Stochastic Spatial Areal Models for Mapping Territorial Vulnerability

Leonardo Cefalo

February 14, 2026

Abstract

This doctoral thesis investigates territorial disparities in Italy by means of Bayesian statistical inference, with a focus on municipality-level vulnerability in two applied domains, namely education and assistance to violence against women. The empirical work consists of a total of three case studies, motivated by the need for granular and uncertainty-aware evidence to support territorial planning and policy-making.

On the methodology side, the thesis adopts hierarchical areal models including conditional autoregressive priors, considering the municipal partitioning of Italy as an irregular discrete spatial domain. Inference is carried out primarily through the Integrated Nested Laplace Approximation (INLA), which leverages the sparse precision structure of conditional autoregressive models and enables efficient model estimation. Particular attention is devoted to the issue of spatial confounding in the interpretation of regression outcomes.

On the application side, the first original research work covered herein consists in an exploratory analysis of the Italian school system data, facilitated by the implementation of an open and reproducible data infrastructure integrated in the `SchoolDataIT` R package, which downloads and harmonises information from multiple data sources and is accessible on CRAN.

The second research work studies the relationship between the students' abilities in the second year of high school and the infrastructural endowment in all Italian municipalities. Municipal student scores are obtained by averaging standardised, spatially homogeneous indicators of student outcomes provided by the Invalsi Institute for two subjects: Italian and Mathematics. To explain spatial variability in Invalsi scores, we employ a bivariate intrinsic conditional autoregressive latent model, defined alternatively at the higher aggregation levels of provinces or catchment areas of infrastructural poles. We find that alongside a significant association with infrastructural indicators chosen as explanatory variables, spatially structured latent effects are still necessary to explain the variability of student outcomes across municipalities.

The third research work analyses the access to local anti-violence centers in Apulia in 2021–2024 by spatio-temporal Poisson regression, comparing standard conditional autoregressive models, namely intrinsic CAR, proper CAR, Leroux proper CAR and the scaled Besag-York-Mollié model, and proposing a joint penalised complexity prior construction for all the models employed, extending the application field of penalised complexity priors in areal modelling. Results suggest the presence of spatial constraints in seeking help by anti-violence centers, with access declining as distance from anti-violence centers increases; in addition, lower education levels may also contribute to under-reporting in disadvantaged areas, while higher economic development may be associated with lower incidence of reported violence.

Overall, this thesis shows how Bayesian spatial modelling can be useful to quantify uncertainty, stabilise inference in sparse territorial settings, and suggest policy-relevant insights on territorial inequalities in Italy.

Contents

0.1 General Introduction

The present doctoral thesis is aimed at analysing the matter of territorial disparities in the Italian context by means of Bayesian spatial statistics, through two main fields of application, namely education and assistance to violence against women; a total of three case studies correspond to such application fields.

Even though the attention to territorial vulnerability is not unprecedented in the history of a country like Italy, whose society has never ceased to be marked by dramatic geographical divides, the current post-pandemic political context has been calling out throughout all of Europe for a wave of systematic territorial planning efforts. Such a framework of coordinated data-informed policy intervention could naturally integrate statistical analysis as an upstream operational phase, and this work has been developed indeed in the wake of measuring infrastructure and service gaps at the local level.

Three out of infinitely many legitimate questions addressing this research proposal would be: why statistics, why Bayesian, and why spatial? In this thesis, we argue that while the relevance of statistics in decision processes is enhanced by the abundance of data characterising the contemporary society, still we should not exempt ourselves from considering that statistical inference is not the sole way in which researchers can take advantage of the information extracted by data, due to the remarkable leaps automated learning and artificial intelligence are making in the complex task of prediction. Therefore, as statisticians, we concentrate our efforts on interpretable inference of the relationships amongst observed phenomena, quantifying uncertainty about the predictions that statistical models allow us to make.

Once the data are processed, we deem it useful to update *a posteriori* the model assumptions we formulate *a priori*, by computing the posterior probability distributions of the parameters defining the model. This is what Bayesian inference does at the very core of its workflow. Treating model parameters as random variables and comparing their prior and posterior probability densities is a normative choice, not universally more legitimate than seeking the closest estimator to the ontologically true value of such parameters, but only more convenient in the context of this work, also considering the increased flexibility allowed for by the choice of the prior distribution of parameters, which becomes crucial as model complexity grows.

Lastly, in contexts such as addressing territorial vulnerability, the position of an observational unit in space becomes a relevant part of information. The idea traditionally conveyed to justify, on an intuitive basis, the application of spatial statistical models is the so-called first law of geography, famously 'invoked' by Tobler (1970): *everything is related to everything else, but near things are more related than distant things*. In mathematical terms, a simple formalisation of this generic principle could be allowed after defining the probabilistic model describing the spatial process in scope.

Spatial statistics distinguishes three main classes of processes (see e.g. Cressie, 1993; Banerjee et al., 2014): *i)* point-referenced or geostatistical data, i.e., stochastic processes defined over a bounded and continuously indexed spatial domain; *ii)* areal data, i.e., stochastic processes defined over a discrete spatial domain partitioned into a countable set of areas; *iii)* point patterns, i.e., processes consisting of the occurrence of a given event over a random point in space; in this latter case it is not the value a variable takes in space that is modelled. Models for point-referenced data (ibidem) are usually characterised by the covariance between pairs of observations, which under some regularity

assumptions is a non-increasing function of distance. Regarding areal data, two popular ways to describe dependence between nearby sites are the simultaneous and conditional autoregressive models (see e.g. Besag and Kooperberg, 1995; Cressie, 1993); the value that either of these processes is expected to take on a given site only depends on the values of neighbouring sites, which conceptually can be considered as a formalisation of Tobler's law in the areal context.

In the case studies covered in the present thesis, territory is intended as partitioned into non-overlapping administrative units, specifically in municipalities, which constitute a discrete and irregular spatial domain. The appropriate class of models to describe spatial dependence thus appears to be that of areal models.

As mentioned above, spatial dependence can be modelled through two different specifications (Besag, 1974), namely simultaneous (SAR) and conditional (CAR) autoregressive models. In SAR models, dating back to Whittle (1954), spatial dependence may occur either through an endogenous dependence structure, as in the spatial lag model in which the SAR structure is assumed for the dependent variable (Anselin, 1988), or through error propagation as in the spatial error model, where the error term follows a SAR model (ibidem). In both cases, the SAR process depends *jointly* on its spatially lagged values, which means indeed that the auto-regression takes place *simultaneously* in fact.

In CAR processes, instead, the auto-regression is assumed to happen *locally*, and *conditionally* to all neighbouring sites for any site, characterising a first-order Markov dependence (Besag, 1974). In this setting, spatial dependence is not interpreted as an error propagation, but rather as a *borrowing strength* by each spatial unit from nearby units. The different specification of spatial dependence also implies a different *joint* distribution of CAR models with respect to SAR models (Cressie, 1993).

SAR models are typically employed in econometric analyses (page 33 Anselin, 1988), while CAR models have a long tradition in spatial statistics, with several applications in fields like disease mapping (Martinez-Beneito and Botella-Rocamora, 2019) and image processing (Besag et al., 1991). Additionally, the Markov properties of CARs allow for some computational advantages in model estimation, as discussed by Rue and Held (2005). In this thesis, we adopt CAR models for convenience reasons, depending both on the computational ease and the flexibility of hierarchical modelling with latent CAR fields.

The present thesis gathers three original research works, each one providing a case-study on territorial vulnerability mapping. Chapters 3 and 4 are two applied spatial analyses based on the same methodological and computational frameworks. For this reason, we dedicate the first two chapters of this thesis to the theoretical and computational aspects of these works.

Chapter 1 is mostly theoretical, and provides an overview on the statistical methodologies employed in later chapters, describing the statistical models we fit to the data, and the procedure with which models are estimated. The first part of the chapter is intended as a minimal toolbox for the four standard multivariate conditional autoregressive (CAR) areal models, starting from the simplest Intrinsic CAR (Besag, 1974; Besag and Kooperberg, 1995), then

describing three possible extensions of it, each of which introduces an additional parameter to model the strength of spatial dependence: the proper CAR (Cressie, 1993; Gelfand and Vounatsou, 2003), which can be compared to the conditional equivalent to the Simultaneous Autoregressive (Cressie, 1993; Wall, 2004; Ver Hoef et al., 2018), the Leroux proper CAR (Leroux et al., 1999) and the scaled Besag-York-Mollié convolution model (Besag et al., 1991; Riebler et al., 2016). We focus on multivariate model definitions, while the corresponding univariate case can be easily derived by deduction.

Fitting hierarchical Bayesian models implies deriving the posterior distribution of unknown estimand variables and, often, of fitted values. Evaluating posterior distributions in the present setting is hardly possible with analytical solutions, hence we rely on approximate inference. In the case of hierarchical regression models where one layer of latent effects is assumed *a priori* to be Gaussian, an extremely convenient way to carry out approximate Bayesian inference is using the Integrated Nested Laplace Approximation or INLA (Rue et al., 2009). For models with a high dimensionality of latent Gaussian effects, the INLA proves to be particularly efficient when such effects have a sparse precision matrix *a priori* (ibidem). The second part of Chapter 1 is dedicated to the INLA and its primary field of application, namely, Gaussian Markov Random Fields. We firstly characterise Gaussian Markov Random Fields, illustrating how the local Markov property implies a sparse precision structure; then, we describe all the phases of the INLA algorithm, joining the theory behind it and the practical functioning of a popular software implementing the INLA for a broad class of likelihoods and any possible Gaussian random field prior model, namely the `INLA R` package. The chapter is not intended as a compendium on areal data modelling, as it would inevitably result redundant with the most notable handbooks in the field, such as Banerjee et al. (2014) and Martinez-Beneito and Botella-Rocamora (2019). Chapter 1 is intended instead as a short summary of the theory on which the remainder of the thesis relies.

The remaining chapters constitute original research output. Chapter 2.7 presents a mostly descriptive overview on the status of the public school system in Italy, directing attention to regional disparities in school infrastructure system and to the availability of detailed data provided by institutions themselves. The availability of validated and open territorial data motivates us to construct an organic set of territorial data merging the information provided by all the relevant sources, namely the Ministry of Education, the Ministry of Enterprises, and the Italian National Institute for the Evaluation of the Education System - INVALSI, in Italian. To this aim, we developed a dedicated `R` package, named `SchoolDataIT` and published it on the Comprehensive R Archive Network. The third chapter provides a broad description of the package's functioning and includes a simple usage example to define a dataset that will be subsequently employed for a statistical analysis application, consisting of the spatial modelling of middle school INVALSI scores.

Chapter 3 describes a case study in the wake of territorial gaps in education quality and school infrastructure. Leveraging on the data shared by institutional sources and processed using the `SchoolDataIT` library, we conduct

a spatial analysis of INVALSI scores in high schools defined at the municipal level, studying territorial patterns and association with some indicators of the infrastructural state of municipalities. Throughout this study, we propose a spatial statistics approach to the analysis of education quality and students' achievements. On the methodological side, we rely on multilevel spatial regression, motivated by the sparse structure of observations, and address the issue of avoiding correlation between a latent spatial effect and some explanatory variables by extending the graph-based version of the Spatial+ methodology (Dupont et al., 2022) proposed by Urdangarin et al. (2024) to a multi-level framework. Applying such a methodology requires fitting a high number of regression models, which makes efficient model-fitting strategies necessary. The use of INLA proves crucial to match such a requirement. A general finding of this work of research is that infrastructural poles exhibit a significant advantage with respect to the remainder of the country, even when spatial trends are modelled explicitly.

Lastly, Chapter 4 addresses the challenge of modelling the spatial variability in violence against women across municipalities in the Apulia region by proposing a Bayesian spatio-temporal Poisson regression model. Using data on access to Local Anti-Violence Centers in Apulia from 2021 to 2024, we investigate the impact of municipality-level socioeconomic characteristics and local vulnerabilities on both the incidence and reporting of gender-based violence, by comparing the four standard autoregressive models, ICAR, PCAR, LCAR and scaled BYM (usually referred to as BYM2 in the literature). We extend the penalised complexity prior for the AR(1) autocorrelation parameter (Sørbye and Rue, 2017) to the spatio-temporal case, and fill the gap in the literature for prior distributions on the PCAR and LCAR hyperparameters by introducing the PC prior for the spatial autoregressive and precision mixing parameters, respectively. Our findings indicate that access to support services decreases with distance from the residential municipality, highlighting spatial constraints in reporting and the strategic importance of support center locations. Furthermore, lower education levels appear to contribute to under-reporting in disadvantaged areas, while higher economic development may be associated with a lower incidence of reported violence. This study emphasises the critical role of spatial modelling in capturing reporting dynamics and informing policy interventions.

Chapter 1

Methodological overview

1.1 Introduction

In this chapter, we summarise the statistical modelling framework followed throughout the next chapters. Specifically, we review the fundamental features of conditional autoregressive modelling, which will be applied in chapters 3 and 4. Most of the attention of this chapter is devoted to multivariate models, bearing in mind that once the multivariate model is defined, the univariate one follows straightforwardly. Additionally, we provide an overview of the Integrated Nested Laplace Approximation (Rue et al., 2009; Van Niekerk et al., 2023, INLA hereinafter), a computational methodology extensively used throughout this thesis for modelling convenience.

Models for areal data explain spatial dependence on a given spatial domain, which can be conveniently represented as a graph in which each spatial unit corresponds to a node, and each edge between two nodes corresponds to the link between two units. In this thesis, we focus on four standard models, starting from the simplest and most intuitive one, the Intrinsic Conditional Autoregressive (ICAR), which describes smooth spatial processes with autocorrelation equal to 1. Due to the high smoothness of this model, some extensions have been proposed to handle random noise; three standard extensions are the Proper CAR (PCAR), the Leroux proper CAR (LCAR), and the scaled Besag-York-Mollié convolution model (BYM2). All these models extend the ICAR by introducing a parameter controlling for the strength of spatial dependence. In a multivariate setting, this parameter can be either common across all variables or multivariate, taking a specific value for each CAR field. In this latter case, the appropriate model class is referred to as *M-models* and has been introduced in the field of disease mapping by Botella-Rocamora et al. (2015).

CAR models are usually employed as latent variables in spatial regression, tipically to include a spatially structured latent random variable for the presence of an unmeasured explanatory variable following a spatial pattern. When a spatial effect is included in a regression model to account for unmeasured

covariates, the latter may be correlated with some observed covariates. This issue can be traced to the more general problem of spatial confounding. Several approaches have been developed in two decades of literature (Urdangarin et al., 2023; Dupont et al., 2023), starting from the intuitive solution of constraining random effects to be linearly independent of observed explanatory variables (Reich et al., 2006; Hodges and Reich, 2010), which is known as restricted spatial regression (RSR); this is done by projecting the spatial field onto a subspace orthogonal to the covariates' matrix. Restricted regression manages to remove the problem, but to do so, it yields to estimated covariate effects approximating those estimated with a non-spatial model. As discussed by Khan and Calder (2022), applying such a constraint would yield undesirable implications for inference. In particular, the estimation of covariate effects approximates that of a non-spatial model, while the standard error of covariate effects is reduced, which implies a higher exposure of the so-called type-S errors, which can be thought of as the Bayesian equivalent of type-1 errors (in regression, accepting that a coefficient is different from zero when it is actually zero). This shrinkage of credibility intervals of covariate effects around roughly the same expectation has been calling out for different methods to overcome spatial confounding (Urdangarin et al., 2023). In areal data modelling, a promising method has been introduced by (Urdangarin et al., 2024) and is based on filtering out spatial trends from explanatory variables. This particular methodology appears to be particularly useful in the statistical applications of chapters 3 and 4.

In this context, inference on the statistical models involves a few variables of interest, namely regression coefficients, latent CAR fields, and a small set of parameters controlling either the distribution of the response variable or of the latent CAR fields. In Bayesian inference, analytical integration of the posterior distribution of these variables of interest is only possible 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 approximating 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, in the specific regression setting of the next chapters, based on tuning and running several models featuring high-dimensional spatially structured latent fields, employing MCMC would become computationally expensive and possibly unfeasible. On the other hand, in the context of additive models and hierarchical regression, it is possible to resort to an alternative approximation methodology, specifically suited for latent Gaussian models, possibly satisfying the local Markov property, namely the Integrated Nested Laplace Approximation or INLA.

The core idea behind INLA is to approximate the joint posterior distribution of the latent variables of interest through the Laplace approximation and then to compute the relevant marginals by numerical integration. To do so, it is necessary to outline a hierarchical regression framework in which the dependent variable y depends on a first layer of of parameters, say ϑ on which a Gaussian

distribution is assumed *a priori*, and a second layer of parameters or hyperparameters, say Ψ, on which ϑ may or may not depend as well. The INLA works in three steps:

1. Laplace-approximate the joint posterior distribution of hyperparameters
2. Approximate the joint posterior distribution of the latent Gaussian field ϑ conditional on Ψ
3. Compute the posterior marginals of ϑ and Ψ by integrating numerically the joint posteriors over Ψ

In INLA, Laplace approximation is said *nested* as it directly regards only the upstream joint posteriors, which in turn will undergo numerical integration to bring out the marginals, which are the typical inferential goal.

The intuition of Laplace approximation in posterior inference dates back to the proposal by Tierney and Kadane (1986) to approximate the first and second moments of the parameters, and leverages on the convergence of the full posterior to a Gaussian distribution when the parameters are Gaussian *a priori*, the approximation error being in $\mathcal{O}(n_y^{-1})$ where n_y is the length of the target variable.

The INLA is implemented in a comprehensive and self-sufficient `R` environment, the `INLA R` package (Gómez-Rubio, 2020; Wang et al., 2018). It will be referred to as `R-INLA` hereinafter. Since most developers and maintainers of the package are also among the main developers of the INLA methodology, the distinction between the theory of INLA and the `R-INLA` functionalities is not always meaningful.

In the literature, several handbooks have been devoted to `R-INLA`. Wang et al. (2018) and Gómez-Rubio (2020) offer a general and thorough overview of both the INLA method and the functionalities and application areas of the package. Blangiardo and Cameletti (2015) describes in detail the use of `R-INLA` in spatial modelling, with focuses on zero inflation, spatio-temporal modelling, and how to map continuously-indexed Matérn fields to discretely-index Gaussian Markov Random Field through the Stochastic Partial Difference Equation (SPDE) approach; we will return to this matter in Section 2.6. Krainski et al. (2018), lastly, deep-dives on the SPDE method with some detailed tutorials on its practical implementation in `R-INLA`.

The chapter is structured as follows. Section 1.2 provides some fundamental preliminaries to the conditional autoregressive theory, introducing Gaussian Markov Random Fields (Besag, 1974; Rue and Held, 2005), the Markovian dependency structure, and characterising the adjacency graph corresponding to the CAR spatial domain. Section 1.3 introduces the intrinsic CAR model, while section 1.4 describes the PCAR, LCAR and BYM2, both in the simpler block-factorisable versions and under the M-model extension. Section 1.5 discusses the issue of spatial confounding in the context of areal data. Next, we move our focus to the Integrated Nested Laplace Approximation, to which section 1.6 is dedicated. Throughout this section, we summarise how the approximations

work, how integration is carried out, and the main features of the `R` software which implements the INLA. Lastly, in section 1.7 we introduce the class of Penalised Complexity priors, which is particularly useful in INLA modelling, and which will be employed to various extents in all regression models in chapters 2.7, 3, and especially in chapter 4.

1.2 Gaussian Markov Random Fields

In this section, we outline some fundamental elements of the theory of conditional autoregression models, namely the reference spatial structure and the sparseness of the precision matrix implied by the assumed Markov dependence.

1.2.1 Spatial structure

Areal data can be considered as random variables defined over a discrete spatial domain, which we can represent as an undirected graph $\mathcal{G} := (\mathcal{S}, \mathcal{E})$ where $\mathcal{S} := \{s_i\}_{i=1}^n$ is the set of nodes or sites of size n; $\mathcal{E}$ is the set of edges, such that a generic element ϵ_{ij} denotes the edge between sites s_i and s_j; and $\partial_i := \{s_j : \epsilon_{ij} \in \mathcal{E}\}$ is the set of neighbours of site i.

The graph can be described as:

- The neighbourhood matrix $\boldsymbol{W}$, assumed to be binary, with $w_{ij} = 1 \iff s_j \in \partial_i$ and $w_{ij} = 0$ otherwise; notice that by convention, $w_{ii} = 0$ (cfr. Chapter 1.2).
- The diagonal degree matrix $\boldsymbol{D}$, whose elements correspond to the number of edges of each node or, alternatively, to the row or column sums of $\boldsymbol{W}$, hence $d_i = \sum_{j=1}^n w_{ij} = \text{card}(\partial_i)$
- The Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, which is singular by construction and more precisely has rank deficiency equal to the number of connected components in $\mathcal{G}$ (see e.g. Besag and Kooperberg, 1995; Hodges et al., 2003; Von Luxburg, 2007).
- The number G of connected components, a connected component being the largest possible partition of a graph in which a path can be found between any possible pair of nodes. A graph consisting of a unique connected component is said a *connected* graph, while a node with no neighbours - the smallest possible connected component - is said a *singleton*.

