.925	13508.602	13390.671	117.931	240.193	17.936
Pole	VB	Base ICAR	6694.435	13387.776	13388.726	13300.529	88.197	232.320	12.178
Pole	VB	ICAR RSR	6754.037	13505.502	13507.315	13388.609	118.706	239.313	14.593
Pole	VB	ICAR S+(3)	6694.443	13387.747	13388.711	**13298.501**	90.210	231.811	13.932
Pole	VB	ICAR S+(4)	6694.055	13387.018	13387.948	13298.904	89.044	231.838	13.276
Null	SL	Null	6979.612	13959.098	13959.450	13942.372	17.078	346.156	4.137
Prov	SL	IID	6776.608	13524.444	13525.161	13369.837	155.324	234.058	11.304
Prov	SL	Ind. ICAR	6725.692	13444.008	13444.536	13355.918	88.619	239.603	14.828
Prov	SL	Base ICAR	6691.699	13379.533	13380.625	13307.040	73.585	235.470	24.948
Prov	SL	ICAR RSR	6769.474	13527.510	13527.747	13441.562	86.184	251.937	40.036
Prov	SL	ICAR S+(1)	6691.473	13379.438	13380.316	13307.059	73.257	235.465	24.181
Prov	SL	ICAR S+(2)	6691.157	13379.047	13380.227	13306.657	73.570	235.362	33.475
Pole	SL	IID	6857.701[1]	13669.789	13669.272	13453.829	215.442	236.662	13.597
Pole	SL	Ind. ICAR	6764.696	13509.405	13509.590	13391.686	117.904	240.333	20.426
Pole	SL	Base ICAR	6696.355	13388.114	13389.168	13300.653	88.515	232.351	31.058
Pole	SL	ICAR RSR	6761.415	13506.393	13508.790	13389.828	118.962	239.459	54.796
Pole	SL	ICAR S+(3)	6695.756	13388.190	13389.662	13298.914	90.748	**231.808**	31.176
Pole	SL	ICAR S+(4)	6696.531	13387.640	13388.553	13299.852	88.701	232.006	33.554

Table 7: Model diagnostics for all the combinations of approximation approach to $\pi(\vartheta|y)$, aggregation level of z, and model employed for z, either null (z not included, nonspatial model), IID (random intercept), independent bivariate ICAR, or dependent bivariate ICAR under either the base formulation, RSR or Spatial+2.0. Models are compared through negative Log Posterior Marginal Likelihood, Watanabe-Akaike information criterion, Deviance Information Criterion, expected deviance, effective number of parameters, Mean Square Error of posterior predictive response and computational time in seconds.

.1.1 Estimates of β under the nonspatial model and under Restricted Regression

In Table 8 the estimated covariate effects under the model with no spatial effect (nonspatial) and under RSR is shown, with z defined across provinces. Both models are estimated with the VB correction to the Gaussian approximation ("new" `R-INLA` version).For what concerns the nonspatial model, explaining y without a spatial latent field has the effect of raising estimates of β with respect to the (unrestricted) spatial model.

		Nonspatial model				RSR model			
	Subj	mean	sd	LB	UB	mean	sd	LB	UB
Central	MAT	4.340	1.116	2.151	6.530	4.347	0.944	2.496	6.198
Central	ITA	3.656	1.109	1.489	5.839	3.513	0.988	1.583	5.458
Peripheral	MAT	-5.055	1.205	-7.418	-2.692	-5.035	1.019	-7.034	-3.036
Peripheral	ITA	-4.092	1.161	-6.368	-1.813	-4.090	1.028	-6.105	-2.072
BB Activation	MAT	4.605	1.283	2.088	7.121	4.622	1.085	2.493	6.750
BB Activation	ITA	3.564	1.225	1.168	5.974	3.346	1.102	1.190	5.512
Urban transport	MAT	4.281	1.160	2.005	6.556	4.310	0.981	2.385	6.235
Urban transport	ITA	3.744	1.126	1.539	5.957	4.017	0.991	2.075	5.964

Table 8: Posterior summaries of covariate effects under the nonspatial model and the RSR-ICAR model defined at the province level

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards Applied Mathematics Series 55, tenth printing edition, 1972.