In this work, we will additionally assume that graphs are undirected, which means that if the value a process defined on $\mathcal{G}$ assumes in the j-th node depends on the value of the i-th node, then also the value of the i-th node depends on the value of the j-th node, and implies that $\boldsymbol{W}$ is symmetric. As a consequence, the Laplacian will be symmetric and semipositive definite. Some authors refer to it as the *unweighted* Laplacian. We will omit this specification as no other definition of Laplacian matrix will be employed presently.

1.2.2 Conditional Independence in Gaussian Random Fields

The models described throughout this chapter have the feature of being Gaussian by definition. In this section, we are going to show that if two Gaussian 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 Rue and Held (2005). It follows that the joint precision matrix of a Gaussian field satisfying conditional independence is typically sparse, namely most of its entries are typically zeroes. Sparseness implies several computational advantages, as it implies a lower computational cost of many matrix operations, including Cholesky factorisation (Rue and Held, 2005; Rue et al., 2009) which is required extensively throughout the INLA procedure (see section 1.6.

Let us consider a generic n-dimensional Gaussian random field $X = (X_1, X_2 ... X_n) \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$. Then, a deductive proof is given for this property:

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

For any $i \neq j, \quad i, j \in [1, n]$; q_{ij} is the element in position (i, j) in the precision matrix $\boldsymbol{Q} := \boldsymbol{\Sigma^{-1}}$. We can choose any set of indices $A \subset \{1, 2, \dots, n\}$ to partition X into X_A and X_B, being $B = \{1, \cdots, n\} \setminus A$ with $\text{card}(A) := m < n$, $\text{card}(B) = n - m$, namely:

$$X := \begin{pmatrix} X_A \\ X_B \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \text{ and } \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{AA}} & \boldsymbol{\Sigma_{AB}} \\ \boldsymbol{\Sigma_{BA}} & \boldsymbol{\Sigma_{BB}} \end{pmatrix}$$

Matrix $\boldsymbol{\Sigma_{AA}}$ has dimensions $m \times m$, $\boldsymbol{\Sigma_{BB}}$ is $(n - m) \times (n - m)$, $\boldsymbol{\Sigma_{AB}}$ is $m \times (n - m)$ and $\boldsymbol{\Sigma_{BA}} = \boldsymbol{\Sigma_{AB}^\top}$. Moreover, we presently assume that $\boldsymbol{\Sigma}$ is nonsingular, and so are its diagonal blocks. Let us define $\boldsymbol{\Sigma_{A|B}} := \boldsymbol{\Sigma_{AA}} - \boldsymbol{\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}}$, i.e. the Schur complement of $\boldsymbol{\Sigma_{BB}}$ in $\boldsymbol{\Sigma}$, and analogously $\boldsymbol{\Sigma_{B|A}} := \boldsymbol{\Sigma_{BB}} - \boldsymbol{\Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}}$. With some standard algebra, we see that the precision $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ reads as follows:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{\Sigma_{A|B}^{-1}} & \boldsymbol{\Sigma_{A|B}^{-1}\Sigma_{AB}\Sigma_{BB}^{-1}} \\ \boldsymbol{\Sigma_{BB}^{-1}\Sigma_{AB}\Sigma_{A|B}^{-1}} & \boldsymbol{\Sigma_{B|A}^{-1}} \end{pmatrix} \tag{1.2}$$

A fundamental property of conditional moments in Gaussian distributions (see Johnson and Wichern, 2007, Result 4.6) states that

$$\mathbb{VAR}\left[X_A | X_B\right] = \boldsymbol{\Sigma_{A|B}}$$

Yet, as we see in equation 1.2, also the element of the joint precision matrix corresponding to the subset A is equal to $\boldsymbol{\Sigma_{A|B}^{-1}}$. This means that by choosing the index set A as any couple (i, j) with $i \neq j$, such that $X_A = (X_i, X_j)^\top$ we have

$$X_i \perp X_j \Longrightarrow \boldsymbol{\Sigma_{A|B}} = \begin{pmatrix} \sigma^2_{i|B} & 0 \\ 0 & \sigma^2_{j|B} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma_{A|B}}^{-1} = \begin{pmatrix} \sigma^{-2}_{i|B} & 0 \\ 0 & \sigma^{-2}_{j|B} \end{pmatrix}$$

For two conditional variances $\sigma^2_{i|B}$ and $\sigma^2_{j|B}$. The marginal precision elements $\text{Prec}[X_i|X_B]$ and $\text{Prec}[X_j|X_B]$ are the off-diagonal entries in $\boldsymbol{\Sigma_{A|B}}^{-1}$, which are, however, two zeroes. This implies that whenever two Gaussian variables are conditionally independent, the corresponding off-diagonal precision element is zero, proving equation 1.1.

1.2.3 The Markov Property

In section 1.2.2, we have seen that conditional pairwise independence in Gaussian random fields implies the corresponding entry in the joint precision matrix 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 (Rue and Held, 2005, 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 defined as at the beginning of section 1.3. 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 *local* Markov property (Hammersley and Clifford, 1971) means that:

$$\pi(X_i|X_{-i}) = \pi\left(X_i|\{X_j : s_j \in \partial_i\}\right) \tag{1.3}$$

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

$$X_i \perp X_j | X_{-i} \quad \forall j \notin \partial_i$$

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 1.3) 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 $\boldsymbol{Q}$ which only has nonzero entries corresponding to neighbouring pairs, namely $\text{card}(\mathcal{E})$. In time series and processes defined on regular lattices $\text{card}(\mathcal{E})$ is a multiple of n; for irregular lattices, there is no such exact relationship, still the number of edges is *typically* in the order of n in connected adjacency graphs.

1.3 The intrinsic CAR model

A k-variate intrinsic CAR field defined over an undirected graph of order n is a random variable taking values $z_i \in \mathbb{R}^k$ at each i-th node, for $i \in \{1, n\}$, whose

conditional probability distribution, following Mardia (1988), (see also Banerjee et al., 2014, chapter 10), is expressed as follows:

$$z_i|z_{-i}, \boldsymbol{\Sigma} \sim \mathcal{N}_k \left(\sum_{j=1}^{n} \frac{w_{ij}}{d_i} z_j, \frac{1}{d_i} \boldsymbol{\Sigma} \right) \tag{1.4}$$

Here, $\boldsymbol{\Sigma}$ denotes the $k \times k$ variance-covariance matrix between ICAR fields. Please notice that this formulation corresponds to a special case of the CAR model by Mardia (1988), who originally proposed the more general form $z_i|z_{-i}, \boldsymbol{\Sigma} \sim \mathcal{N}_k \left(\sum_{j=1}^{n} \boldsymbol{B_{ij}} z_j, \frac{1}{d_i} \boldsymbol{\Sigma} \right)$ and presented $\boldsymbol{B_{ij}} = \frac{w_{ij}}{d_i} \boldsymbol{I_p}$ as a simplifying assumption (see Corollary 2 to Theorem 2.1, ibidem). In this model, we see that the distribution of z_i only depends on its neighbouring sites and is independent on all other ones. Recalling equation 1.3, it is clear that this class of processes follows the local Markov property.

The joint distribution of a multivariate ICAR field is:

$$\pi\left(z|\boldsymbol{\Sigma}\right) = \left(\frac{1}{2\pi}\right)^{\frac{k(n-G)}{2}} \sqrt{|\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{L}|_+} \; e^{-\frac{1}{2} z^\top \left(\boldsymbol{\Sigma^{-1}} \otimes \boldsymbol{L}\right) z} \tag{1.5}$$

Namely, an improper Normal distribution with zero mean and singular precision given by $\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{L}$, where the first factor is the precision between individual ICAR fields (inverse of the between-fields covariance matrix), and the second factor (the Laplacian) is the common precision matrix within each ICAR field. This distribution is not proper, as the Laplacian matrix is singular (Besag and Kooperberg, 1995; Hodges et al., 2003; Banerjee et al., 2014), and an ICAR model can only be identified by applying a number of linear constraints equal to the rank deficiency of $\boldsymbol{L}$. In particular, a sum-to-zero constraint is imposed to each connected component of the graph. The exponent in the normalising constant depends on the rank deficiency G of the Laplacian, as proved by Hodges et al. (2003). The $|\cdot|_+$ symbol denotes the pseudo-determinant of a matrix, i.e., the product of its non-zero eigenvalues.

1.3.1 Precision scaling

As can be seen in equation 1.5, the ICAR precision is the product of two terms, the global precision (inverse scale) and the Laplacian. This hinders parameter interpretation at the global level; to begin, let us first consider the case of an ICAR defined on a connected graph. Denoting with $\boldsymbol{\Omega} = \boldsymbol{L}^+$, the marginal variance of a generic i-th realisation of the latent effect for the h-th variable is:

$$\mathbb{VAR}[z_{ih}|\boldsymbol{\Sigma}] = \sigma_h^2 \omega_{ii}$$

Where σ_h^2 is the marginal variance of the h-th variable, corresponding to the h-th diagonal element of $\boldsymbol{\Sigma}$; in other words, ω_{ii} would be the marginal variance of z_i if the relevant global precision was equal to 1. Hence, by definition, the ICAR variance incorporates a factor determined by the neighbourhood structure. At

the global level, this consideration implies that a deterministic scale factor depending on the neighbourhood structure contributes to the process variability. An intuitive solution to make the precision parameter easier to interpret is scaling the ICAR model, as shown by Sørbye and Rue (2014). Following their operational proposal, we take the geometric mean of the diagonal of $\mathbf{\Omega}$ as scaling factor - or, in other words, the geometric mean of marginal variances if all global scale parameters (diagonal entries of $\mathbf{\Sigma}$) were equal to 1, namely

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

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 $\mathbf{\Sigma}$ from the effect on variance induced by the graph structure. To do so, the scaled model does not employ $\boldsymbol{L}$ as the structure matrix, but the scaled matrix $\boldsymbol{L}_{\text{s}} := \bar{\sigma}^2 \boldsymbol{L}$. On the other hand, the covariance matrix becomes $\bar{\sigma}^2 \Sigma$, which actually expresses the variability of the process independently of how the graph is structured.

When the graph includes $G > 1$ connected components, its Laplacian matrix is (a permutation of) the direct sum of the Laplacian matrices of its components. 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 Freni-Sterrantino et al. (2018):

$$\boldsymbol{L}_{\text{s}} = \mathcal{P} \left(\bar{\sigma}_1^2 \boldsymbol{L_1} \oplus \bar{\sigma}_2^2 \boldsymbol{L_2} \oplus \ldots \oplus \bar{\sigma}_G^2 \boldsymbol{L_G} \right) \mathcal{P}^\top$$

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

1.4 Some extensions of the Intrinsic CAR

The ICAR model has a relatively small number of parameters, i.e., $k(k+1)/2$, corresponding to the entries of the precision matrix, and is easy to interpret. However, it can be seen that it is not a flexible model, in that it assumes a spatial autocorrelation parameter equal to 1. It can be thought of as the spatial equivalent of the order 1 Random Walk process in time series analysis: it is suited to detect strong spatial structures, but is misspecified when spatial trends are not particularly smooth. In such cases, the possible outcomes are two: either the model fails at capturing random noise in data, or it does adapt to data but at the price of a higher complexity (overfitting). For this reason, some extensions have been developed in the literature to account for noise. Here we introduce three popular extensions: the so-called Proper CAR (Gelfand and Vounatsou, 2003), the Leroux proper CAR (Leroux et al., 1999), and the scaled Besag-York-Mollié (BYM2 Riebler et al., 2016) convolution model.

1.4.1 The Proper CAR

An intuitive generalisation of the ICAR is obtained introducing a parameter, say $\rho \in [0, 1[$, controlling for spatial autocorrelation (Banerjee et al., 2014, chapter 10). This would yield the conditional prior distribution:

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

Please notice that ρ is a scalar here. As proved in appendix .1, the joint distribution of PCAR fields is:

$$z | \boldsymbol{\Sigma}, \rho \sim \mathcal{N}_{kn} \left(0, \boldsymbol{\Sigma} \otimes (\boldsymbol{D} - \rho \boldsymbol{W})^{-1} \right) \tag{1.6}$$

In the absence of singletons, i.e., nodes with zero neighbours, this is a proper distribution in that, considering the multivariate case, for $\rho \in [0, 1[$ the within-fields precision matrix $\boldsymbol{D} - \rho \boldsymbol{W}$ is symmetric and positive definite[1]. This can be seen starting from the eigendecomposition of the row-stochastic matrix $\boldsymbol{D^{-1}W}$. The Gershgorin disk theorem (Horn and Johnson, 2012, section 6.1) implies all its eigenvalues lie within the unit disk. This matrix is similar to $\boldsymbol{D^{-1/2}WD^{-1/2}}$, hence these two matrices share the same eigenvalues. If $\rho \in [0, 1[$, the eigenvalues of the symmetric matrix $\boldsymbol{I}_n - \rho \boldsymbol{D^{-1/2}WD^{-1/2}}$ must then be all strictly positive. The within-fields precision matrix is given by the quadratic form $\boldsymbol{D^{1/2}} \left(\boldsymbol{I}_n - \rho \boldsymbol{D^{-1/2}WD^{-1/2}} \right) \boldsymbol{D^{1/2}}$ and, if $\rho \in [0, 1[$, it is positive definite for any full-rank degree matrix $\boldsymbol{D}$ (see also Horn and Johnson, 2012, Observation 7.1.8).

One advantage of this model is the ease in interpretation, as ρ quantifies the extent to which the value of z_j depends on the neighbouring sites. Clearly, the limit case for $\rho = 1$ is the ICAR, while the limit case for $\rho = 0$ is an independent but *not* identically distributed model, in which the variance is proportional to the inverse degree matrix $\boldsymbol{D}^{-1}$, hence a site with a higher number of neighbours will have a smaller prior conditional variance.

Throughout this thesis, the main reference for the proper CAR will be Gelfand and Vounatsou (2003), even though the model can be traced back in a more general formulation at least to Cressie (1993); this choice is due to the more specific model construction provided by the former reference.

1.4.2 The Leroux proper CAR

As we have seen, a partial drawback of the PCAR model is that the limit case in the absence of spatial autocorrelation is not an IID process. A simple solution would be to employ a rather similar model proposed by Leroux et al.

[1] For this reason, this model is known in the literature as the 'proper' CAR. Properness is far from an exclusive feature of this model, and is not even limited to the models listed in this section, yet this naming has become a convention, and we respect this convention for notation brevity.

(1999). This time, the idea is to balance the precision matrix between the ICAR precision (proportional to the Laplacian) and the precision of an IID process (proportional to the identity matrix). To do so, a precision mixing parameter is introduced, which we label as $\lambda \in [0, 1]$. The starting point is the conditional prior:

$$z_i|z_{-i}, \boldsymbol{\Sigma} \sim \mathcal{N}_k \left(\sum_{j=1}^{n} \frac{\lambda w_{ij}}{1 - \lambda + \lambda d_i} z_j, \frac{1}{1 - \lambda + \lambda d_i} \boldsymbol{\Sigma} \right)$$

As can be seen, for $\lambda = 0$ the conditional prior becomes $\mathcal{N}_k(0, \boldsymbol{\Sigma}) \quad \forall i$, and for $\lambda = 1$ we revert to the ICAR. After some algebra, the joint distribution reads as:

$$z|\boldsymbol{\Sigma}, \lambda \sim \mathcal{N}_{kn} \left(0, \boldsymbol{\Sigma} \otimes [\lambda \boldsymbol{L} + (1 - \lambda)\boldsymbol{I}_n]^{-1}\right) \tag{1.7}$$

Interpreting this model is less simple, as it involves understanding the meaning of precision mixing. As we have seen in section 1.2, the precision matrix of a Gaussian Markov Random Field informs us ofal the *conditional* dependence structure of the spatial domain, and in both the LCAR and PCAR models the off-diagonal precision elements take the form $-\alpha$ or $-\lambda$ times the global precision parameter for neighbouring sites, hence roughly speaking both these parameters can be read as the strength of spatial association.

1.4.3 The Besag-York-Mollié model

The original version of this model proposed by Besag et al. (1991) was the convolution of a univariate ICAR, say $u' \sim \mathcal{N}(0, \sigma_u^2 \boldsymbol{L}^+)$, and a IID process, say $v \sim \mathcal{N}_n(0, \sigma_v^2 \boldsymbol{I_n})$. The underlying idea is to control for both spatial variation and random noise. This is particularly useful in disease mapping contexts, in which the dependent variable is usually Poisson distributed and the linear predictor needs a term to account for random noise, which is not captured by the multivariate ICAR component alone. Some drawbacks can be found, however, in this model. The two components u' and v are not separable and their convolution $z =: u' + v$ is what needs being modelled.

This formulation hinders the interpretation of the extent to which process variability is driven either by spatial autocorrelation or by noise. For this reason, Dean et al. (2001) introduced a reparametrisation of the BYM which allows to model a unique scale parameter σ^2 in place of the two parameters σ_u^2 and σ_v^2, replacing the former with $\phi\sigma^2$ and the latter with $(1 - \phi)\sigma^2$. To do so, z is defined as the weighted sum of an ICAR field with unit global variance and an IID field, introducing the variance mixing parameter $\phi \in [0, 1]$, such that $z \sim \mathcal{N}_n \left(0, \sigma^2 \phi \boldsymbol{L}^+ + \sigma^2 (1 - \phi)\boldsymbol{I}_n\right)$ [2]

Additionally, Riebler et al. (2016) proposed a further reparametrisation allowing to scale the variance of the ICAR component, as seen in section 1.3.1. We will refer to this model as scaled BYM or BYM2. The univariate BYM2

[2] Authors labelled the variance mixing parameter as λ but we maintain the ϕ notation to avoid confusion with the LCAR model.

model becomes thus:

$$\begin{cases} z = \sigma\sqrt{\phi}u + \sigma\sqrt{1-\phi}v \\ u \sim \mathcal{N}_n\left(0, \boldsymbol{L}_s^+\right) \\ v \sim \mathcal{N}_n\left(0, \boldsymbol{I}_n\right) \end{cases}$$

Since the Laplacian matrix is scaled, the two addenda are now defined on the same scale of variability, which implies the mixing parameter can be interpreted as the balancing of variability between spatial structure and noise.

To characterise the multivariate BYM distribution, let us consider $\boldsymbol{\Sigma} = \boldsymbol{M}^\top \boldsymbol{M}$ for a generic full-rank square matrix $\boldsymbol{M}$. In the multivariate case, suppose all the ICAR fields are independent and have the same distribution, i.e. $U_j \sim \mathcal{N}_n(0, {\boldsymbol{L}_\boldsymbol{s}}^+) \quad \forall j \in [1, k]$ and $\boldsymbol{U} \sim \mathcal{N}_{kn}(0, \boldsymbol{I_k} \otimes \boldsymbol{L}^+)$. The same assumption holds for the IID component, i.e. $\boldsymbol{V} \sim \mathcal{N}_{kn}(0, \boldsymbol{I_k} \otimes \boldsymbol{I_n})$. Then, let us define the multivariate BYM field, arranged in matrix form, as

$$\boldsymbol{z} = \sqrt{\phi}\boldsymbol{U}\boldsymbol{M} + \sqrt{1-\phi}\boldsymbol{V}\boldsymbol{M}$$

In this formula, we rely on the simplifying assumption of the same variance-covariance parameter for both the ICAR and the IID component. This is not as immediate as in the univariate case, in which the univariate variance parameter could be easily rescaled leveraging on the presence of the mixing parameter. Under this assumption, the joint distribution of $\boldsymbol{z}$ reads as:

$$\boldsymbol{z}|\boldsymbol{\Sigma}, \phi \sim \mathcal{N}_{kn}\left(0, \Sigma \otimes (\phi \boldsymbol{L}_s^+ + (1-\phi)\boldsymbol{I}_n)\right) \tag{1.8}$$

As can be seen, the precision of the BYM2 field is a dense matrix. Operations such as inversion, eigendecomposition, or multiplication, would then have cubic computational cost, which in most real-world applications is prohibitive. Riebler et al. (2016), therefore, suggest to model z and u jointly, as a process of length $2n$ (or $2kn$ in the multivariate case). The resulting joint process has a sparse precision matrix, retaining the same rank deficiency as $\boldsymbol{L}$, which requires the sum-to-zero constraint on the ICAR component U. To sum up, the BYM2 model boasts some desirable features, namely:

- Flexibility (as the PCAR and LCAR)
- Limit cases being the ICAR and the IID models for $\phi = 1$ and $\phi = 0$, respectively (as for the LCAR)
- Straightforward interpretation (as for the PCAR)
- Scalability (as for the ICAR)

This explains its popularity in areal data analysis and especially in the disease mapping literature.

1.4.4 M-models

For all the autoregressive models discussed above, we observed that the precision (or the covariance) matrix can be conveniently factorised as the Kronecker

product of a precision (or scale, respectively) parameter and a structure matrix, which is deterministic for the ICAR and unknown but independent on the precision parameter for the other models; in this latter case, the structure matrix only depends on one scalar-valued parameter (the AR parameter for the PCAR, the precision mixing for the LCAR and the variance mixing for the BYM2). We may thus refer to these models as *block-factorisable*, and their joint variance-covariance matrix can be written as

$$\mathbb{VAR}\left[z|\Psi\right] = \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega}$$

Where Ψ is the full set of model parameters, $\boldsymbol{\Sigma}$ is the *between*-spatial fields covariance matrix, and $\boldsymbol{\Omega}$ is the *within*-spatial fields covariance matrix, specifically:

- $\boldsymbol{\Omega} = \boldsymbol{L}^{+}$ for the ICAR
- $\boldsymbol{\Omega} = (\boldsymbol{D} - \rho\boldsymbol{W})^{-1}$ for the PCAR
- $\boldsymbol{\Omega} = [\lambda\boldsymbol{L} + (1-\lambda)\boldsymbol{I_n}]^{-1}$ for the LCAR
- $\boldsymbol{\Omega} = \phi\boldsymbol{L}^{+} + (1-\phi)\boldsymbol{I_n}$ for the BYM2

However, the hypothesis of a constant spatial dependence parameter for *all* the spatial fields can be strong. A more flexible class of models, introduced by Botella-Rocamora et al. (2015), leverages on the decomposition of the scale parameter Σ into $\boldsymbol{M}^{\top}\boldsymbol{M}$, where $\boldsymbol{M}$ is a generic $p \times k$ full-rank matrix for any p; these models are known in literature as M-models. M-models generalise proper autoregressive models by relaxing the hypothesis of a scalar-value hyperparameter controlling the strength of spatial association, and allow for p different hyperparameters.

The core idea is defining $\boldsymbol{z} := \boldsymbol{z}^{(0)}\boldsymbol{M}$, where $\boldsymbol{z}^{(0)} = (z_1^{(0)}, z_2^{(0)} \dots z_p^{(0)})$ and the $z_1^{(0)} \dots z_p^{(0)}$ fields are independent, each one following a prior of the form: $z_j^{(0)} \sim \mathcal{N}_n(0, \Omega_j) \quad \forall j \in \{1, \dots, p\}$ where $\Omega_j$is the j-th *within*-fields spatial covariance matrix, i.e. the one specific of the j-th spatial field.; taking the PCAR model as an example we would have $\Omega_j = (\boldsymbol{D} - \rho\boldsymbol{W})^{-1}$.

Although the $\boldsymbol{M}$ matrix does not have to be square, if the number of variables is not prohibitive, models with square $\boldsymbol{M}$, i.e. $p = k$, are easier to interpret and allow for exactly one spatial association parameter for each variable.

Hence, in general, M-models feature the following representation:

$$\text{M - models:} \qquad \mathbb{VAR}\ \left[z|\Psi\right] = \left(\boldsymbol{M}^{\top} \otimes \boldsymbol{I_n}\right)\ \boldsymbol{\Omega}\ \left(\boldsymbol{M} \otimes \boldsymbol{I_n}\right) \tag{1.9}$$

With $p = k$, in the PCAR model the $\boldsymbol{\Omega}$ matrix is

$$\text{PCAR:} \qquad \boldsymbol{\Omega} = (\boldsymbol{I_p} \otimes \boldsymbol{D} - \boldsymbol{\rho} \otimes \boldsymbol{W})^{-1}$$

where $\boldsymbol{\rho} = \text{diag}(\rho_1, \rho_2, \dots, \rho_k)$ (in the block-factorisable case seen in subsection 1.4.1, with an autoregressive parameter common to all the k spatial fields, we had $\boldsymbol{\rho} = \rho\boldsymbol{I_k}$). Likewise for the CAR:

$$\text{LCAR:} \qquad \boldsymbol{\Omega} = [\Lambda \otimes \boldsymbol{L} + (\boldsymbol{I}_p - \Lambda) \otimes \boldsymbol{I}_n]^{-1}$$

Where, again, $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_k)$. Regarding the BYM2, as in section 1.4.3, we rely on the simplifying hypothesis that:

$$\text{BYM2}: \qquad \boldsymbol{z} = \boldsymbol{U}\boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{M} + \boldsymbol{V}(\boldsymbol{I_k} - \boldsymbol{\Phi})^{\frac{1}{2}}\boldsymbol{M} \tag{1.10}$$

i.e. both the ICAR and IID components share the same between-fields covariance matrix. This hypothesis is not admitted, e.g., by Corpas-Burgos et al. (2019), who did not employ the BYM2 parametrisation and admitted a different scale parameter for each of the two BYM2 components. On the other hand, the unique $\boldsymbol{M}$ approach is followed by Urdangarin et al. (2024), who indeed extend the parametrisation of Riebler et al. (2016) to the M-model case. This is the approach to which we will commit. Equation 1.10 then implies

$$\text{BYM2:} \qquad \boldsymbol{\Omega} = \left[\boldsymbol{\Phi} \otimes \boldsymbol{L_s^+} + (\boldsymbol{I_k} - \boldsymbol{\Phi}) \otimes \boldsymbol{I}_n\right] \tag{1.11}$$

Lastly, in Appendix .2 we show that sparse parametrisation is also possible for the multivariate BYM2 when z and U are modelled jointly, allowing to avoid dense precision issues.

1.5 Spatial Confounding

Spatial confounding can occur when a regression model includes a spatially structured latent random variable, say z, correlated with some explanatory variables, say $\boldsymbol{X}$. This issue implies a competition between $\boldsymbol{X}$ and z in explaining y, introducing bias in the estimation of β. The consideration that $\boldsymbol{X}$ is directly observed and is part of the available information and the estimation β is frequently the main inferential target may encourage the analyst to restrict z to a subspace in which it cannot be confounded with observed covariates. Restricted regression serves this issue precisely (Reich et al., 2006; Hodges and Reich, 2010). In the following paragraphs, we explain how restricted regression works, we prove that it implies the same posterior expectation for β as the non-spatial model in Gaussian settings, and discuss an alternative method to address spatial confounding with areal data.

1.5.1 Preliminary: Latent Gaussian model outline

To understand restricted regression, let us briefly outline a general latent Gaussian linear model and assume:

$$\begin{cases} \mathbb{E}\left[y|\eta, \Psi\right] = g^{-1}(\eta) \\ \eta = \boldsymbol{A}\vartheta \\ \vartheta|\Psi \sim \mathcal{N}(0, \boldsymbol{Q_\Psi}^{-1}) \end{cases}$$

where y denotes the response variable consisting of n_y observations; η is the linear predictor linked by a function g to the expected value of y; ϑ is a generic vector of latent variables of interest; $\boldsymbol{A}$ is a known design matrix; Ψ is the array of hyperparameters, and $\boldsymbol{Q_\Psi}$ is the prior precision matrix of ϑ. Let us assume

that the linear predictor, only depends on an array of effects of explanatory variables, say $\beta \in \mathbb{R}^p$ and an array of spatial latent effects [3], say $z \in \mathbb{R}^{n_z}$ for two integers p and n_z. if z is multivariate, in regression model we employ vec z instead of its matrix representation. The design matrix and the vector of latent variables can be then written as

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

where $\boldsymbol{X}$ is a matrix of explanatory variables or covariates of size $n_y \times p$, and $\boldsymbol{\xi}$ is the design matrix of latent effects of size $n_y \times n_z$. In many cases, such as in typical spatial disease mapping models, $n_y = n_z$, though this assumption is not necessary and is violated e.g., in multilevel models, such as the case study in Chapter 3.

Let us further assume that $(y_j \perp\!\!\!\perp y_k)|\vartheta, \Psi$ and $(y_j \perp\!\!\!\perp \eta_k)|\vartheta, \Psi$ for any $j, k \in [1, n_y]$ with $j \neq k$, so each observation y_j depends on ϑ only through one value η_j and the likelihood can be factorised as:

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

For illustration purposes, let us now assume that y is Gaussian and the latent spatial effect z has prior mean 0 and precision matrix $\boldsymbol{Q}(\Psi) = \tau \boldsymbol{R}$, where $\boldsymbol{R}$ is a positive semidefinite structure matrix and $\tau \in \Psi$ is a global precision parameter; for univariate spatial autoregressive models described in sections 1.3 and 1.4, $\boldsymbol{R} = \boldsymbol{\Omega}^+ = \boldsymbol{L}$ for the ICAR $\boldsymbol{R} = \boldsymbol{\Omega}^{-1}$ if the model is proper. In such Gaussian regression framework, the regression model becomes:

$$\begin{cases} y = \boldsymbol{X}\beta + \boldsymbol{\xi}z + \varepsilon \\ z \sim \mathcal{N}_{n_z}(0, \tau^{-1}\boldsymbol{R}^{-1}) \\ \beta \sim \mathcal{N}_p(0, N\boldsymbol{I_p}) \\ \varepsilon \sim \mathcal{N}_{n_y}(0, \tau_\varepsilon^{-1}\boldsymbol{I_{n_y}}) \end{cases} \tag{1.12}$$

Where $\tau_\varepsilon \in \Psi$ is the error precision parameter and N is an arbitrarily large scalar. Let us firstly define the following quantities to make formulas more tractable:

$$\begin{cases} \bar{r} := \frac{\tau}{\tau_\varepsilon} \\ r_\varepsilon := \frac{1}{\tau_\varepsilon N} = \frac{\sigma_\varepsilon^2}{N} \\ \boldsymbol{G} := (\boldsymbol{X}^\top \boldsymbol{X} + r_\varepsilon \boldsymbol{I_p}) \\ \boldsymbol{S_{11}} := (\boldsymbol{\xi}^\top \boldsymbol{\xi} + \bar{r}\boldsymbol{R} - \boldsymbol{\xi}^\top \boldsymbol{X}\boldsymbol{G}^{-1}\boldsymbol{X}^\top \boldsymbol{\xi}) \end{cases} \tag{1.13}$$

[3] Labelling β as "fixed" and z as "random" effects is quite frequent in hierarchical regression applications, but due to the polysemic 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 $\boldsymbol{X}$ or $\boldsymbol{\xi}$), 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 ($\boldsymbol{X}$) and y, while z is an unobserved process shaping the structure of y itself.

Then, the posterior expectation of β and z, conditional on hyperparameters, can be derived analytically. The proof is left in Appendix .3. For the covariate effects, we have:

$$\mathbb{E}[\beta|y,\tau_\varepsilon,\bar{r}] = \boldsymbol{G}^{-1}\left[\boldsymbol{X}^\top + \boldsymbol{X}^\top\boldsymbol{\xi}{\boldsymbol{S_{11}}}^{-1}\boldsymbol{\xi}^\top\left(\boldsymbol{X}\boldsymbol{G}^{-1}\boldsymbol{X}^\top - \boldsymbol{I_{n_y}}\right)\right] y \tag{1.14}$$

In a non-spatial regression model, with $\boldsymbol{A} = \boldsymbol{X}$ and $\vartheta = \beta$, this expectation reduces to $\boldsymbol{G}^{-1}\boldsymbol{X}^\top y$.

Regarding the latent spatial field, we have:

$$\mathbb{E}[z|y,\tau_\varepsilon,\bar{r}] = {\boldsymbol{S_{11}}}^{-1}\left(\boldsymbol{\xi}^\top - \boldsymbol{\xi}^\top\boldsymbol{X}\boldsymbol{G}^{-1}\boldsymbol{X}^\top\right) y \tag{1.15}$$

Moreover, if a non-informative prior is defined on β, we have $N \longrightarrow \infty$, with $\boldsymbol{G} \longrightarrow \boldsymbol{X}^\top\boldsymbol{X}$. In this case, we see that the limits for equations 1.14 and 1.15 are the results stated by (Reich et al., 2006, page 1199). Under a non-spatial model with a flat prior on β, instead, we obtain the familiar ordinary least square estimator as the posterior mean of β. As we see in this particular case with a Gaussian likelihood, the posterior expectation of latent effects and covariate effects can be expressed with a closed formula. This will be useful in the forthcoming section.

1.5.2 Restricted Spatial Regression

In this section, we describe how restricted spatial regression (RSR) works under the simplifying hypothesis that $\boldsymbol{\xi} = \boldsymbol{I_{n_y}}$, which also implies $n_y = n_z := n$. For a generic full column rank $\boldsymbol{\xi}$ model matrix, we would need to replace $\boldsymbol{X}$ with $\left(\boldsymbol{\xi}^\top\boldsymbol{\xi}\right)^{-1}\boldsymbol{\xi}^\top\boldsymbol{X}$ in the forthcoming calculations, as we will do in chapter 3 when applying RSR. Let us now consider the projection matrix of y onto its least squares fitted values, and its orthogonal complement, respectively:

$$\boldsymbol{P} := \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}X^\top \qquad \text{and} \qquad \boldsymbol{P}_\perp := \boldsymbol{I_{n_y}} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{I_{n_y}} - \boldsymbol{P}$$

The crucial step of all the subsequent analysis is the eigendecomposition of $\boldsymbol{P}_\perp$, which is useful to project z on a subspace orthogonal to covariates. Firstly, consider that its first $n_y - p$ eigenvalues are ones, and the latter p are zeroes. To see that the eigenvalues of $\boldsymbol{P}_\perp$ are either zeros or ones, first consider that $\boldsymbol{P}_\perp = \boldsymbol{P}_\perp\boldsymbol{P}_\perp$. Then for a generic eigenvalue-eigenvector pair λ, v it holds that $\boldsymbol{P}_\perp v = \lambda v$. With no loss of generality it is also true that $\boldsymbol{P}_\perp\boldsymbol{P}_\perp v = \lambda\boldsymbol{P}_\perp v$, but since $\boldsymbol{P}_\perp$ is idempotent, it follows that $\boldsymbol{P}_\perp v = \lambda^2 v$, implying $\lambda v = \lambda^2 v$ for a nontrivial v, hence $\lambda = \lambda^2$. The number of nonzero eigenvalues is the rank of $\boldsymbol{P}_\perp$, which is in turn $n_y - \mathrm{rk}(\boldsymbol{P})$, hence $n - p$.

To prove the implication of restricted regression on the posterior expectation of β, let us partition the eigenvector matrix of $\boldsymbol{P}_\perp$ into the matrices $(\boldsymbol{E}\,\boldsymbol{K})$; $\boldsymbol{E}$ are the first $n_y - p$ eigenvectors corresponding to nonzero eigenvalues, while $\boldsymbol{K}$ belongs to the null space of $\boldsymbol{P}_\perp$. Thus, the decomposition of $\boldsymbol{P}_\perp$ is:

$$\boldsymbol{P}_\perp = \begin{pmatrix}\boldsymbol{E} & \boldsymbol{K}\end{pmatrix}\begin{pmatrix}\boldsymbol{I_{n_y-p}} & \boldsymbol{0_{(n_y-p)\times p}} \\ \boldsymbol{0_{p\times(n_y-p)}} & \boldsymbol{0_{p\times p}}\end{pmatrix}\begin{pmatrix}\boldsymbol{E}^\top \\ \boldsymbol{K}^\top\end{pmatrix} = \boldsymbol{E}\boldsymbol{E}^\top$$

Notice that:

$$\begin{pmatrix} \boldsymbol{E}^\top \\ \boldsymbol{K}^\top \end{pmatrix} (\boldsymbol{E}\,\boldsymbol{K}) = (\boldsymbol{E}\,\boldsymbol{K}) \begin{pmatrix} \boldsymbol{E}^\top \\ \boldsymbol{K}^\top \end{pmatrix} = \boldsymbol{I}_{\boldsymbol{n_y}}\,, \quad \boldsymbol{K}\boldsymbol{K}^\top = \boldsymbol{P}, \quad \boldsymbol{K}^\top \boldsymbol{E} = 0$$

Without loss of generality, it holds that

$$\ln \pi(z|\tau) = C - \frac{\tau}{2} z^\top \boldsymbol{R} z = C - \frac{\tau}{2} z^\top (\boldsymbol{E}\,\boldsymbol{K}) \begin{pmatrix} \boldsymbol{E}^\top \\ \boldsymbol{K}^\top \end{pmatrix} \boldsymbol{R} \, (\boldsymbol{E}\,\boldsymbol{K}) \begin{pmatrix} \boldsymbol{E}^\top \\ \boldsymbol{K}^\top \end{pmatrix} z$$

Where C is an additive constant. Let us now define $\zeta = \begin{pmatrix} \boldsymbol{E}^\top \\ \boldsymbol{K}^\top \end{pmatrix} z$, such that $z = (\boldsymbol{E}\,\boldsymbol{K})\,\zeta$ and decompose the newly defined spatial effect ζ as

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

The distribution of ζ then reads as:

$$\zeta \sim \mathcal{N}_{n_z} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tau^{-1} \begin{pmatrix} \boldsymbol{E}^\top \boldsymbol{R}\boldsymbol{E} & \boldsymbol{E}^\top \boldsymbol{R}\boldsymbol{K} \\ \boldsymbol{K}^\top \boldsymbol{R}\boldsymbol{E} & \boldsymbol{K}^\top \boldsymbol{R}\boldsymbol{K} \end{pmatrix}^{-1} \right)$$

Restricted spatial regression, then, works by constraining z onto the null space of $\boldsymbol{P}$, which is orthogonal to $\boldsymbol{X}$:

$$\text{RSR} : \zeta_0 = \boldsymbol{K}^\top z = 0$$

Specifically, the regression model under RSR reads as:

$$\text{RSR} : \quad y = \boldsymbol{X}\beta + \boldsymbol{E}\zeta_1 + \varepsilon \tag{1.16}$$

In this new regression framework, recalling equation 1.16 and considering the property of the conditional variance in Gaussian distributions (Result 4.6 of Johnson and Wichern, 2007, see also section 1.2), we obtain that ζ_1 is Normal with precision $\tau \boldsymbol{E}^\top \boldsymbol{R}\boldsymbol{E}$. Therefore, substituting the model matrix $\boldsymbol{\xi} = \boldsymbol{E}$ into equation 1.14, we see that the expectation of covariate effects reads as:

$$\text{RSR} : \quad \mathbb{E}[\beta|y, \tau_\varepsilon, \bar{r}] = \boldsymbol{G}^{-1} \boldsymbol{X}^\top y \tag{1.17}$$

Which is exactly the same results, in terms of expectation, that one could obtain with a much simpler non-spatial regression model involving a smaller number of parameters to estimate.

1.5.3 Beyond Restricted Regression

We have just seen that under RSR the posterior means of covariate effects approximate those of a non-spatial model, and even more so when y is Gaussian, the two estimates are equal. Interestingly, Dupont et al. (2023) 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 as a model ignoring the existence of the spatial component. In terms of interpretation, adding a spatial effect forced to be orthogonal to covariates would not significantly alter the point estimation of covariate effects.

A more recent strand of literature criticising RSR has therefore begun developing, starting from Khan and Calder (2022), who noticed that RSR yields similar coefficient estimates as nonspatial regression, *yet* it underestimates coefficient variability, resulting in reduced coverage with respect to nonspatial regression.

For this reason, alternative approaches have been sought in recent years. Among our aims, the most useful one is represented by filtering out spatial trends from the explanatory variables. Using a non-parametric correction (Urdangarin et al., 2024) is particularly convenient for our aims. Before diving into details, we find it useful to discuss another method proposed by Hughes and Haran (2013) with the same purpose as RSR, which leverages on the spectral properties of the graph.

Filtering based on the Moran operator

Several alternatives to RSR have been proposed in the last fifteen years. One of the earliest ones was introduced by Hughes and Haran (2013) and can be thought of as a refinement of RSR. In brief, authors noticed that the constraints used in RSR are not informative on the spatial structure of the data; they developed a restriction methodology based on the eigenvectors of what authors define as the generalised Moran operator, rather than the projection matrix.

To gain some context, let us first recall that the Moran index $I(z)$ referring to a vector z of length n taking values over a graph with neighbourhood matrix $\boldsymbol{W}$ is given by

$$I(\boldsymbol{W}) = \frac{n_z}{\mathbf{1}^\top \boldsymbol{W} \mathbf{1}} \frac{z^\top \mathcal{C} \boldsymbol{W} \mathcal{C} z}{z^\top \mathcal{C} z} \tag{1.18}$$

where $\mathcal{C}$ is the idempotent centering matrix $\mathcal{C} := \boldsymbol{I}_{\boldsymbol{n_z}} - \frac{1}{n_z}\mathbf{1}\mathbf{1}^\top$, hence $\mathcal{C}z = z - \bar{z}$.

The Moran coefficient plays a crucial role in hypothesis testing, as under the assumption of spatial independence it follows (for Normal z Cliff and Ord, 1981) or converges in distribution to (for non-Normal x Griffith, 2010) a Normal distribution with known mean and variance; which allows to test easily for the absence of spatial autocorrelation across an observed variable.

In the context of addressing spatial confounding, Hughes and Haran (2013) introduce what they refer to as the *generalised Moran statistic for $\boldsymbol{X}$ with respect to the graph*, which is given by replacing the centering matrix $\mathcal{C}$ in the above formula with $\boldsymbol{P}_\perp = \boldsymbol{I}_{\boldsymbol{n_y}} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. To this aim, they also introduce the *generalised Moran operator for $\boldsymbol{X}$*, defined as $\boldsymbol{P}_\perp \boldsymbol{W} \boldsymbol{P}_\perp$. As an example, this gif shows the structure of the 30 eigenvectors associated with the largest eigenvalues of a Moran operator defined on a set of 4 randomly generated Uniform variables over a 30×30 regular lattice; the neighbourhood matrix in this case follows the rook criterion (neighbours must share an edge). What Hughes and Haran (2013) propose is to project z onto the space spanned by a

small number of low-frequency eigenvectors of the generalised Moran operator, which are still orthogonal to explanatory variables, but already constitute a set of spatial trends. This approach is particularly interesting in that the collinearity between the unmeasured confounder and the explanatory variables is ruled out by means of a spatial filter and not by automatic constraints.