[2] Tommaso Agasisti and Giorgio Vittadini. Regional economic disparities as determinants of student's achievement in italy. *Research in Applied Economics*, 4(2):33, 2012.

[3] J. D. Angrist, E. Battistin, and D. Vuri. In a small moment: Class size and moral hazard in the italian mezzogiorno. *American Economic Journal: Applied Economics*, 9(4):216–249, 2017.

[4] T. Appelhans, F. Detsch, C. Reudenbach, and S. Woellauer. mapview: Interactive viewing of spatial data in R, 2016. EGU General Assembly Conference Abstracts, R package version 2.11.2.

[5] Adelchi Azzalini and Antonella Capitanio. *The Skew-Normal and Related Families*, volume 3. Cambridge University Press, 2014.

[6] C. Bagnarol and S. Donno. Analisi spaziale degli apprendimenti scolastici (invalsi) in inglese: un confronto tra le regioni del nord d'italia. In *XIII ESPAnet conference*, 2020.

[7] S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2nd edition, 2014.

[8] G. Barbieri, C. Rossetti, and P. Sestito. Teacher motivation and student learning. *Politica economica*, 33(1):59–72, 2017.

[9] Peter Barrett, Alberto Treves, Tigran Shmis, and Diego Ambasz. The impact of school infrastructure on learning: A synthesis of the evidence. *World Bank Publications*, 2019.

[10] F. Bernardi and R. C. Keivabu. Poor air quality at school and educational inequality by family socioeconomic status in italy. *Research in Social Stratification and Mobility*, 91:100932, 2024.

[11] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

[12] Julian Besag and Charles Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 12 1995.

[13] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20, 1991.

[14] Roger Bivand, Micah Altman, Luc Anselin, Renato Assunção, Olaf Berke, Andrew Bernat, and Guillaume Blanchet. Package 'spdep': Spatial dependence: Weighting schemes, statistics, r package version, 2017. R package version 1.3-7.

[15] P. Blatchford, P. Bassett, and P. Brown. Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and instruction*, 21(6):715–730, 2011.

[16] Massimiliano Bratti, Daniele Checchi, and Antonio Filippin. Territorial differences in italian students' mathematical competencies: evidence from pisa 2003. *Giornale Degli Economisti e Annali Di Economia*, 66(3):299–333, 2007.

[17] C. Brühwiler and P. Blatchford. Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and instruction*, 21(1):95–108, 2011.

[18] Mauro Bucci, Luigi Gazzano, Elena Gennari, Adele Grompone, Giorgio Ivaldi, Giovanna Messina, and Giacomo Ziglio. Per chi suona la campan (ell) a? la dotazione di infrastrutture scolastiche in italia (for whom the bell tolls? the availability of school infrastructure in italy). *Politica economica*, pages 1–50, 2023.

[19] S. Cattari, S. Alfano, V. Manfredi, B. Borzi, M. Faravelli, F. Di Meo, Da Porto, A. Saler, A. Dall'Asta, L. Gioiella, M. Di Ludovico, C. Del Vecchio, C. Del Gaudio, G. Verderame, G. Gattesco, I. Boehm, E. Speranza, M. Dolce, Lagomarsino F., and A. Masi. National risk assessment of italian school buildings: The mars project experience. *International Journal of Disaster Risk Reduction*, page 104822, 2024.

[20] J. Cheng, B. Karambelkar, Y. Xie, H. Wickham, K. Russell, K. Johnson, and V. Agafonkin. Package 'leaflet', 2019. version 2.2.2.

[21] A. D. Cliff and J. K. Ord. *Spatial Processes: Models and Applications*. Pion, London, 1981.

[22] R. Crescenzi, M. Giua, and G.V. Sonzogno. Mind the covid-19 crisis: An evidence-based implementation of next generation eu. *Journal of Policy Modeling*, 43(2):278–297, 2021.