1.5.4 Spatial+ and its variants

Spatial confounding arises when one or more observed covariates are linearly dependent on an unmeasured covariate, which needs being estimated with a latent spatial model. Both restricted regression and the methodology of Hughes and Haran (2013) set some conditions on the latent spatial field. The "Copernican revolution" of treating spatial confounding came from Dupont et al. (2022), who proposed, on the contrary, to remove the spatial trend from observed covariates. Authors labelled this methodology as *Spatial+*. The original application of Spatial+ involved point-referenced data and thin plate (TP) splines; the authors fitted a TP-spline to the covariate and employed the resulting residual as the new explanatory variable. However, the method can be extended with no effort to areal models, provided the appropriate prior model specification (Marques and Wiemann, 2023).

In Bayesian inference, Spatial+ requires assuming a nested parametric model on covariates. On the one side, this tends to propagate uncertainty (Marques and Wiemann, 2023) and complicate the probabilistic structure of the model, as the covariate itself is now treated as a stochastic process.

A variant of Spatial+ has recently been proposed by Urdangarin et al. (2024) to circumvent spatial modelling of covariates, though in the specific context of graph-based models. In line with (Lamouroux et al., 2024), we label this method as *Spatial+2.0*. As we have seen in the previous paragraph, the graph itself can be informative about the spatial structure of a variable defined over it, and graph-related information can be useful to define a spatial filter. The Spatial+2.0 method provides a non-parametric solution to spatial confounding which does not add any layer of uncertainty to covariates.

As we have seen, one way of defining a spatial filter is through the eigenvectors of the generalised Moran operator. However, if filtering is to be applied on covariates, this method is not feasible as the generalised Moran operator depends itself on covariates. Urdangarin et al. (2024), therefore, propose to generate deterministic spatial trends from the smoothest eigenvectors of the Laplacian matrix.

The intuition that spatial confounding arises when covariates are not independent on the smoothest Laplacian eigenvectors can be traced back to Reich et al. (2006) and is also briefly discussed by Khan and Calder (2022). The use of Laplacian eigenvectors, on the other hand, is a milestone in the spatial clustering culture, as thoroughly discussed by Von Luxburg (2007).

A property of the Laplacian matrix is that for any vector V with length n equal to the order of $\boldsymbol{L}$, it holds that $V^\top \boldsymbol{L} V = \sum_i^n \sum_{j \in \partial_i} (v_i - v_j)^2$ where ∂_i is the set of neighbours of the i-th site (see, e.g. Von Luxburg, 2007). Specifically,

if V is chosen as an eigenvector of $\boldsymbol{L}$, this equality implies that

$$\ell = \sum_{i}^{n} \sum_{j \in \partial_i} (v_i - v_j)^2 \tag{1.19}$$

where ℓ is the Laplacian eigenvalue associated with V. Equation 1.19 implies, in turn, that each Laplacian eigenvalue equals the sum of squares of the deviations of each value of the corresponding eigenvector from the average value of its neighbours. Then, small eigenvalues are associated to eigenvectors with small local deviation, representing smoother spatial trends. More specifically, it is possible to observe that the number of oscillations of eigenvectors grows as the eigenvalue size grows.

The log-density of an ICAR process is proportional to $-\sum_i^n \sum_{j \in \partial_i} (z_i - z_j)^2$, hence ICAR models tend to penalise local deviations as well. Intuitively, the Laplacian eigenvectors with which an ICAR field may confound are therefore the smoothest ones. This penalisation is less strong in PCAR, LCAR and BYM2 models, as random noise is allowed for.

Therefore, in the case $\boldsymbol{\xi} = \boldsymbol{I_n}$, with $n = n_z = n_y$, it is possible to express $\boldsymbol{X}$ as a linear combination of Laplacian eigenvectors, i.e. $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{B}$ for a $n \times p$ matrix $\boldsymbol{B}$. By convention, we choose to sort eigenvalues in descending order, and associated eigenvectors are ordered consequently. To filter out spatial trends from $\boldsymbol{X}$, let us label the rank deficiency of $\boldsymbol{L}$ as G, and consider a value K strictly smaller than $n - G$. Then, we can partition the eigenvector matrix as $\boldsymbol{V} = \left(\boldsymbol{V^{(NS)}}\, \boldsymbol{V^{(S)}}\, \boldsymbol{V^{(0)}}\right)$ where the non-spatial component $\boldsymbol{V^{(NS)}}$ includes the first $n - K - G$ eigenvectors, the spatial component $\boldsymbol{V}^{(S)}$ includes the eigenvectors indexed $]n - K - G, n - G]$, i.e. the K smallest-frequency non-constant ones, and $\boldsymbol{V^{(0)}}$ is constant across each connected component. Likewise, we can partition the rows of $\boldsymbol{B}$ into $\left(\boldsymbol{B^{(NS)}}^\top\, \boldsymbol{B^{(S)}}^\top\, \boldsymbol{B^{(0)}}^\top\right)^\top$. This partitioning allows us to rewrite the covariate matrix as

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

with $\boldsymbol{X^{(NS)}} = \boldsymbol{V^{(NS)}}\boldsymbol{B^{(NS)}}$, $\boldsymbol{X^{(S)}} = \boldsymbol{V^{(S)}}\boldsymbol{B^{(S)}}$ and $\boldsymbol{X^{(0)}} = \boldsymbol{V^{(0)}}\boldsymbol{B^{(0)}}$. In the regression model, the dependent variable is regressed on $\boldsymbol{X^{(NS)}} + \boldsymbol{X^{(0)}}$ or, as in Urdangarin et al. (2024), on $\boldsymbol{X^{(NS)}}$ alone. If the graph is connected, no difference arises between the effects of these two covariate matrices as they only differ by an additive constant.

If $\boldsymbol{\xi} \neq \boldsymbol{I}$, and the observation length of covariates is not the same as z, $\boldsymbol{X}$ can be mapped onto the spatial domain of z and this procedure can thus be applied on $\left(\boldsymbol{\xi}^\top\boldsymbol{\xi}\right)^{-1}\boldsymbol{\xi}^\top\boldsymbol{X}$. An application for this case is offered by Chapter 3. A simpler application in a single-level regression model is shown in Chapter 4.

1.6 The Integrated Nested Laplace Approximation

1.6.1 The Approximations Framework

Context: hierarchical regression

Throughout this thesis, we will extensively resort to the INLA in the framework of hierarchical linear regression models, that can be generalised as:

$$\begin{cases} E[y|\eta, \Psi] = g^{-1}(\eta) \\ \eta = \boldsymbol{A}\vartheta \\ \vartheta \sim \mathcal{N}\left(0, {\boldsymbol{Q}_{\boldsymbol{\Psi}}}^{-1}\right) \end{cases}$$

Where y is the dependent variable, η is the linear predictor, $g()$ is a link function, $\boldsymbol{A}$ is a known model matrix, ϑ is a latent Gaussian random field (GRF), and Ψ are model hyperparameters.[4]. As in the previous sections, if ϑ includes a multivariate latent field, it is intended as stacked columnwise rather than represented in matrix form.

This regression model is hierarchical as we are layering two arrays of unknown variables: the first layer ϑ of dimension n_ϑ, which enters the linear predictor and whose prior distribution is Gaussian and may depend on Ψ; and the second layer of unknown variables Ψ of dimension n_Ψ, from which either the likelihood or the prior of ϑ or both depend; sometimes the prior distribution of Ψ is referred to as the hyper-prior. In Gaussian regression, for instance, Ψ includes the variance-covariances parameters of $y|\vartheta$ and ϑ; in contrast, in Poisson regression, Ψ only influences the prior of ϑ. In this regression framework, we can assume that ϑ includes a set of covariate effects of size p and a latent GRF of size n, hence $\text{card}(\vartheta) = n + p$. What researchers are typically interested in are the posterior marginal distributions $\pi(\vartheta_i|y)$ and $\pi(\Psi_j|y)$ $\forall i \in [1, n+p]$ and $j \in [1, q]$ with q being the number of hyperparameters. Posteriors are the solutions of the integrals:

$$\begin{aligned} \pi(\vartheta_i|y) &= \int_\Psi \pi(\vartheta_i|y, \Psi)\pi(\Psi|y)d\Psi \\ \pi(\Psi_j|y) &= \int_{\Psi_{-j}} \pi(\Psi|y)\, d\Psi_{-j} \end{aligned} \tag{1.20}$$

Integration is feasible provided that q is small (say, less than 20). However, this operation can hardly be completed in closed form. This calls for the need of approximations, which we are going to discuss in the following paragraphs.

Despite our interest lies in linear models, we deem it worthy recalling that they are not the exclusive field of application of INLA. Nonlinear regression can

[4] In most of the INLA-related literature, the latent GRF is labelled as X or f, and θ denotes hyperparameters. Throughout this thesis we chose to derogate to the X, θ or f, θ notation in favour of the ϑ, Ψ notation in order to avoid confusion with explanatory variables being already labelled as X.

be achieved through the `inlabru` framework, implemented in the homonym R package (Bachl et al., 2019). `inlabru` implements a linearisation algorithm to approximate the predictor with a linear one. The steps of the algorithm can be summarised in this way: first, the predictor is Taylor approximated up to the linear term around a linearisation point; given the new linear predictor the posteriors are evaluated; the next linearisation point is set at the mode of the posterior of ϑ (Lindgren et al., 2024).

Approximating the joint Hyperparameters Posterior

Considering Equation 1.20, the first task is computing

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

The numerator is known *a priori*; $\pi(y)$ does not depend on the parameters of interest and can be treated here as a normalising constant. What is hard to find analytically, instead, is the full conditional $\pi(\vartheta|y,\Psi)$, as it is derived as:

$$\pi(\vartheta|y,\Psi) = \frac{\pi(\vartheta,y|\Psi)}{\int \pi(\vartheta,y|\Psi)d\vartheta}$$

and the integral can only be solved in closed formula for conjugate priors on ϑ. The intuition behind the INLA is, therefore, to replace the integrand function with its Gaussian approximation, which we denote as $\tilde{\pi}_{\text{G}}(\vartheta,y|\Psi)$

$$\pi(\vartheta|y,\Psi) \approx \tilde{\pi}_{\text{G}}(\vartheta|y,\Psi) = \frac{\pi(\vartheta,y|\Psi)}{\int \tilde{\pi}_{\text{G}}(\vartheta,y|\Psi)\,d\vartheta} \tag{1.21}$$

Here we describe more in detail how the Gaussian approximation works. For brevity, define $\gamma(\vartheta) := \ln \pi(\vartheta,y|\Psi)$. Then, a Taylor approximation truncated at the second order is applied:

$$\gamma(\vartheta) \approx \gamma(\mu_0) + \nabla\gamma(\mu_0)^\top(\vartheta-\mu_0) + \frac{1}{2}(\vartheta-\mu_0)^\top \boldsymbol{H}_{\boldsymbol{\gamma}}(\mu_0)(\vartheta-\mu_0)$$

Here, $\boldsymbol{H}_{\boldsymbol{f}}(x_0)$ denotes the Hessian matrix of a generic scalar-valued function $f(\cdot)$ evaluated at x_0 [5].

[5] This expression can be further simplified. To see this, first recall that

$$\boldsymbol{H}_{\boldsymbol{\gamma(\vartheta)}} = \boldsymbol{H}_{\boldsymbol{\ln \pi(\vartheta|\Psi)}} + \sum_{j=1}^{N} \boldsymbol{H}_{\boldsymbol{\ln \pi(y_j|\vartheta,\Psi)}}$$

The first addendum is the negative prior precision of ϑ. Regarding the second term, notice that instead of ϑ inference can be made on the first N elements of the vector $\theta := ((\eta+\epsilon)^\top, \vartheta^\top)^\top$, which is itself a GMRF (Rue et al., 2017). The term ϵ is a Gaussian error with arbitrarily small variance employed to make the distribution of the augmented predictor $\eta+\epsilon$ proper. We then have, $\forall\, i,r,c \in [1,N]$, that $\frac{\partial^2 \ln \pi(y_i|\theta_i,\Psi)}{\partial\theta_r \partial\theta_c}$ is nonzero only for $r=c=i$, hence under this parametrisation $\boldsymbol{H}_{\boldsymbol{\ln \pi(y_j|\vartheta,\Psi)}}$ is diagonal.

The point μ_0 is set as the maximum of $\gamma(\vartheta)$, denoted as $\mu_0(y, \Psi)$ to highlight its dependence on observed data and hyperparameters. This choice implies that $\nabla\gamma(\vartheta)$ equals an array of zeroes at $\mu_0(y, \Psi)$, and it is straightforward to observe that $\pi_{\text{G}}(\vartheta|y, \Psi)$ is proportional to a Gaussian density with mean $\mu_0(y, \Psi)$ and precision equal to the negative Hessian:

$$\tilde{\pi}_{\text{G}}(\vartheta, y|\Psi) = \gamma\left(\mu_0(y, \Psi)\right)\ e^{\frac{1}{2}[\vartheta - \mu_0(y,\Psi)]^\top \boldsymbol{H}_{\boldsymbol{\gamma}}(\mu_0(y,\Psi))[\vartheta - \mu_0(y,\Psi)]} \tag{1.22}$$

More in detail, considering the latent Gaussian model defined in section 1.6.1, let us define, in a preliminary stance, for $j = 1, \dots, N$:

$$\delta'_j := \frac{\partial \ln \pi(y_j|\eta_j, \Psi)}{\partial \eta_j} \quad \text{and} \quad \delta''_j := \frac{\partial^2 \ln \pi(y_j|\eta_j, \Psi)}{\partial \eta_j^2}$$

and stack these values in the arrays $\Delta_1 = (\delta'_1, \delta'_2, \dots, \delta'_N)^\top$ and $\boldsymbol{\Delta_2} = \text{diag}\,(-\delta''_1, -\delta''_2, \dots, -\delta''_N)$. It follows that

$$\begin{cases} \nabla\gamma(\vartheta) = \boldsymbol{A}^\top \Delta_1 - \boldsymbol{Q_\Psi}\vartheta \\ \boldsymbol{H}_{\gamma(\vartheta)} = -\boldsymbol{A}^\top \boldsymbol{\Delta_2} \boldsymbol{A} - \boldsymbol{Q_\Psi} \end{cases}$$

The mode $\mu_0\left(y, \Psi\right)$, i.e. the zero of the gradient in the formula above, can be computed iteratively.
The remaining problem in the formula for $\pi(\Psi|y)$ is that it still depends on ϑ. The joint posterior distribution of hyperparameters is thus approximated via Laplace approximation, locating ϑ at its mode conditional on Ψ and y:

$$\pi(\Psi|y) \approx \tilde{\pi}_{\text{LA}}(\Psi|y) \propto \left.\frac{\pi(y|\eta, \Psi)\pi(\vartheta|\Psi)\pi(\Psi)}{\pi_{\text{G}}(\vartheta|y, \Psi)}\right|_{\vartheta=\mu_0(\Psi,y)} \tag{1.23}$$

Approximating the joint Gaussian field Posterior

To approximate the marginals $\pi(\vartheta_i|y, \Psi)\forall i$, several solutions are possible. We provide an overview of the four methods developed so far, distinguishing between the original methods proposed in the seminal paper of Rue et al. (2009), and the latest alternative method developed by Van Niekerk and Rue (2024).

The original framework

The original INLA framework relied on a set of three alternative strategies, following a trade-off in accuracy and computational cost. We will not focus in depth on these strategies as in this thesis we mainly rely on the novel approach.

Gaussian approximation A rough solution would be using the Gaussian approximation in equation 1.21. Even though this is a computationally cheap operation, it may suffer low accuracy, as it ignores all information about skewness or higher moments in the posterior. The following two alternatives have been indeed developed to improve accuracy in approximate inference.

Full Laplace approximation This strategy 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 $\pi(\vartheta_{-i}|\vartheta_i,y,\Psi)$. This is the most rigorous choice but is computationally demanding.

Simplified Laplace approximation The third 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 $\ln\pi(\vartheta,y|\Psi)$ at the third order term, while still locating the approximation at the mode of ϑ. This allows to fit a Skew-Normal density to the full conditional.

In the original framework, inference was originally made, rather than on ϑ, on η, by leveraging the fact that if ϑ is a GMRF, so is η (though with a singular precision). For more details on inference made on η, a valid reference is Rue et al. (2017).

The Variational Bayes method

A recent strand of literature introduced a different approach to approximating the joint conditional posterior of the latent Gaussian field. While the former methods consisted of three proposed functional forms of this approximation, the new strategy takes a further step in the field of optimisation. In this context, optimisation must be read as proposing an approximation as close as possible to the true joint posterior $\pi(\vartheta|y,\Psi)$. To this aim, the simplest Gaussian approximation $\pi_{\text{G}}(\theta|y,\Psi)$ remains *per se* a rigorous proposal as a functional form, while the elements on which attention is drawn are now its moments.

Intuitively, approximating the true but unknown density $\pi(\vartheta|y,\Psi)$ with a known density $\pi_{\text{G}}(\vartheta|y,\Psi)$ implies a loss in information. As argued by Zellner (1988), statistical inference consists, in its very essence, of the processing of two prior assumptions, the likelihood and parameter prior density, into two outputs, the marginal likelihood and the posterior distribution of parameters. The Bayes rule itself implies that no loss of information occurs when the output quantities are computed exactly. But prior conjugacy is a fortunate case, and if it held in general, the whole scientific matter of approximate inference would even lose its reason to be.

Since at this stage all inferential interest lies in ϑ, we will omit the conditioning on Ψ throughout the remainder of this section. In a statistical model yielding the posterior $\pi^*(\vartheta|y)$, the priors or *pre-data* inputs (Zellner, 1988) are given by

$$\Pi_{\text{IN}} = \int_{-\infty}^{+\infty} \left[\ln\pi(\theta) + \ln\pi(y|\theta)\right]\pi^*(\vartheta|y)d\vartheta$$

while the model output is:

$$\Pi_{\text{OUT}} = \int_{-\infty}^{+\infty} \left[\ln\pi(y) + \ln\pi^*(\vartheta|y)\right]\pi^*(\vartheta|y)d\vartheta = \ln\pi(y) + \int_{-\infty}^{+\infty} \ln\pi^*(\vartheta|y)\pi^*(\vartheta|y)d\vartheta$$

Where $\pi(y)$ is the marginal likelihood, obtained by integrating out ϑ from $\pi(y, \vartheta)$. The metric Zellner (1988) proposed to measure information loss through the information processing scheme is the difference between output and input information. If the statistical model allowed to compute $\pi^*(\vartheta|y) \equiv \pi(\vartheta|y)$, the difference between output and input information would be zero. Otherwise, a loss in information would arise, and the optimisation problem shaping this novel approximation scheme in INLA consists precisely in minimising such an information loss. In particular, if the posterior of ϑ is approximated via a known functional form like the Gaussian approximation, $\pi_{\text{VB}}(\vartheta|y)$, where the VB subscript stands for Variational Bayes, the loss in information is measured as:

$$\Pi_{\text{OUT}} - \Pi_{\text{IN}} = -\mathbb{E}_{\pi_{\text{VB}}(\vartheta|y)}\left[\ln \pi(y|\vartheta)\right] + \int_{-\infty}^{+\infty} \ln \frac{\pi_{\text{VB}}(\vartheta|y)}{\pi(\vartheta)} \pi_{\text{VB}}(\vartheta|y) d\vartheta + C =$$

$$= -\mathbb{E}_{\pi_{\text{VB}}(\vartheta|y)}\left[\ln \pi(y|\vartheta)\right] + \text{KLD}\left(\pi_{\text{VB}}(\vartheta|y)||\pi(\vartheta)\right) + C =$$

Where $C = \ln \pi(y)$ does not depend on ϑ and is thus of no interest in this optimisation scheme.

Specifically, Van Niekerk and Rue (2024) propose to employ

$$\pi_{\text{VB}}(\vartheta|y, \Psi) := \varphi\left(\vartheta|\mu_1, (-\boldsymbol{H}_{\boldsymbol{\gamma}})^{-1}\right) \tag{1.24}$$

where $-\boldsymbol{H}_{\boldsymbol{\gamma}}$ is the precision matrix of the Gaussian approximation $\pi_G(\vartheta|y)$, as seen in section 1.6.1, and

$$\mu_1 := \mu_0(y, \Psi) + \Sigma_I \lambda$$

where Σ_I is a projection matrix determined from a subset of the columns of the covariance $(-\boldsymbol{H}_{\boldsymbol{\gamma}})^{-1}$ and λ is the correction factor whose only nonzero elements are those belonging to the index set I. Then, λ is determined as the minimiser to the following objective function:

$$\begin{aligned} \lambda := \arg\min_{\lambda} \Bigg\{ & KLD\left(\pi_{VB}(\vartheta|y, \Psi)||\pi(\vartheta|\Psi)\right) + \\ & - \int_{-\infty}^{+\infty} \sum_{i=1}^{N} \ln \pi(y_i|\eta_i, \Psi) \pi_{VB}(\vartheta|y, \Psi)\, d\vartheta \Bigg\} \end{aligned} \tag{1.25}$$

To improve computation efficiency, Van Niekerk and Rue (2024) suggest to choose I as a small subset of the column indexes of the precision; for this reason, this methodology is referred to as low-rank correction. Since λ is multiplied by a square matrix, it is straightforward to figure out that any nonzero element in λ will be propagated across multiple values of ϑ.

The integral can be computed via Gauss-Hermite quadrature, and Taylor approximated at the second order, as shown by Van Niekerk and Rue (2024). Regarding the KLD, we have two Gaussian densities with the same mean and

different precision: it is then available in closed formula using the result in Appendix .4.

Once the VB correction terms are computed, deriving the marginal conditionals $\pi_{\text{VB}}(\vartheta_i|y,\Psi)$ from the joint distribution as defined in equation 1.24 is straightforward, as computing the marginal distributions of a multivariate Gaussian only requires the diagonal entries of the inverse precision matrix.

This methodology appears promising, as it attempts to face the essential issue of approximate inference, i.e. the very fact that the output information is an approximate function. However, it bears two limits. First, only the mean of the posterior is corrected. Second, the posterior functional form is assumed rigidly. To this aim, some refinements are being carried forward (Dutta et al., 2025). On the first issue, a Variational Bayes additive correction has been proposed for the precision matrix, and it is seen to improve the accuracy of the posterior approximation for some likelihoods, such as the t-Student and the generalised Pareto in small samples. For the Poisson likelihood with a moderate-size sample (300 observations), instead, authors observed that the variance correction did not yield substantive improvements.

Additionally, imposing the Gaussian functional form to the posterior forces the third moment to zero. We have seen that the "old" INLA framework allowed to fit a Skew-Normal distribution to the joint posterior. Dutta et al. (2025) propose to integrate the Variational Bayes approach with a third-order correction. The full joint then becomes a product of Normal and Skew-Normal marginals and a Gaussian copula, taking into account the posterior correlations among the elements of ϑ. Partitioning the posterior precision and variance-covariance matrices allows to fit skew-Normal marginals with a scalable computational cost.

Though these two developments may improve the accuracy of INLA in several scenarios, e.g., with small sample sizes or with heavy-tailed likelihoods, they lie beyond the scope of this work, and we limit the analyses in forthcoming chapters to the VB correction to the mean.

1.6.2 The Integration Scheme

Once the integrand functions in equation 1.20 are defined, integration is carried out numerically. This requires to compute $\tilde{\pi}(\Psi|y)$ across a given set of points, which in this case is a multidimensional grid. In particular, this grid exploration will not regard Ψ itself, but a transformation defined over $\mathbb{R}$, whereas many hyperparameters usually range on a subset of $\mathbb{R}$ (think of variance and precision parameters which are defined on the nonnegative axis, or correlations which range -1 to 1).

The fundamental assumption to make is that $\tilde{\pi}(\Psi|y)$ is unimodal. In a first stance, unimodality implies that the hyperparameter posterior tends to zero as Ψ departs from its mode. Thus, numerical integration can be carried out within a neighbourhood of hyperparameter mode and not on its entire domain. In a second stance, under this assumption only one value for the local maximum of the hyperparameter posterior needs to be found.

The output of the previous steps of the INLA procedure are functions. Here, we are looking for for a numerical array, i.e. the posterior mode of Ψ. This is clearly an optimisation problem, which requires to explore the hyperparameter space with an iterative scheme, in which at each iteration t, the directions of $\Psi^{(t)}$ are typically proportional either to $\nabla\tilde{\pi}\left(\Psi^{(t)}|y\right)$ (gradient descent) or to $\boldsymbol{H}^{-1}_{\tilde{\pi}(\Psi^{(t)}|y)}\nabla\tilde{\pi}(\Psi^{(t)}|y)$ (Newton-Raphson). The most intuitive way to define the numerical gradient is using a canonical basis, i.e considering that $\nabla\tilde{\pi}(\Psi|y) = \sum_{j=1}^{n_\Psi}\frac{\partial\tilde{\pi}(\Psi|y)}{\partial\Psi_j}e_j$ where n_Ψ is the number of hyperparameters and e_j is the j-th vector of the canonical basis in n_Ψ dimensions. Due to the intractable form of the hyperparameter posterior, derivatives must be computed numerically, with either forward or central differences. Fattah et al. (2022) proposed to define directions with a non-canonical basis which, conceptually, "learns" from the previous iterations of the algorithm. In this framework, the gradient takes the form:

$$\tilde{\nabla}_{\tilde{\boldsymbol{d}}}\tilde{\pi}\left(\Psi^{(t)}|y\right) = \tilde{\boldsymbol{G}}^{(t)^{-1^\top}}\tilde{\nabla}_{\boldsymbol{e}}\tilde{\pi}\left(\Psi^{(t)} + \tilde{\boldsymbol{G}}^{(\boldsymbol{t})}\delta|y\right)|_{\delta=0}$$

Where notation $\tilde{\nabla}_{\boldsymbol{v}}f(x)$ stands for the approximate gradient of the function $f(x)$ using directions $\boldsymbol{v}$; $\boldsymbol{e}$ denotes the canonical basis. The matrix $\tilde{\boldsymbol{G}}^{(t)}$ whose columns form the orthonormal basis $\tilde{\boldsymbol{d}}$ is constructed by modified Gram-Schmidt orthogonalisation[6] of the matrix $\boldsymbol{G}^{(\boldsymbol{t})} := \left(\Psi^{(t)} - \Psi^{(t-1)}, d_1^{(t-1)}, \cdots, d_{n_\Psi-1}^{(t-1)}\right)$, where columns $\left(d_1^{(t-1)}, \cdots, d_{n_\Psi-1}^{(t-1)}\right)$ depend on the previous $n_\Psi - 1$ iterations. Taking second derivatives clearly allows to approximate the Hessian matrix as well.

Authors labelled this methodology "smart gradient approach" and assess it to be more accurate than the approximation built on the canonical basis (i.e., in the case of $\tilde{\boldsymbol{G}} = \boldsymbol{I}_{\boldsymbol{n_\Psi}}$).

Now, expliciting equation 1.23 and omitting the argument from $\mu_0(y,\Psi)$ for brevity, the Laplace approximation to the joint hyperparameter posterior reads as:

$$\tilde{\pi}_{\text{LA}}\left(\Psi|y\right) \propto \pi(\Psi) \times \prod_{j=1}^{n_y}\pi\left(y_j|\Psi,\eta_j = \boldsymbol{A}\mu_0\right) \times e^{-\frac{1}{2}\mu_0^\top\boldsymbol{Q}_\Psi\mu_0} \times \frac{\sqrt{|\boldsymbol{Q}_\Psi|}}{\sqrt{|-\boldsymbol{H}_{\boldsymbol{\gamma}}|}}$$

Where $\boldsymbol{H}_{\boldsymbol{\gamma}} = -\boldsymbol{A}^\top\boldsymbol{\Delta_2}\boldsymbol{A} - \boldsymbol{Q}_{\boldsymbol{\Psi}}$ is the precision of $\vartheta|y,\Psi$. To compute the gradient with finite differences, the posterior must be evaluated $2n_\Psi$ times, and to compute the Hessian, which is required in Newton-Raphson schemes, ${n_\Psi}^2+2$ iterations are required. Since the gradient (and the Hessian if required) must be evaluated, in turn, at each iteration of the optimisation algorithm, the computational cost of $\tilde{\pi}_{\text{LA}}(\Psi|y)$ is crucial for the efficiency of INLA. In the following, we discuss the asymptotic cost of each factor. In doing so, we can

[6]Even though the original paper uses the notation $\tilde{\boldsymbol{G}}^{-1^\top}$, it should be noticed that since $\tilde{\boldsymbol{G}}$ is constructed on an orthonormal basis $\tilde{\boldsymbol{G}}^{-1} \equiv \tilde{\boldsymbol{G}}^\top$, so this notation could be simplified to $\tilde{\boldsymbol{G}}^{-1^\top} \equiv \tilde{\boldsymbol{G}}$.

treat $1/\pi(y)$ as a proportionality constant, in that it does not depend on Ψ. We define n_ϑ as the size of ϑ and n_y as the size of y.

- The cost of the joint prior $\pi(\Psi)$ strictly depends on how it is defined. Except for the infrequent case of distributions requiring e.g., the eigendecomposition or the inversion of dense matrices such as e.g. in non-irreducible KLD structures (as mentioned in Section **??**), this is a negligible cost when the size of Ψ is in the order of 10^1.

- The likelihood $\prod_{j=1}^{n_y} \pi\left(y_j|\Psi, \eta_j = \boldsymbol{A}\mu_0\right)$ is the product of n_y scalars. It can be noticed that the cost of each $\pi(y_j|\eta_j, \Psi)$ is driven by $\eta_j = \boldsymbol{A_j}.\vartheta$. In hierarchical regression, only $\Theta(1)$ elements in each row of $\boldsymbol{A}$ are nonzero, and we do not actually need to perform as many operations as the values of ϑ. This is possible by treating $\boldsymbol{A}$ as a sparse matrix, i.e., only storing in the internal memory the position of its nonzero entries. Due to sparse matrix algebra, the cost of computing the likelihood is $\Theta(n_y)$, i.e., the procedure is scalable.

- $\mu_0^\top \boldsymbol{Q_\Psi} \mu_0$. Two products occur. The first is between a vector and a square matrix. Again, the sparsity pattern of $\boldsymbol{Q_\Psi}$ plays a crucial role. If $\boldsymbol{Q_\Psi}$ has $\Theta(n_\vartheta)$ entries, i.e., it is a sparse matrix, the number of operations is in $\Theta(n_\vartheta)$; for dense matrices, the cost is in $\mathcal{O}(n_\vartheta^2)$, which is prohibitive. Then, the cross-product of the two vectors has $\Theta(n_\vartheta)$ operations. This operation is scalable as long as $\boldsymbol{Q}_\Psi$ is sparse.

- The two determinants are usually the most complex operations in hierarchical regression. In a context where $\boldsymbol{A} = (\boldsymbol{X}, \boldsymbol{\xi})$ where $\boldsymbol{X}$ is the matrix of explanatory variables and $\boldsymbol{\xi}$ is the model matrix for a latent Gaussian random field, the matrix $\boldsymbol{A}^\top \boldsymbol{\Delta_2} \boldsymbol{A}$ is block-diagonal, with a dense upper-left block of order equal to the column of $\boldsymbol{X}$, and a sparse bottom-right block (for instance, it is diagonal in the model used in chapter 3). So, if the number of explanatory variables has a dimensional order lower than the size of the latent Gaussian field, the computational cost of both determinants depends on the sparsity pattern of $\boldsymbol{Q_\Psi}$. The determinant of a matrix can be computed as the square product of the diagonal of its Cholesky factor. It turns out that Cholesky factorisation for sparse precision matrices implies significant computational advantages, considering that in the most general case the determinant of a matrix has a cubic cost. As pointed out by Rue and Held (2005) and Rue et al. (2009), the factorisation has a cost in the order of n_ϑ for time series (one-dimensional dependency structure), $n_\vartheta^{3/2}$ for spatial models (two dimensions), and n_ϑ^2 for three-dimensional structures. With sparse precision matrices, this procedure is efficient, but not scalable, except for one-dimensional structures.

As can be seen, the sparsity pattern in $\boldsymbol{Q_\Psi}$ is a crucial requirement for the efficiency of INLA. This motivates the emphasis we have given to this matter in section 1.2.

Once the hyperparameter posterior mode is located, numerical integration can be carried out to approximate the solution to equation 1.20 as:

$$\begin{cases} \tilde{\pi}_{\text{G}}\left(\vartheta_i|y\right) = \sum_{k=1}^{K} \tilde{\pi}\left(\vartheta_i|y,\Psi^{(k)}\right) \tilde{\pi}_{\text{LA}}\left(\Psi^{(k)}|y\right) \Delta\Psi^{(k)} \\ \tilde{\pi}_{\text{LA}}\left(\Psi_j|y\right) = \sum_{k=1}^{K} \tilde{\pi}_{\text{LA}}\left(\Psi^{(k)}|y\right) \Delta\Psi^{(k)}_{-j} \end{cases}$$

For all i and all j. $\Delta\Psi^{(k)}_{-j}$ are the integration weights and $\Psi^{(k)}$ are the integration points chosen along the parameter space of Ψ. $\tilde{\pi}\left(\vartheta_i|y,\Psi^{(k)}\right)$ denotes any of the four approximations introduced in section 1.6.1. In the case of Gaussian approximations to the full posterior, marginalising from a joint Gaussian only requires the diagonal of the inverse Hessian, which can be computed efficiently through sparse Cholesky factorisation, and is only required once for all ϑ_i. The cost of this operation has thus the same order as $\tilde{\pi}_{\text{LA}}(\Psi)$, implying that the evaluation of the "integrand" functions in the formulae above is comparable to a single iteration of the optimisation algorithm to locate the mode of hyperparameters. This highlights that the number K of integration points is a crucial matter. While the number of iterations of the optimisation algorithm is not always easy to control, the number of integration points depends on the user's choice among three alternative integration schemes proposed by Rue et al. (2009). We sort them in order of complexity.

Empirical Bayes The cheapest among possible solutions relies on the so-called empirical Bayes method, and consists in assuming that the hyperparameter posterior is sharply peaked around its mode, hence the integrals in equation 1.20 are replaced by the integrand functions evaluated at the mode of Ψ. In practice, the integration point is only one. This rather extreme solution rules out all the uncertainty on the joint distribution of hyperparameters but has a negligible computation cost.

Central Composite Design A more accurate integration scheme consists in employing $\mathcal{O}\left(2^{n_\Psi}\right)$ integration points, where n_Ψ is the number of hyperparameters. These points are determined through the composite central design method and lie on the surface of the n_Ψ-dimensional sphere of radius $\sqrt{n_\Psi}$. Specifically, the set of points includes, net of a subsequent scaling, some vertices of the hypercube inscribed in the sphere, i.e., an embedded fractional factorial design, and $2n_\Psi + 1$ star points, i.e., vectors lying on axes having a form $\sqrt{n_\Psi}\, e_j$ where e_j is the j-th basis vector in the n_Ψ-dimensional space. In full factorial design, the number of points would be 2^{n_Ψ}; the term *fractional* comes from the feature of reducing the exponent for $n_\Psi \geq 5$. For instance, for $n_\Psi \in \{9, 10, 11\}$ the fractional factorial design includes only 2^7 points, to which $2n_\Psi + 1$ points have to be added from the intersection of the sphere with the axes. Details on how fractional design simplifies the number of points are thoroughly explained in Sanchez and Sanchez (2005).

Multidimensional grid The most rigorous approach consists in locating integration point along a regular n_Ψ-dimensional grid. If m points are explored in

each dimension, a total of m^{n_Ψ} evaluations of the integrand functions is needed. In practice, this strategy may imply a computational cost comparable to the optimisation algorithm when the number of hyperparameters is moderate. In the current R package version, this option can only be activated with 2 or less hyperparameters.

1.6.3 The `R-INLA` package

The complex architecture of the INLA methodology finds its practical implementation in the `R-INLA` library. While `R` provides the user interface and the wrapper functions, the core of calculations is carried out through a system of inner `.dll` libraries. Due to the size of such libraries, the package is not available on CRAN, but only in the dedicated repository https://inla.r-inla-download.org/R/.

The main `R-INLA` functionality the user relies on is to run a regression model. This is done with the `INLA::inla()` function, whose first argument is the predictor formula and follows the standard syntax of `stats::lm()` and `stats::glm()`. As we have seen in section 1.6.2, the joint posterior of the first layer of parameters can be retrieved analytically if the likelihood comes from a finite set of functions. For this reason, it is preferable to choose it from the models already implemented within the software. In the package version `2025.06.22-1`, the total number of available likelihoods is 105. They can be consulted using the command `INLA::inla.list.models("likelihood")`. Each likelihood model has a default hyperprior associated, but the user can specify a different one; for instance, the Gaussian likelihood has, by default, a Gamma$(1, 10^{-5})$ default prior.

If the regression model is hierarchical, the prior model for the latent Gaussian field (z in our notation) can be passed to the function through the `INLA::f()` argument. The first argument of this function is an array of indexes of the same length as y mapping each element z_i with $i = 1, \ldots, n_z$ to η_j with $j = 1, \ldots, n_y$. Another crucial argument to this function is `model`, specifying the prior model for z. Some models are automatically implemented in `R-INLA` and the user simply needs to call them by name; the full list can be scanned using `INLA::inla.list.models("latent")`. As for the likelihood specification, though each model has its default hyperpriors, the user is free to deviate from the default.

Differently from the likelihood function, since z is treated as a Gaussian variable, the software is required no extra effort if the user specifies a customised model manually. This can be done in two ways. For a precise subset of simple precision structures, a user-friendly solution is specifying the model as one of `"generic0"`, `"generic1"`, `"generic2"`, `"generic3"`. Otherwise, more generic latent Gaussian models can be defined with the function `INLA::inla.rgeneric.define()`, which requires the user to provide the precision matrix, the graph structure, the prior distribution of hyperparameters on which z depends (e.g. for the PCAR model, the autocorrelation ρ and either the scale or precision parameter), and the initial values for the optimisation algorithm. Multivariate models must be stacked columnwise. Chapter 11 of

Gómez-Rubio (2020) provides a step-by-step explanation of how models must be defined internally. The high degree of operational flexibility allowed for by the `inla.rgeneric.define()` has played a crucial role in establishing the popularity of `R-INLA`. In the context of areal modelling, the univariate ICAR, LCAR, and BYM2 are already implemented in the package and can be fitted by choosing as model `"besag", "besagproper2"`, and `bym2` respectively; notably, the LCAR model features, by default, a Normal prior on logitλ. The PCAR is not implemented, but chapter 11 of Gómez-Rubio (2020) uses it as an instructional example of how to manually define a model for Gaussian latent fields.

Regarding multivariate models, due to the high number of parameters and the more complex hyperprior combinations, no ready-made model is implemented, but some helper `R` packages have been developed to fit them automatically.

An early contibution is the `INLAMSM` package (Palmí-Perales et al., 2021), implementing multivariate ICAR, PCAR, and M-model PCAR, with a weakly informative Wishart prior on the precision in the former two cases and a Wishart prior on the scale in the latter, and a Uniform prior on the PCAR autocorrelation.

The package `bigDM` (Vicente et al., 2023), which is available on CRAN (`https://CRAN.R-project.org/package=bigDM`), instead, implements ICAR, PCAR, and LCAR M-models using a Wishart prior on the scale parameter through the Bartlett decomposition (Kabe, 1964). The Bartlett decomposition consists in decomposing the variance-covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$, where $\boldsymbol{B}$ is a triangular matrix whose squared diagonal entries $b_{ii}^2 \quad \forall i \in \{1, \cdots, k\}$ are assigned a chi-squared distribution and whose off-diagonal elements $b_{ij} \forall i \in \{2, \cdots, k\}, j \in \{1, \cdots, i\}$, are assigned a standard Normal prior. This parametrisation is particularly useful in the `R-INLA` context when exploring the parameter space of Ψ to locate its mode, as for any combination of values in the $\boldsymbol{B}$ matrix, as long as it is full-rank, $\boldsymbol{\Sigma}$ will be positive-definite, as it is a quadratic form. If, instead, $\boldsymbol{\Sigma}$ results not to be positive definite, the system crashes due to calculation failures to evaluate the joint posterior. Also, differently from the Cholesky decomposition, the Bartlett decomposition is invariant to the ordering of the entries of the CAR fields. Lastly, as for `INLAMSM`, the PCAR autocorrelation and the LCAR precision mixing are assigned a Uniform prior.

INLA can internally compute some information criteria and preserve the hyperparameter configurations at all integration points, which can be useful but may require a large storage space. The user can ask the system what to compute in addition to posteriors with argument `control.compute()`. Options regarding the optimisation algorithm, the numerical integration scheme and other more technical issues can be specified with the argument `control.inla()`.

Except for the case of one-dimensional GMRFs, we have seen that the optimisation algorithm is not scalable. This fact may become problematic in the case of large latent structures. A solution to this more-than-linear cost problem is in parallelising computations across multiple cores of the local machine of the user (Gaedke-Merzhäuser et al., 2023; Van Niekerk et al., 2023). This higher efficiency, however, comes at the cost of losing exact replicability of re-

sults, which stems from the algorithm being run on several cores. To achieve results that are perfectly replicable using the same operating system and the same package version, it is necessary to force the number of cores to 1. This can be done with argument `num.threads = 1`, and setting `control.inla = list(..., internal.opt = FALSE)` (Wang et al., 2018).

The four approaches described in Section 1.6.1 to approximate $\pi(\vartheta_i|y)$ 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. Using the new package setup also implies that the hyperparameters posterior is integrated with points defined with the regular grid for $n_\Psi \leq 2$ and with the CCD otherwise.