[23] C. De la Porte and M. D. Jensen. The next generation eu: An analysis of the dimensions of conflict behind the deal. *Social Policy & Administration*, 55(2):388–402, 2021.

[24] S. Donno, C. Bagnarol, and M. Marsili. Analisi spaziale degli apprendimenti scolastici negli istituti del sud italia. Technical report, INVALSI WP 46/2020, 2020.

[25] Emiko Dupont, Isa Marques, and Thomas Kneib. Demystifying spatial confounding. *arXiv preprint*, pages 1–36, 2023.

[26] Emiko Dupont, Simon N Wood, and Nicole H Augustin. Spatial+: a novel approach to spatial confounding. *Biometrics*, 78(4):1279–1290, 2022.

[27] Eurostat. Nuts - nomenclature of territorial units for statistics, 2024. last access December 17th 2024.

[28] Anna Freni-Sterrantino, Massimo Ventrucci, and Håvard Rue. A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and spatio-temporal epidemiology*, 26:25–34, 2018.

[29] G. A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452, 2019.

[30] M. Garlaschi. L'edilizia scolastica in italia: un confronto regionale. *Osservatorio Conti Pubblici Italiani*, 2022. Last accessed December 17th 2024.

[31] Seymour Geisser and William F Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

[32] Alan E Gelfand, Dipak K Dey, and Hong Chang. Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian statistics 4*, pages 147–168, 1992.

[33] Alan E Gelfand and Penelope Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15, 2003.

[34] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition, 2004.

[35] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[36] Orazio Giancola, Luciano Benadusi, Rita Fornari, et al. Così vicine, così lontane. la questione dell'equità scolastica nelle regioni italiane. *Scuola democratica*, 1:52–79, 2010.

[37] Orazio Giancola and Luca Salmieri. Family background, school-track and macro-area: the complex chains of education inequalities in italy, 2020. working paper.

[38] Virgilio Gómez-Rubio. *Bayesian Inference with INLA*. Chapman and Hall/CRC, 2020.

[39] J.M Hammersley and P. Clifford. Markov fields on finite graphs and lattices. *Unpublished Manuscript*, 1971.

[40] Leonhard Held, Birgit Schrödle, and Håvard Rue. Posterior and cross-validatory predictive checks: a comparison of mcmc and inla. In Thomas Kneib and Gerhard Tutz, editors, *Statistical Modelling and Regression Structures*, pages 91–110. Physica-Verlag HD, 2010.

[41] James S. Hodges, Brian P. Carlin, and Q. Fan. On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, 59(2):317–322, 2003.

[42] James S. Hodges and Brian J. Reich. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334, 2010.

[43] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2nd edition, 2012.

[44] Infratel Italia. Schools dashboard, part of the ultra-broadband activation plan, 2024. last access December 19th 2024.

[45] Invalsi - Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione. National standardized assessment data (invalsi censuary surveys), 2024. Last accessed April 3rd 2025.

[46] ISTAT - Italian National Institute of Statistics. La geografia delle aree interne nel 2020: vasti territori tra potenzialità e debolezze, 2022.

[47] ISTAT - Italian National Institute of Statistics. Confini delle unità amministrative a fini statistici (administrative units borders for statistical purposes), 2025. Last accessed April 3rd 2025.

[48] ISTAT - Italian National Institute of Statistics. Sistema informativo territoriale delle unità amministrative e statistiche (territorial informative system of administrative and statistical units), 2025. Last accessed April 3rd 2025.

[49] Italian Ministry of Economic Development. Decree of july 7th 2020, 2020.

[50] Italian Ministry of Education. Ministerial decree No. 90 of May 19th 2023, 2023.

[51] Italian Ministry of Education, University and Research. Portale unico dei dati sulla scuola (unique school data portal), 2024. last accessed on December 17th 2024.

[52] Italian Official Journal. Law No. 176 of October 25th 2007, 2007.

[53] Italian Official Journal. Decree of the President of the Republic No. 81 of March 20th 2009, 2009.

[54] Italian Official Journal. Law no. 107 of july 15th 2015, 2015.

[55] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, 6th edition, 2007.