1.7 Penalised Complexity Priors

So far, in section 1.6 most attention in prior distributions has been devoted to $\pi(\vartheta \mid \Psi)$. An appealing feature of Bayesian inference we have not focused on yet is the flexibility in setting hyperpriors, namely $\pi(\Psi)$. Built-in models in `R-INLA`, both for likelihood and for latent Gaussian random fields, feature some default hyperpriors (for the precision in the univariate ICAR, for instance, under version `2025.06.22-1` the default prior is a weekly informative Gamma with scale parameter = 1 and rate parameter = 5×10^{-5}); however, prior beliefs should always be calibrated in the specific framework analysis, and the amount of subjective belief to be expressed by the prior is a delicate issue, requiring user judgement.

Since INLA is inherently suitable for complex models with a large number of parameters, it contributed as a field of application to a class of prior distributions designed for hierarchical models, namely Penalised Complexity (PC) priors, introduced by Simpson et al. (2017) to penalise *a priori* the departure of a hierarchical model from a somewhat simpler one in an explicit and directly interpretable way. The functional form of this distribution family is model-dependent, and owes to the principle of parsimony; quoting authors, *"simpler model formulations should be preferred until there is enough support for a more complex model"*.

Consider two statistical models for a random variable y, say $\mathcal{M}_1(y|\psi)$ which allows the parameter ψ to vary within a given parameter space, and $\mathcal{M}_0(y|\psi = \psi_0)$, a simpler one, in which the parameter ψ is fixed to a deterministic value.

As an illustrative example, let us consider the case in section 3.3 of Simpson et al. (2017). If y is defined on a discrete spatial domain and is characterised by a smooth spatial trend, one could regress y either on a set of observed covariates (say, using a generalised linear model), or on the union of a set of observed

covariates *and* a latent ICAR field, with prior distribution $z \mid \mathcal{N}\left(0, \sigma^2 \boldsymbol{L}^+\right)$. The latter model is more flexible, including some regression coefficients (as the simple one), the latent ICAR field z, and the unknown scale parameter of the ICAR. σ^2. The simpler one, instead, is equivalent to a model in which the latent Gaussian field has mean and variance exactly zero. In this case, a PC-prior can be assigned to σ^2 by penalising the departure of $\sigma^2 > 0$ from $\sigma^2 = 0$.

Formalising the concepts of *departure* and *penalisation* is a normative choice. Simpson et al. (2017) propose to measure the departure between $\mathcal{M}_1(y|\psi)$ and $\mathcal{M}_0(y|\psi = \psi_0)$, which can be interpreted as a measure of complexity indeed, through the distance function $D(\psi) := \sqrt{2\text{KLD}(\mathcal{M}_1||\mathcal{M}_0)}$. On the other hand, this measure of complexity is penalised by assigning it a density function decaying at a constant rate. This can be done through an exponential prior, say:

$$\pi(D(\psi)) = \gamma e^{-\gamma D(\psi)}$$

for a rate parameter γ. For univariate parameters, the PC prior is then computed by a simple change of variable if the square root of the KLD is a monotonic function of ψ:

$$\pi(\psi) = \gamma e^{-\gamma D(\psi)} \frac{1}{D(\psi)} \text{abs}\left(\frac{dD(\psi)}{d\psi}\right) \tag{1.26}$$

Choosing an exponential prior for the distance measure $D(\psi)$ allows the PC-prior to be particularly simple to interpret. From the properties of the exponential distributions, we know that for a given boundary value U and left-tail probability α, $\text{Prob}\{D(\psi) \leq U\} = 1 - e^{-\gamma U} = \alpha$, hence it is sufficient to fix α and U to have $\gamma = \frac{1}{U} \ln\left(\frac{1}{1-\alpha}\right)$. Therefore, defining a PC prior on ψ ultimately implies to assume a precise probabilistic statement, namely that with a probability α, ψ does not exceed a given boundary value of U.

In `R-INLA`, some model-specific functions are implemented to compute the PC-prior for some specific models, such as the precision parameter for univariate Gaussian models (function `INLA::inla.pc.dprec`), or the variance mixing parameter in the univariate BYM2. Notably, the default prior for the BYM2 is a joint PC-prior on the precision and mixing parameters (function `INLA:::inla.pc.bym.phi`).

As we have seen, this class of distributions represent a flexible, interpretable and principle-based way to define univariate priors. A drawback of this approach, instead, regards multivariate parameters, as to derive a PC-prior we need to apply a change of variable from the KLD (which is always a scalar) to the hyperparameter itself. If the hyperparameter has dimension $n_\Psi > 1$, the change of variable from $\mathbb{R}^1$ to $\mathbb{R}^{n_\Psi}$ becomes challenging and requires specific solutions to circumvent the multivariate change of variable problem, unless the distance function $D(\Psi)$ has either a linear or quadratic form (Simpson et al., 2017). Some promising solutions include decomposing the joint prior into a series of univariate conditionals through the Hammersley-Clifford theorem (Battagliese, 2020), modelling the dependence between hyperparameters with copulas (ibidem), or defining a graphical dependence structure among hyperparameters

(Freni-Sterrantino et al., 2025). In this thesis, we also bring our contribution to extending PC-priors to multivariate parameters using a stepwise-defined prior, based on the simple chain rule decomposition of the joint prior, as discussed in section **??**. (what will be the next section 4.4)

.1 Joint distribution of multivariate PCAR

To compute the joint density $\pi(z|\boldsymbol{\Sigma})$ we rely on the Brook's Lemma (Hammersley and Clifford, 1971; Besag, 1974):

$$\pi(z) = \prod_{i=1}^{n} \frac{\pi\left(z_i|x_1, x_2, \ldots, x_{i-1},\ z_{i+1}, \ldots, z_n\right)}{\pi\left(x_i|x_1, x_2, \ldots, x_{i-1},\ z_{i+1}, \ldots, z_n\right)} \pi(x) \tag{27}$$

Where x is a set of known variables with the same cardinality as z and satisfying the positivity condition, namely if $\pi(x_i) > 0 \forall i$, then $\pi(x) > 0$ as well (Besag, 1974). If the positivity condition holds, $\pi(x)$ is a constant term which can be treated a a normalising factor. With no loss of generality, defining $x_j = 0 \quad \forall\, i \in \{1, \ldots, n\}$ and introducing the row-wise vectorising operator vec_r, which stacks row-wise the elements of a $n \times k$ matrix into a $kn \times 1$ vector, with some algebra (see also Mardia, 1988; Banerjee et al., 2014; Gelfand et al., 1992) it is possible to see that :

$$\pi(z|\boldsymbol{\Sigma}^{-1}) \propto e^{-\frac{1}{2}(\text{vec}_r\, z)^\top \left[(\boldsymbol{D}-\rho\boldsymbol{W})\otimes\boldsymbol{\Sigma}^{-1}\right] \text{vec}_r z} \tag{28}$$

In practical applications such as the regression models in sections 1.5 or 1.6.1 it is convenient to rearrange z column-wise through the vec operator Palmí-Perales et al. (2021), such that

$$\text{vec}\, z = (z_{11}, \ldots, z_{n1}, z_{12}, \ldots, z_{n2}, \ldots, z_{1k}, \ldots, z_{nk})^\top$$

To do so, we are actually using an orthogonal permutation matrix $\mathcal{P}$, such that

$$\text{vec}\, z = \mathcal{P}\text{vec}_r z$$

To define $\mathcal{P}$, consider the i-th basis vector of $\mathbb{R}^p$, e_i, the i-th row of $\boldsymbol{I_n}$, whose i-th element is 1 and all other elements are zero, $\forall i \in [1, n]$. The permutation matrix is:

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

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

$$\mathcal{P}\left[(\boldsymbol{D} - \rho\boldsymbol{W}) \otimes \boldsymbol{\Sigma}^{-1}\right]\mathcal{P}^\top = \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{D} - \rho\boldsymbol{W})$$

Which can be proved with some algebra. Hence, in total equivalence with equation 28, we obtain equation 1.6

.2 Sparse parametrisation for M-model BYM2 extension

Here we generalise the sparse-precision parametrisation of Riebler et al. (2016). As mentioned in section 1.4.4, we use a unique scale parameter $\boldsymbol{\Sigma}$ for the two components, and consider $\boldsymbol{M}$ as a square and invertible matrix of order k such that $\boldsymbol{M}^\top \boldsymbol{M} = \boldsymbol{\Sigma}$.

Consider the model in equation 1.11. Keeping up with the approach of Riebler et al. (2016), and defining for the sake of brevity $\bar{\Phi} := \boldsymbol{I}_k - \Phi$, we know that:

$$\operatorname{vec} \mathbb{E}[\boldsymbol{z}|U, \Phi, \boldsymbol{\Sigma}] = \operatorname{vec}(\boldsymbol{U}\Phi^{\frac{1}{2}}\boldsymbol{M}) = [(\boldsymbol{M}^\top \Phi^{\frac{1}{2}}) \otimes \boldsymbol{I}_n] \operatorname{vec} \boldsymbol{U}$$

and similarly

$$\mathbb{VAR}[\boldsymbol{z}|\boldsymbol{U}, \Phi, \boldsymbol{\Sigma}] = \left[(\boldsymbol{M}^\top \bar{\Phi}^{1/2}) \otimes \boldsymbol{I}_n\right] \mathbb{E}\left[\operatorname{vec}\boldsymbol{V} (\operatorname{vec}\boldsymbol{V})^\top\right] \left[(\bar{\Phi}^{1/2}\boldsymbol{M}) \otimes \boldsymbol{I}_n\right] = \\ = \left(\boldsymbol{M}^\top \bar{\Phi} \boldsymbol{M}\right) \otimes \boldsymbol{I}_n$$

The distribution of $\boldsymbol{z}|\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}$ then reads:

$$\begin{aligned} -2 \ln \pi \left(z|\boldsymbol{U}, \boldsymbol{\Sigma}, \phi\right) = \\ = C + (\operatorname{vec} \boldsymbol{z})^\top \left[\left(\boldsymbol{M}^{-1} \bar{\Phi}^{-1} {\boldsymbol{M}^{-1}}^\top\right) \otimes \boldsymbol{I}_n\right] \operatorname{vec} \boldsymbol{z} + \\ -2(\operatorname{vec} \boldsymbol{z})^\top \left[\left(\boldsymbol{M}^{-1} \bar{\Phi}^{-1} \Phi^{1/2}\right) \otimes \boldsymbol{I}_n\right] \operatorname{vec} \boldsymbol{U} + \\ +(\operatorname{vec} \boldsymbol{U})^\top \left[\left(\Phi \bar{\Phi}^{-1}\right) \otimes \boldsymbol{I_n}\right] \operatorname{vec} \boldsymbol{U} \end{aligned}$$

Now, for brevity let us define the following $p \times p$ matrices:

$$\boldsymbol{q}_{11} := \boldsymbol{M}^{-1} \bar{\Phi}^{-1} {\boldsymbol{M}^{-1}}^\top; \quad \boldsymbol{q}_{12} := \boldsymbol{M}^{-1} \bar{\Phi}^{-1} \Phi^{\frac{1}{2}}; \quad \boldsymbol{q}_{22} := \Phi \bar{\Phi}^{-1}$$

Hence

$$\begin{aligned} -2 \ln \pi \left(\operatorname{vec} \boldsymbol{z}|\boldsymbol{U}, \boldsymbol{\Sigma}, \Phi\right) = C + \operatorname{vec}(\boldsymbol{z})^\top \left(\boldsymbol{q}_{11} \otimes \boldsymbol{I}_n\right) \operatorname{vec} \boldsymbol{z} + \\ -2(\operatorname{vec} \boldsymbol{z})^\top \left(\boldsymbol{q}_{12} \otimes \boldsymbol{I}_n\right) \operatorname{vec} \boldsymbol{U} + \\ +(\operatorname{vec} \boldsymbol{U})^\top \left(\boldsymbol{q}_{22} \otimes \boldsymbol{I}_n\right) \operatorname{vec} \boldsymbol{U} \end{aligned}$$

And

$$\begin{aligned} -2 \ln \pi \left(\boldsymbol{z}, \boldsymbol{U}|\boldsymbol{\Sigma}, \boldsymbol{\Phi}\right) = C + (\operatorname{vec} \boldsymbol{z})^\top \left(\boldsymbol{q}_{11} \otimes \boldsymbol{I}_n\right) \operatorname{vec} \boldsymbol{z} + \\ -2\operatorname{vec} \boldsymbol{z}^\top \left(\boldsymbol{q}_{12} \otimes \boldsymbol{I_n}\right) \operatorname{vec} \boldsymbol{U} + \\ +(\operatorname{vec} \boldsymbol{U})^\top \left(\boldsymbol{q}_{22} \otimes \boldsymbol{I_n} + \boldsymbol{I_k} \otimes \boldsymbol{L}\right) \operatorname{vec} \boldsymbol{U} \end{aligned}$$

Hence, with some straightforward algebra, it can be concluded that:

$$\begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{U} \end{pmatrix} \sim \mathcal{N}_{2pn}\left(0, \begin{pmatrix} \boldsymbol{q}_{11} \otimes \boldsymbol{I}_n & -\boldsymbol{q}_{12} \otimes \boldsymbol{I}_n \\ -\boldsymbol{q}_{12}^\top \otimes \boldsymbol{I}_n & \boldsymbol{q}_{22} \otimes \boldsymbol{I_n} + \boldsymbol{I_k} \otimes \boldsymbol{L} \end{pmatrix}^{-1}\right) \tag{29}$$

Which consists of four blocks; three are diagonal and the lower-right one has the same sparsity pattern as $\boldsymbol{L}$.

.3 Expectation of covariate effects and latent fields in Gaussian models

The model in equation 1.12 can be parametrised in a more general way (see e.g. Spiegelhalter et al., 2002):

$$\begin{aligned} y &\sim \mathcal{N}_{n_y}\left(\boldsymbol{A_1}\vartheta_1,\, \tau_\varepsilon^{-1}\boldsymbol{I_{n_y}}\right) \\ \vartheta_1 &\sim \mathcal{N}_{n_\vartheta}\left(\boldsymbol{A_2}\vartheta_2,\, \tau_\varepsilon^{-1}\boldsymbol{C}\right) \end{aligned} \tag{30}$$

Where $\boldsymbol{A_1} = (\boldsymbol{X}\,\boldsymbol{\xi})$ is the model matrix, $\vartheta_1 = (\beta^\top z^\top)^\top$ is the vector of parameters, and $n_\vartheta = n_z + p$. The term $\boldsymbol{A_2}\vartheta_2$ has by construction both mean and variance zero. Moreover,

$$\boldsymbol{C}^{-1} = \tfrac{\tau}{\tau_\varepsilon}\begin{pmatrix} \frac{1}{\tau N}\boldsymbol{I_p} & \boldsymbol{0_{p\times n_z}} \\ \boldsymbol{0_{n_z\times p}} & \boldsymbol{R} \end{pmatrix}$$

Since the error term ε and the latent effects z have different precision parameters, the matrix $\boldsymbol{C}^{-1}$ cannot be considered known. To see this, consider the prior precision of the spatial effects defined as $\tau\boldsymbol{R}$; $\tau_\varepsilon\boldsymbol{C}^{-1}$ must thus be structured as $\begin{pmatrix} \frac{1}{N}\boldsymbol{I_p} & \boldsymbol{0_{p\times n_z}} \\ \boldsymbol{0_{n_z\times p}} & \tau\boldsymbol{R} \end{pmatrix}$ where N is the known variance parameter of β, but τ is unknown and isolating τ to express $\tau_\varepsilon\boldsymbol{C}^{-1} = \tau\begin{pmatrix} \frac{1}{\tau N}\boldsymbol{I_p} & \boldsymbol{0_{p\times n_z}} \\ \boldsymbol{0_{n_z\times p}} & \boldsymbol{R} \end{pmatrix}$ would leave it dependent on an unknown parameter.

Now, it is possible to show (Lindley and Smith, 1972) that the posterior mean of ϑ_1, conditional on the model hyperparameters, is

$$\mathbb{E}[\vartheta_1|y,\tau,\tau_\varepsilon] = \left(\boldsymbol{A_1}^\top\boldsymbol{A_1} + \boldsymbol{C^{-1}}\right)^{-1}\boldsymbol{A_1}^\top y \tag{31}$$

Based on how we defined $\boldsymbol{A_1}$ and $\boldsymbol{C}$, consider that

$$\begin{aligned} \left(\boldsymbol{A_1}^\top\boldsymbol{A_1} + \boldsymbol{C}^{-1}\right)^{-1} &= \begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{X} + r_\varepsilon\boldsymbol{I_p} & \boldsymbol{X}^\top\boldsymbol{\xi} \\ \boldsymbol{\xi}^\top\boldsymbol{X} & \boldsymbol{\xi}^\top\boldsymbol{\xi} + \bar{r}\boldsymbol{R} \end{pmatrix}^{-1} = \\ &= \begin{pmatrix} \boldsymbol{G^{-1}} + \boldsymbol{G^{-1}}\boldsymbol{X}^\top\boldsymbol{\xi}\boldsymbol{S_{11}}^{-1}\boldsymbol{\xi}^\top\boldsymbol{X}\boldsymbol{G}^{-1} & -\boldsymbol{G}^{-1}\boldsymbol{X}^\top\boldsymbol{\xi}\boldsymbol{S_{11}}^{-1} \\ -\boldsymbol{S_{11}}^{-1}\boldsymbol{\xi}^\top\boldsymbol{X}\boldsymbol{G}^{-1} & \boldsymbol{S_{11}}^{-1} \end{pmatrix} \end{aligned} \tag{32}$$

Where $\bar{r} = \tau/\tau_\varepsilon$, $r_\varepsilon = 1/N\tau_\varepsilon$, $\boldsymbol{G} = \left(\boldsymbol{X}^\top\boldsymbol{X} + r_\varepsilon\boldsymbol{I_p}\right)$ and $\boldsymbol{S_{11}}$; the matrix $\boldsymbol{S_{11}} := (\boldsymbol{\xi}^\top\boldsymbol{\xi} + \bar{r}\boldsymbol{R} - \boldsymbol{\xi}^\top\boldsymbol{X}\boldsymbol{G}^{-1}\boldsymbol{X}^\top\boldsymbol{\xi})$ is the Schur complement (Horn and Johnson, 2012, paragraph 0.7.3) of the upper-left block in $\boldsymbol{A_1}^\top\boldsymbol{A_1} + \boldsymbol{C}^{-1}$, $\bar{r} := \dfrac{\tau}{\tau_\varepsilon}$.

Putting equations 31 and 32 together, the posterior expectation of ϑ_1 conditional on hyperparameters, as detailed in equations 1.14 and 1.15 follows as an exercise of algebra.

.4 KLD between Gaussian densities

Here, the formula for the KLD between two Gaussian densities is proved; with no loss of generality, let us compare $\varphi_1(z) := \varphi(z|\mu_1, \mathrm{Var}_1)$ against $\varphi_0(z) :=$

$\varphi(z|\mu_0, \text{Var}_0)$, where z is a Gaussian random vector of size n and μ and Var are generic terms for mean and variance. The only assumption we make is that Var_1 and Var_0 are positive definite. Then we have, by definition:

$$\text{KLD}(\varphi_1||\varphi_0) = \int \left\{ \frac{n}{2} \ln \frac{|\text{Var}_0|}{|\text{Var}_1|} - \frac{1}{2} \left[(z-\mu_1)^\top \text{Var}_1^{-1} (z-\mu_1) - (z-\mu_0)^\top \text{Var}_0^{-1} (z-\mu_0) \right] \right\} \varphi_1(z)\, dz =$$

$$= \frac{n}{2} \ln \frac{|\text{Var}_0|}{|\text{Var}_1|} - \frac{1}{2} \int (z-\mu_1)^\top \text{Var}_1^{-1} (z-\mu_1)\, \varphi_1(z)\, dz +$$

$$+ \frac{1}{2} \int (z-\mu_1+\mu_1-\mu_0)^\top \text{Var}_0^{-1} (z-\mu_1+\mu_1-\mu_0) \varphi_1(z) dz$$

Consider that the first integrand function is a scalar, and a scalar trivially equals its trace. Then, leveraging on the cyclic property of trace, we see that

$$-\frac{1}{2} \int (z-\mu_1)^\top \text{Var}_1^{-1} (z-\mu_1)\, \varphi_1(z)\, dz = -\frac{1}{2} \int \text{tr}\left((z-\mu_1)^\top \text{Var}_1^{-1} (z-\mu_1) \right) \varphi_1(z)\, dz =$$

$$= -\frac{1}{2} \int \text{tr}\left((z-\mu_1)(z-\mu_1)^\top \text{Var}_1^{-1}\, \varphi_1(z)\, dz \right) =$$

$$= -\frac{1}{2} \text{tr}\left(\int (z-\mu_1)(z-\mu_1)^\top \varphi_1(z)\, dz \right) \text{Var}_1^{-1} =$$

$$= -\frac{1}{2} \text{tr}\left(\mathbb{VAR}[z|z \sim \mathcal{N}(\mu_1, \text{Var}_1)] \text{Var}_1^{-1} \right) = -\frac{n}{2}$$

Regarding the second integral, we can write it as:

$$\frac{1}{2} \int (z-\mu_1)^\top \text{Var}_0^{-1} (z-\mu_1) \varphi_1(z)\, dz - \frac{1}{2} \int (\mu_1-\mu_0)^\top \text{Var}_0^{-1} (z-\mu_1) \varphi_1(z)\, dz + \frac{1}{2} (\mu_0-\mu_1)^\top \text{Var}_0^{-1} (\mu_0-\mu_1)$$

The first of these three addenda can be simplified to:

$$\frac{1}{2} \int (z-\mu_1)^\top \text{Var}_0^{-1} (z-\mu_1) \varphi_1(z)\, dz =$$

$$= \frac{1}{2} \text{tr}\left(\text{Var}_0^{-1} \mathbb{VAR}[z|z \sim \mathcal{N}(\mu_1, \text{Var}_1)] \right) = \frac{1}{2} \text{tr}\left(\text{Var}_0^{-1} \text{Var}_1 \right)$$

The second addendum is proportional to $\mathbb{E}[z-\mu_1|z \sim N(\mu_1, \text{Var}_1)] := 0$ and cancels out by definition. Putting together these results, it follows that:

$$\text{KLD}(\varphi_1(z)||\varphi_0(z)) = \frac{n}{2} \ln \frac{|\text{Var}_0|}{|\text{Var}_1|} - \frac{n}{2} + \frac{1}{2} \text{tr}\left(\text{Var}_0^{-1} \text{Var}_1 \right) + \frac{1}{2} (\mu_1-\mu_0)^\top \text{Var}_0^{-1} (\mu_1-\mu_0) \tag{33}$$

Chapter 2

A Comprehensive Analysis of the Italian School System using the SchoolDataIT R Package

2.1 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 Barrett et al. (2019). 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 (Barrett et al., 2019).

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 summarised by factors such as illumination, indoor air quality (the main threat being the concentration of CO_2, which also can under-

mine 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, Garlaschi (2022) and Bucci et al. (2023) 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, Bucci et al. (2023) 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 Agasisti and Vittadini (2012). In particular, this disparity increases along the schooling process, implicitly suggesting that educational gaps tend to accumulate over time (Martini, 2020). 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 (Bagnarol and Donno, 2020; Donno et al., 2020, 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 Bucci et al. (2023).

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 (R Core Team, 2023) which retrieves and harmonizes some relevant institutional databases at the territorial level of either municipalities (LAU hereinafter) and provinces (NUTS-3 henceforth, Eurostat, 2024). 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 (Crescenzi et al., 2021; De la Porte and Jensen, 2021). Some simple datasets were loaded in Amelia in April 2025 and

regard classroom size, Invalsi scores, and a selecion of variables from the school buildings database. The relevant usage example and the full R codes, alongside with the complete metadata in `.csv` format can be found at this link: https://github.com/lcef97/AMELIA_datasets.

Data providers are the Italian Ministry of Education (formerly MIUR, Ministry of Education, University and Research) (Italian Ministry of Education, University and Research, 2024), the Institute for the Evaluation of the Education System (hereinafter Invalsi) (Invalsi - Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione, 2024), the Italian National Institute of Statistics (ISTAT, ISTAT - Italian National Institute of Statistics, 2022, 2025b,a), and the in-house company Infratel SPA on behalf of the Italian Ministry of Enterprises and Made in Italy (MIMIT, Infratel Italia, 2024), 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 permanent storage space in the local machine of the user.

The `SchoolDataIT` software is currently available under version 0.2.12, released on December 17th 2025 on the Comprehensive R Archive Network (CRAN). To ensure constant package maintenance, experimental versions are hosted on my GitHub webpage.

The remainder of this chapter is structured as follows. In Section 2.2, 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 2.3, we describe in detail the structure of the library and the most relevant functions made available for the users. Section 2.4 provides some usage notes and describes how web scraping is carried out. In Section 2.5, we describe the datasets that can be accessed through the package, while including some relevant examples and potentiality. Finally, in Section **??**, an empirical exercise involving the implementation of Bayesian spatial regression models is presented to investigate the student outcomes across the Italian territory.

2.2 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 present `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 Bucci et al. (2023), 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 2.5.3). 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 2.5.1, 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, Bucci et al. (2023) 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 Italian Ministry of Education, University and Research (2024), see also Section 2.5.2, 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 Garlaschi (2022), 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 (Bernardi and Keivabu, 2024) 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 or-

[1] In Italy, the seismic risk of a given area is classified based on the relative peak ground accel-

ganic framework to assess the seismic risk of school buildings, integrating several extant methodological approaches is described in Cattari et al. (2024). 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 Garlaschi (2022) and Bucci et al. (2023) 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 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 Bratti et al. (2007), who show that infrastructural variables can contribute to explaining part of the North-South divide in Programme for International Student Assessment (PISA) test scores (OECD, 2023).

2.3 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` (Müller and Wickham, 2023), thus fully compatible with the `Tidyverse` (Wickham et al., 2019). Figure 2.1 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 either the raw data to be downloaded, or some static `.zip` folders including the relevant inputs. Raw data are then converted to `.csv` and eventually into `R` objects, while `.zip` objects are handled in temporary folders which are deleted immediately after `R` acquires data. In doing so, it is needful to recall that no permanent storage space

eration (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/.

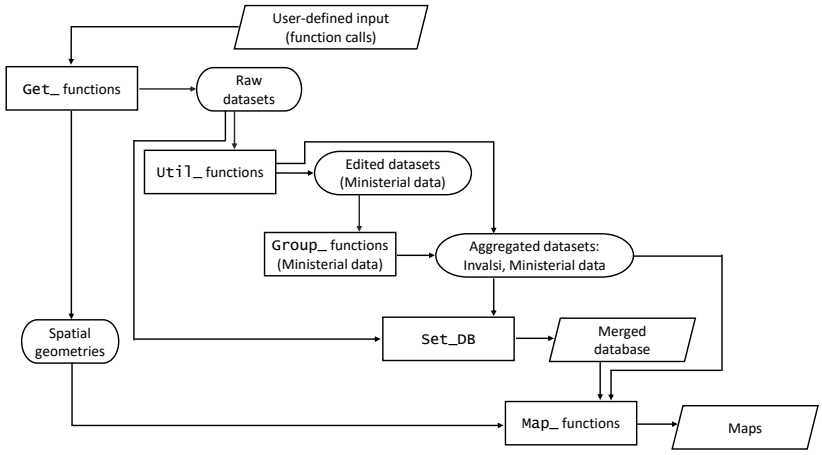

Figure 2.1: Flowchart of the package. Rhomboids denote inputs and R output objects; rectangles denote functions and rounded rectangles denote intermediate R objects.

is required on the local machine of the user outside the working environment of the R working session.

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 2.5.

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` (Wickham, 2011) environment for graphical representation, allowing a simplified export. Interactive maps, obtained through the `leaflet` (Cheng et al., 2019) and `mapview` (Appelhans et al., 2016) libraries, preserve all the information of the dataset to be rendered. Spatial geometries used for mapping are provided by the Italian Institute of Statistics through specific shape files (ISTAT - Italian National Institute of Statistics, 2025a).

2.4 Static web scraping

R packages can be easily distributed and made available to the general public. For instance, the integrated development environment (IDE) R-Studio (R Studio team, 2021), which is frequently employed by R users, even provides a window dedicated to sharing working directories on GitHub, hence R packages can be posted on GitHub and downloaded via the R function `devtools::install_github(...)`. While this is the most practical approach on the developer side, on the user side we deem more effective to employ the canonical repository of `R` extensions, namely the CRAN archive. `R` packages can be download from CRAN immediately. Since `R` is compatible with several operating systems, so must be all packages hosted on CRAN. This is ensured with the remote checks (Wickham and Bryan, 2023, Chapter 22) constantly run on 13 different operating systems; checks fail each time an exported function crashes or returns any error, or if the package exceeds the size of 5 megabytes.

As soon as the package violates the checks on any operating system, an alert is sent to maintainers who are given 10-15 days to solve all issues, otherwise the package is removed from CRAN.

Avoiding system crashes is a particularly sensible issue when `R` interferes with external programs or even more so with internet navigation. The present package relies on scraping static data objects from the web, hence it operates two requests. The first one consists in looking for the hyper-references in the data provider webpage. The second consists in downloading the relevant object including the data to retrieve. An insufficient but necessary condition to send web requests is the availability of a working internet connection. The `AutoAbort=TRUE` option of `Get_` functions is used to abort the whole operation when connection lacks; otherwise the user will be asked to give the system some time to wait for a connection to be established by using the R command line.

In versions until `0.2.12`, the HTML structure of the source page is inspected using either `xml2::read_html()`[2] or `httr::GET()`. In the experimental `0.2.13`, instead, we are switching towards the latter request handler in almost retrieval functions due to a slightly lower computational time. The only exception is `Get_BroadBand()`, for which the `httr` function is less robust. Differences in functions performance are driven by specific factors of the web pages to be scraped. Due to several server-side problems (e.g. the web address has been changed or for any reason the website architecture has changed), the HTML may happen not to be read correctly; in this case, the software is left with a total of 10 attempts; after that, the operation is aborted. If the webpage has been inspected correctly, the data objects are found among the attributes with anchor tag `<a>` (Munzert et al., 2015). These elements are filtered through the function `rvest::html_nodes("a")`. Among all anchored nodes, data objects can be accessed via a static link; therefore the next step is filtering hyper-references tags with `rvest::html_attr("href")`. Some other filters may be necessary to locate the exact link to the data source. Once the data source is identified, if it is a `.csv` file it can be read automatically from `R` from the raw file; if it is compressed in a `.zip` archive, as e.g. the shapefiles, it needs being downloaded and hosted in a temporary repository on the user local machine, which is removed after the operation is done. If any error occurs during the input file reading, the user is informed and the operation is aborted. A feature of an `R` package eligible for CRAN is that aborted operations do not result in `ERROR`s, but the function returns a `NULL` output. In the `R` developers jargon, this "emergency landing" procedure is known as `graceful failure` and if not properly implemented may easily cause remote check failures.

Another requirement from CRAN is limiting the number of dependencies, namely the number of packages imported (Wickham and Bryan, 2023, chapter 11). No more than 20 dependencies are typically allowed. Except for shapefiles, most input data objects are originally provided as `.csv` files, which can be read with some functions from the `readr` package. Two datasets are provided in

[2]The `rvest` package also features a function with the same name. It is basically a wrapper to the `xml2` function which is called internally

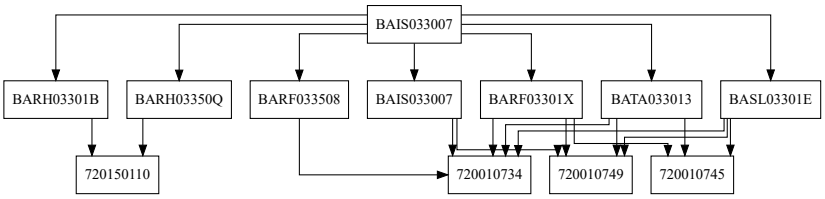

Figure 2.2: Example of the codes of all the schools (middle nodes) and the buildings in which they are located (bottom nodes), pertaining to the same reference institute (top node)

`.xlsx` format (i.e. excel files). In this case, `readr` is not sufficient anymore. The most conservative approach is unzipping the excel files and parsing the underlying `.xml` structure. This is feasible for the inner areas taxonomy, which is a relatively small worksheet, but becomes computationally intensive when dealing with the broadband dataset. In this case, we resort to `readxl` if it is already included in the user library, otherwise the slower unzipping procedure is activated. This allows us to list `readxl` among suggested packages and keep the number of dependencies as low as possible.

2.5 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 (Italian Ministry of Education, University and Research, 2024), an open data portal managed by the Italian Ministry of Education according to Law 107/2015 (Italian Official Journal, 2015). The National Schools Registry (Section 2.5.1), the school buildings database (2.5.2) and the counts of students and teachers (2.5.3) 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 Infratel Italia (2024) has been included in the package as well. Lastly, Invalsi censuary survey data (Invalsi - Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione, 2024) have been included to assess education quality (2.5.5).

Italian schools are officially identified by a 10-digit alphanumeric code. In Fig. 2.2 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).

2.5.1 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 (ISTAT - Italian National Institute of Statistics, 2025b).

2.5.2 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 (ISTAT - Italian National Institute of Statistics, 2022). 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, characterised by a comprehensive and self-sufficient combination of school, health, and transport infrastructure (ISTAT - Italian National Institute of Statistics, 2022); 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 2.3 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 2.5, 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 online Supplementary material of the paper corresponding to this chapter.

2.5.3 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

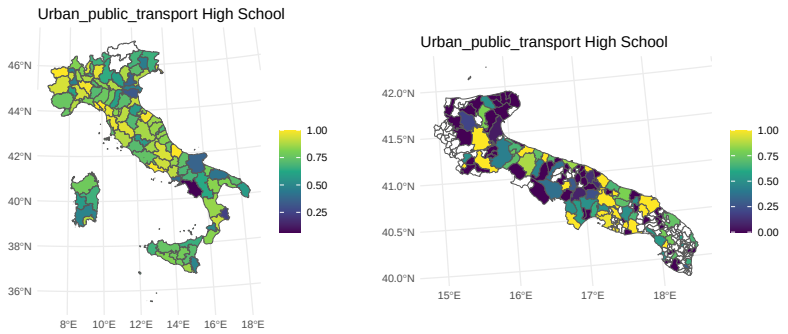

Figure 2.3: Province-level and municipality-level percentage of high schools served by public transport in 2022/23, on the whole national territory and in the Apulia region respectively. Data of the Trentino-Alto Adige region are not provided by the Ministry.

of education quality, which is typically acknowledged to improve as classroom size decreases (Blatchford et al., 2011; Brühwiler and Blatchford, 2011). In the case of Italy, however, caution is needed when studying the relationship between classroom size and student outcomes at the aggregate level (Angrist et al., 2017). 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 2.4 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 **??**.

The function `Util_nstud_wide` rearranges the input dataset into a wide 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 (Italian Ministry of Education, 2023, 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 (Italian Official Journal, 2009, Artt 10, 11). A framework of waivers is established by the Ministry Decree n.90/2023 (Italian Official Journal, 2009), 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

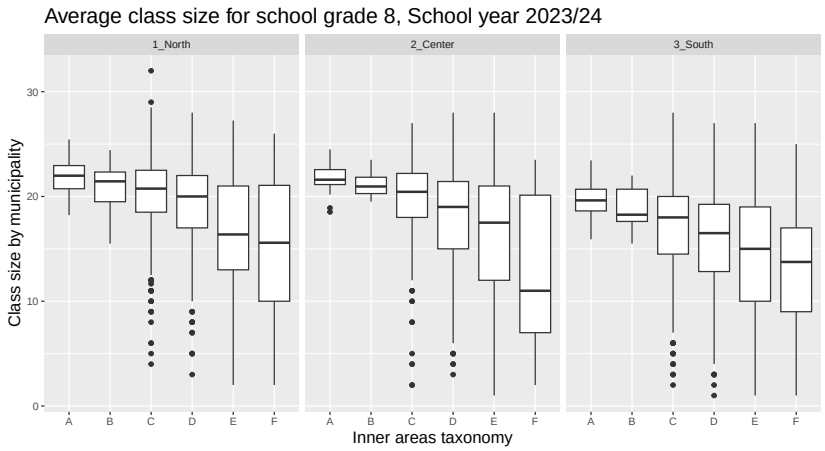

Figure 2.4: Average municipality-level classroom size by Inner Areas taxonomy in 2023/24, last grade of middle schools. Inner area taxonomy follows a descending order of centrality: A - standalone infrastructural poles; B - inter-municipality infrastructural poles; C - belt municipalities; D - intermediate areas; E - peripheral; F - ultra-peripheral.

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 2.5 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.

2.5.4 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 (Italian Ministry of Economic Development, 2020). The Plan aims

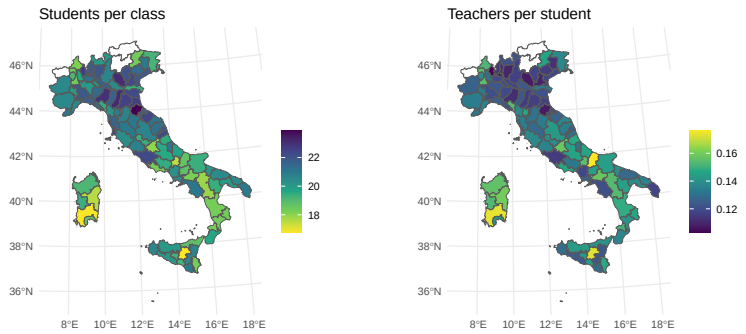

Figure 2.5: Average province-level classroom size in the 2nd year of high school and province-level teachers/students ratio in high schools in 2022/23. Data are not available for the Trentino-Alto Adige and Aosta Valley regions. Additionally, schools with average classroom size of less than 10 or more than 40 students in any grade have been filtered out.

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 (Infratel Italia, 2024). 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 specifies the reference date for checking whether the ultra-broadband connection was activated or not in each school.

2.5.5 Invalsi census survey

To develop a spatially homogeneous indicator of education quality, Italian law No. 176/2007 (Italian Official Journal, 2007) 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 (Invalsi - Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione, 2024), 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 (OECD, 2009) 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 3.

2.6 Example: Student outcomes in Mathematics and classroom size

Here we provide an example of spatial statistical application to the data covered by the `SchoolDataIT` package. Following Section 2.5.3, suppose the user is interested in studying to what extent classroom size is associated with student outcomes, say, in Mathematics at the last year of middle school, for the school year 2023/24. We can start by mapping the Invalsi scores and the average classroom size in figure 2.6. The easiest kind of analysis is regressing Invalsi scores 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; the choice of this interval ensures more classroom size variability than the narrower legal limits, which are, however, subjected to a regime of waivers. With these data at hand, 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.

Preliminary results from the ordinary least squares (OLS) regression return an estimated coefficient for classroom size equal to 4.153, with standard error 0.407. If no additional information is taken into account, classroom size would

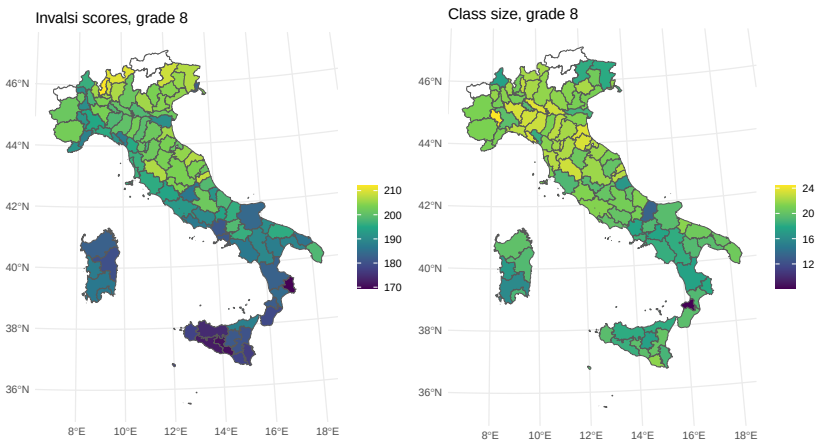

Figure 2.6: Invalsi score in Mathematics and average classroom size, last year of middle school, school year 2023/24. Trentino-Alto Adige and Aosta Valley are not included due to lack of classroom size data.

then appear to have a positive and statistically significant relationship with Invalsi scores. Such a counter-intuitive outcome could stem from either the lack of other explanatory variables, or from the possible selection bias caused by the exclusion of municipalities with a single middle school from the Invalsi data record. A closely linked result, namely the significantly positive association of Invalsi scores with the students/teacher ratio, was noticed by Barbieri et al. (2017), 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 appeared to be no longer significant. This may be suggesting an omitted variable bias in our OLS results.

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 have seen in 2.4, 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:

$$y(s) = \beta_0 + \beta_{\text{nstud}} X_{\text{nstud}}(s) + \beta_I X_I(s) + \beta_\ell \ell(s) + \beta_\phi \phi(s) + u(s) + \varepsilon(s) \quad (2.1)$$

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 $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_n)$. 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 (Banerjee et al., 2014, 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. The Matérn covariance function is described in appendix .2.

Fitting spatial models such as the one in equation 2.1 implies the inversion of the covariance matrix of u, which is typically dense as it can be seen from equation 2. In geostatistical literature, this issue is typically referred to as the big-n problem (Jona Lasinio et al., 2013) 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. Lindgren et al. (2011) 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 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 (Rue and Held, 2005), as seen in Section 1.2.

Mapping the Matérn field onto a GMRF shrinks the asymptotic computational cost to $\boldsymbol{\theta}(N^{3/2})$, due to the sparse precision structure of processes defined on bi-dimensional domains.

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 2.1 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 employ a joint PC-prior on the range r and the global standard deviation σ of the latent spatial field. The behaviour of these distributions is described in Fuglstad et al. (2019), who notably show that the joint density factorises into the relevant marginals. We select two fixed values, σ_0 and r_0 such that, for two fixed values p_σ and p_r, $\text{prob}\,(r < r_0) = p_r$ and $\text{prob}\,(\sigma > \sigma_0) = p_\sigma$.

	mean	s.d.	$\boldsymbol{q}_{0.025}$	$\boldsymbol{q}_{0.975}$
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 2.1: Estimated effects of classroom size, inner area dummy, latitude and longitude on Invalsi scores in Italian, last year of middle school, under model 2.1

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 2.1. 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 2.7.