[56] G. Jona Lasinio, G. Mastrantonio, and A. Pollice. Discussing the "big n problem". *Statistical Methods and Applications*, 22:97–112, 2013.

[57] Kori Khan and Catherine A. Calder. Restricted spatial regression methods: Implications for inference. *Journal of the American Statistical Association*, 117(537):482–494, 2022.

[58] J. Lamouroux, A. Geffroy, S. Leblond, C. Meyer, and I. Albert. Addressing spatial confounding in geostatistical regression models: An r-inla approach. *arXiv preprint*, 2024.

[59] F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4):423–498, 2011.

[60] DV LlNDLEY and AFM Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34:1–41, 1972.

[61] K.V. Mardia. Multi-dimensional multivariate gaussian markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284, 1988.

[62] A. Martini. Il divario nord-sud nei risultati delle prove invalsi, 2020. Invalsi Working Paper n. 52.

[63] M. Matteucci and S. Mignani. Exploring regional differences in the reading competencies of italian students. *Evaluation review*, 38(3):251–290, 2014.

[64] K. Müller and H. Wickham. tibble: Simple data frames, 2023.

[65] W. S. Nobre, A. M. Schmidt, and J. B. Pereira. On the effects of spatial confounding in hierarchical models. *International Statistical Review*, 89(2):302–322, 2021.

[66] OECD. *PISA Data Analysis Manual: SPSS, Second Edition*. OECD Publishing, Paris, 2009.

[67] OECD. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing, 2023.

[68] Francisco Palmí-Perales, Virgilio Gómez-Rubio, Roger S. Bivand, Michela Cameletti, and Haavard Rue. Bayesian inference for multivariate spatial models with inla, 2023.

[69] Francisco Palmí-Perales, Virgilio Gómez-Rubio, and Miguel A. Martinez-Beneito. Bayesian multivariate spatial models for lattice data with inla. *Journal of Statistical Software*, 98(2):1–29, 2021.

[70] L. I. Pettit. The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(1):175–184, 1990.

[71] R Core Team. R: A language and environment for statistical computing, 2020.

[72] Brian J. Reich, James S. Hodges, and Vesna Zadnik. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206, 2006.

[73] H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[74] Haavard Rue, Sara Martino, and Nicholas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: (Methodological)*, 71(2):319–392, 2009.

[75] Haavard Rue, Andrea Riebler, Sigrunn H. Sørbye, Janine B. Illian, Daniel P. Simpson, and Finn K. Lindgren. Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017.

[76] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[77] Daniel Simpson, Haavard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1 – 28, 2017.

[78] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[79] Sigrunn Sørbye and Haavard Rue. Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014.

[80] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

[81] Arantxa Urdangarin, Tomas Goicoa, T. Kneib, and M. D. Ugarte. A simplified spatial+ approach to mitigate spatial confounding in multivariate spatial areal models. *Spatial Statistics*, 59:100804, 2024.

[82] Arantxa Urdangarin, Tomas Goicoa, and Maria Dolores Ugarte. Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*, 36(2):333–360, 2023.

[83] J. Van Niekerk and H. Rue. Low-rank variational bayes correction to the laplace method. *The Journal of Machine Learning Research*, 25(62):1–25, 2024.

[84] Janet Van Niekerk, Elias Krainski, Denis Rustand, and Haavard Rue. A new avenue for bayesian inference with inla. *Computational Statistics and Data Analysis*, 181, 2023.

[85] Janet Van Niekerk and Haavard Rue. Skewed probit regression - identifiability, contraction and reformulation. *Revstat - Statistical Journal*, 19:1–22, 2021.

[86] Xiaofeng Wang, Yu Ryan Yue, and Julian J Faraway. *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC, 2018.

[87] Sumio Watanabe. A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897, 2013.

[88] Peter Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

[89] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. A. McGowan, R. François, and H. Yutani. Welcome to the tidyverse. *Journal of open source software*, 4(43):1686, 2019.

[90] Hadley Wickham. ggplot2. *Wiley interdisciplinary reviews: computational statistics*, 3(2):180–185, 2011. R package version 3.5.1.