The darker zones in the right panel (lower values of $\mathbb{E}[u|y]$) 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.

2.7 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

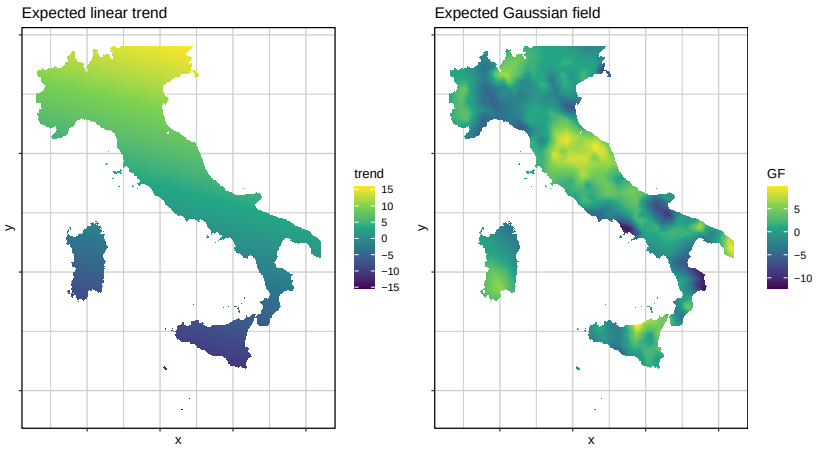

Figure 2.7: Posterior expectation of the linear trend on the left and of the spatial latent variable on the right, the latter being added to the model to explain spatial variability in Invalsi scores after controlling for the linear trend and the explanatory variables.

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 statistical 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

.2 The Matern covariance function

The Matérn covariance function between two generic sites i and j is given by:

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

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

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

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 (Abramowitz and Stegun, 1972, 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, Whittle (1954) 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):

$$(\kappa^2 - \Delta)^{\frac{\nu+1}{2}} \tau u(s) = \epsilon(s) \tag{3}$$

Where Δ is the Laplacian operator and $\epsilon(s)$ denotes a spatial Gaussian white noise process. Parameters ν and with κ are linked to the global variance

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

σ^2 through the relationship:

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

	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 3

Spatial Modelling of High Schools Invalsi Scores

3.1 Introduction

As it has been stressed out in 2.7, 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, OECD, 2009). In Italy, this task is attributed by law (Italian Official Journal, 2007) 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 (Martini, 2020), 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. Giancola et al., 2010; Bratti et al., 2007; Agasisti and Vittadini, 2012; Giancola and Salmieri, 2020) unless more explanatory variables regarding the labour market and socio-demographic dynamics are taken into account in relatively complex econometric models (Bratti et al., 2007). Additionally, Agasisti and Vittadini (2012) 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 (Matteucci and Mignani, 2014) (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 (ISTAT - Italian National Institute of Statistics, 2022), 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 (Besag, 1974), we assume an Intrinsic Conditional Autoregressive structure (hereinafter ICAR, Besag et al., 1991). Since the scores in two subjects are available, a bivariate ICAR (Mardia, 1988) 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 (Urdangarin et al., 2024) 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 (Rue et al., 2009; Van Niekerk et al., 2023), see also section 1.6. In particular, multivariate spatial modelling of areal data is implemented in the `INLAMSM` package (Palmí-Perales et al., 2021, 2023), 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 3.3 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 3.4 we discuss the results of the models implemented.

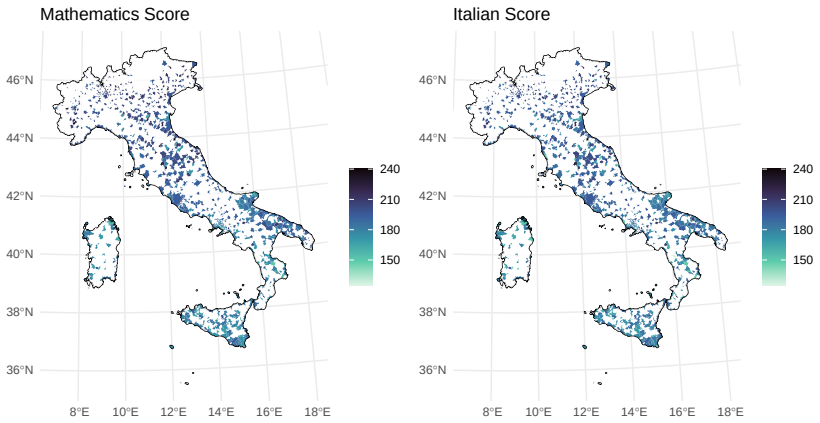

Figure 3.1: Invalsi scores in 2022/2023, 2nd year of high school. Trentino-Alto Adige is absent due to the lack of availability of auxiliary variables.

3.2 Student outcome data and infrastructural endowment

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 2.7.

3.2.1 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 3.1. 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.

3.2.2 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 (Italian Ministry of Education, University and Research, 2024). 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 Infratel Italia (2024). 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 (ISTAT - Italian National Institute of Statistics, 2022) and includes six classes, defined as in Chapter 2.7. 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.

3.2.3 Spatial structure

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

Two alternative definitions of the macro-areas are used, corresponding to two spatial models. The first level are provinces (NUTS-3 units), i.e., $n_z = 105$ 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 and symmetric neighbourhood matrix $\boldsymbol{W}$.

3.2.4 Spatial exploratory analysis of explanatory variables

In this Section, the spatial structure of explanatory variables at the province level is briefly explored. In Figure 3.2 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 (ISTAT - Italian National Institute of Statistics, 2022) 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 3.1 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 (Cliff and Ord, 1981). 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 (Bivand et al., 2017).

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 3.1: 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 3.2: Moran's I values for infrastructural catchment area-level averages of auxiliary variables

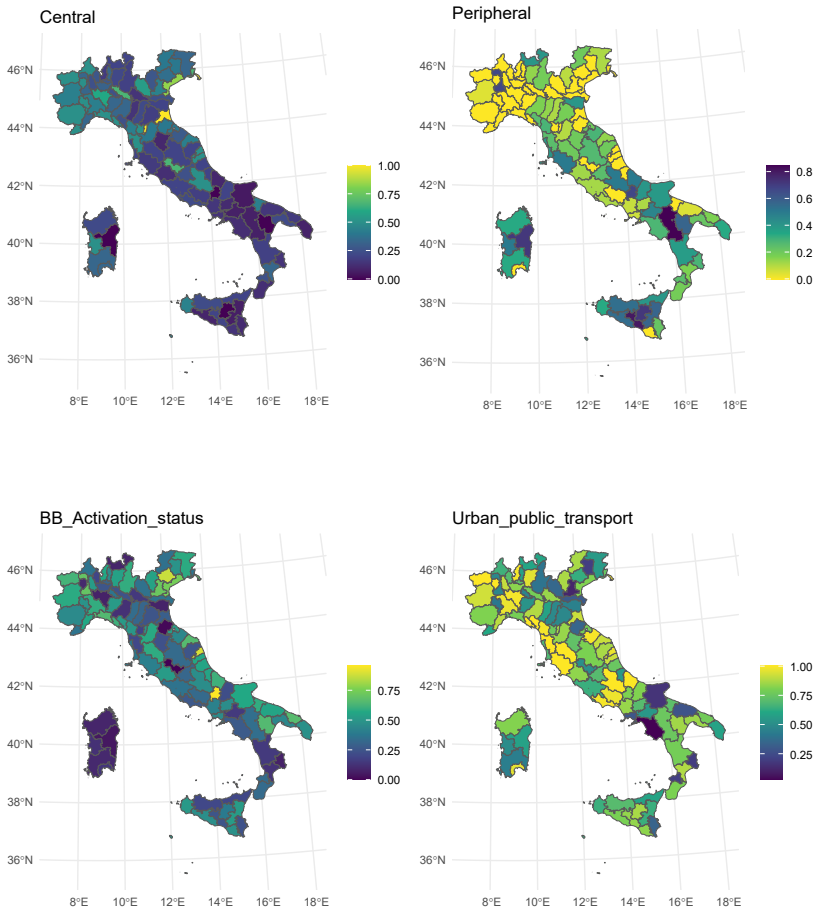

Figure 3.2: Upper panel: proportion of central (left) and peripheral (right) municipalities per province. Lower panel: ultra-broadband availability (left) and urban transport accessibility (right) per province.

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

$$y = \tilde{\boldsymbol{X}}\,\beta + \tilde{\boldsymbol{\xi}}\,\tilde{\boldsymbol{C}}\,\beta_C + \tilde{\boldsymbol{\xi}}\,z + \varepsilon \tag{3.1}$$

where $\tilde{\boldsymbol{X}} := \boldsymbol{I}_2 \otimes \boldsymbol{X}$, $\tilde{\boldsymbol{\xi}} := \boldsymbol{I_2} \otimes \boldsymbol{\xi}$ and $\tilde{\boldsymbol{C}} := \boldsymbol{I}_2 \otimes \boldsymbol{C}$. $\boldsymbol{X}$ is the $n_y \times 4$ matrix of auxiliary variables net of intercepts (see Section 3.2.2), and β is a vector of fixed effects of length 8. The $n_y \times n_z$ matrix $\boldsymbol{\xi}$ is binary and maps the n_z macro-areas onto n_y municipalities; each row of $\boldsymbol{\xi}$ is a unit vector in the n_z-dimensional space. The $n_z \times 3$ matrix $\boldsymbol{C}$ is also binary and denotes which connected component (continent, Sicily, Sardinia) each macro-area belongs to, β_C is the vector of subject-specific and 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 bivariate random error, following the distribution:

$$\begin{cases} \varepsilon_1 | \omega_1 \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \omega_1) \\ \varepsilon_2 | \omega_2, \alpha \overset{\text{iid}}{\sim} \text{SN}(m_{\omega_2,\alpha}, s_{\omega_2,\alpha}, \alpha) \end{cases} \tag{3.2}$$

The Mathematics score is supposed to be Normal in continuity with extant literature. $\text{SN}(\cdot)$ in 3.2 denotes the Skew-Normal distribution (Azzalini and Capitanio, 2014) with location and scale parameters $m_{\omega_2,\alpha}$ and $s^2_{\omega_2,\alpha}$ defined to ensure that $\mathbb{E}[\varepsilon_2|\alpha] = 0$ and $\mathbb{VAR}[\varepsilon_2|\alpha] = \omega_2$, and α is the shape parameter., i.e.

$$\pi(\varepsilon_2|\omega_2, \alpha) = \frac{2}{s_{\omega_2,\alpha}} \varphi\left(\frac{\varepsilon_2 - m_{\omega_2,\alpha}}{s_{\omega_2,\alpha}}\right) \Phi\left(\alpha \frac{\varepsilon_2 - m_{\omega_2,\alpha}}{s_{\omega_2,\alpha}}\right)$$

This choice is due to the negative skewness in municipality-level Italian scores that neither auxiliary variables or spatial effects can explain (see Fig. 3.8 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 Van Niekerk and Rue (2021), α is assigned a Penalised Complexity prior with a given rate parameter λ, the reference model being the Gaussian likelihood. Van Niekerk and Rue (2021) found that the KLD between the Skew-Normal and Normal densities is approximated by an even polynomial in α; regarding the functional form, smaller values of λ tend to move the probability mass away from zero, flattening the distribution of γ_1. $\lambda = 4$ 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 (Van Niekerk and Rue, 2021).

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 3.2.1). Error variances ω_1 and ω_2 have independent Gamma vague priors with shape 10^{-3} and rate 10^{-3}.

3.3.1 Modelling the spatial component

Considering the neighbourhood structure outlined in Section 3.2.3, z is modelled as a bivariate ICAR defined on the graph $\mathcal{G}$, following the prior defined in equation 1.4 In this case, considering the graph structure, the joint prior distribution of the ICAR field becomes

$$\pi\left(z|\boldsymbol{\Lambda}\right) = \left(\frac{1}{2\pi}\right)^{n-3} \sqrt{|\boldsymbol{\Lambda} \otimes \boldsymbol{L}|_+}\; e^{-\frac{1}{2}(\operatorname{vec} \boldsymbol{z})^\top (\boldsymbol{\Lambda} \otimes \boldsymbol{L}) \operatorname{vec} \boldsymbol{z}} \tag{3.3}$$

where $\boldsymbol{L} := \boldsymbol{D} - \boldsymbol{W}$ is the Laplacian matrix of the graph $\mathcal{G}$ with 3 connected components and $\boldsymbol{D}$ is the diagonal degree matrix of $\mathcal{G}$ (as discussed more in detail in Chapter 1). The matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is the 2×2 precision parameter. Since $\boldsymbol{L}$ is singular with rank deficiency 3 (Hodges et al., 2003), it is necessary to constrain z to sum to zero within each connected graph component (Besag and Kooperberg, 1995). This is the reason for adopting component-specific intercepts β_C in equation 3.1.

To ease the interpretation of $\boldsymbol{\Lambda}$ 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 $\boldsymbol{L}$ Sørbye and Rue (2014) and reparametrising the precision of z accordingly. The scaling procedure is described in subsection 1.3.1; this procedure allows to compare the precision parameter across the connected components of the graph.

Since $\mathcal{G}$ is disconnected, each component-specific block of $\boldsymbol{L}$ 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 Freni-Sterrantino et al. (2018) and described in section 1.3.1. It is therefore possible to define a precision parameter $\boldsymbol{\Lambda}_{\text{scaled}}$ which is not confounded with graph-induced effects. The scaled precision is assigned a Wishart prior (Gelman et al., 2004) with $2k + 1$ degrees of freedom and scale parameter equal to the identity matrix, i.e. $\boldsymbol{\Lambda}_{\text{scaled}} \sim \text{Wishart}_k(I_k, 2k + 1)$, with $k = 2$ (Palmí-Perales et al., 2021).

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 $\boldsymbol{W}$ 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 $\boldsymbol{W}$ from the scaled Laplacian.

3.3.2 Spatial confounding treatment

In our multilevel framework, the value of the m-th covariate $\boldsymbol{X_{\cdot m}}$ 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: $\boldsymbol{X} = \boldsymbol{\xi}\bar{\boldsymbol{X}} + \boldsymbol{\Delta X}$, where $\bar{\boldsymbol{X}} = (\boldsymbol{\xi}^\top\boldsymbol{\xi})^{-1}\boldsymbol{\xi}^\top\boldsymbol{X}$. We can then apply the Spatial+2.0 procedure (Urdangarin et al., 2024) described in Section 1.5.4. Consider the eigendecomposition of the Laplacian matrix:

$$\boldsymbol{L} = \boldsymbol{V\Gamma V}^\top$$

where the eigenvalues in $\boldsymbol{\Gamma}$ are in decreasing order and the eigenvectors in $\boldsymbol{V}$ 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 $n_z \times 4$ matrix $\boldsymbol{B}$, we have $\bar{\boldsymbol{X}} = \boldsymbol{VB}$, and the spatial component of $\bar{\boldsymbol{X}}$ is thus determined by the last columns of $\boldsymbol{V}$ (Urdangarin et al., 2024): without loss of generality, $\bar{\boldsymbol{X}}$ is decomposed into:

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

where $\bar{\boldsymbol{X}}^{(\boldsymbol{NS})}$ is the nonspatial component, given by the linear combination of the first $n_z - G - K$ eigenvectors (with $G = 3$ connected graph components), $\bar{\boldsymbol{X}}^{(\boldsymbol{s})}$ is the combination of the eigenvectors associated with the last K nonzero eigenvalues and represents the spatial component, and $\bar{\boldsymbol{X}}^{(\boldsymbol{0})}$ 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, we consider $\xi\left(\bar{\boldsymbol{X}}^{(\boldsymbol{NS})}\right) + \boldsymbol{\Delta X}$ as the covariate matrix in the regression model.

For the present data, the Spatial+2.0 methodology implies a shift in the value of explanatory variables by an additive correction defined at the macro-area level, namely $\bar{\boldsymbol{X}}^{(\boldsymbol{S})} + \bar{\boldsymbol{X}}^{(\boldsymbol{0})}$. For a generic h-th covariate the corresponding deconfounded values are $x_{i;h} - x^{(S)}_{i;h} - x^{(0)}_{i;h}$; if the variable is binary, taking values $x_{i,j;h} \in \{0, 1\}$, the corresponding deconfounded values would belong to the couple $\{-x^{(S)}_{i;h} - x^{(0)}_{i;h},\ 1 - x^{(S)}_{i;h} - x^{(0)}_{i;h}\}$. Subsequently, to allow model comparisons addressed in Section 3.4, deconfounded covariates are scaled to have the same global variance as the input ones.

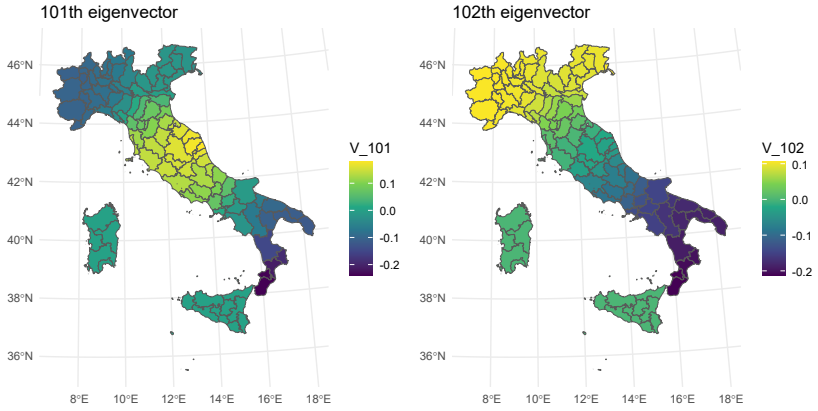

Figure 3.3: Lowest frequency eigenvectors, provinces

In Figure 3.3 we plot the two lowest-frequency eigenvectors of $\boldsymbol{L}$ at the macro-area level of provinces and varying across the continent. Eigenvectors are ordered decreasingly with respect to corresponding eigenvalues. It is possible to see that the 101st eigenvector follows a quadratic trend with one oscillation, whereas the 102nd 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 $\bar{\boldsymbol{X}}^{(\boldsymbol{NS})}$. A documented approach (Urdangarin et al., 2024; Lamouroux et al., 2024) is to choose the value of K minimizing the Watanabe-Akaike Information Criterion (WAIC, Watanabe, 2013). 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.

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 $\bar{\boldsymbol{X}}^{(\boldsymbol{NS})}$ not to display evidence of autocorrelation according to Moran's I index, whose use is encouraged by its robustness to non-Normality Griffith (2010) in datasets with at least 100 observations. 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 3.2.

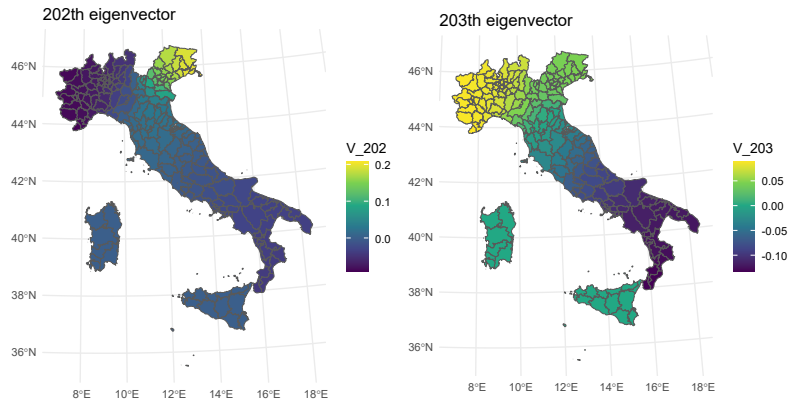

Figure 3.4: Lowest frequency eigenvectors, infrastructural catchment areas

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.

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 3.5 the first two columns of $\bar{\boldsymbol{X}}^{(\boldsymbol{NS})}$ 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 3.2.

3.3.3 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

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

Marginal Likelihood (LPML, Geisser and Eddy, 1979; Gelfand et al., 1992), i.e. minus the logarithmic sum of the Conditional Predictive Ordinates (Pettit, 1990; Held et al., 2010), and the Watanabe-Akaike Information criterion (WAIC, Watanabe, 2013), following the formulation of Gelman et al. (2014), the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002), 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 $j \in [1, n_y]$ and $h \in [1, 2]$, by:

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

Low CPO values denote "surprising" observations and, hence, possible outliers. Details on how the CPO is computed in `R-INLA` are provided by Held et al. (2010). Here, we compare models through the LPML (Geisser and Eddy, 1979), i.e. $\text{LPML} := -\sum_{h=1}^{2}\sum_{j=1}^{n_y} \ln \text{CPO}_{j,h}$.

The Deviance Information Criterion (DIC, Spiegelhalter et al. (2002)) is given by:

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

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, $D(\vartheta, \Psi) = -2 \ln \pi(y|\vartheta, \Psi)$ being the model deviance, whose expectation measures the goodness of fit. Following Rue et al.

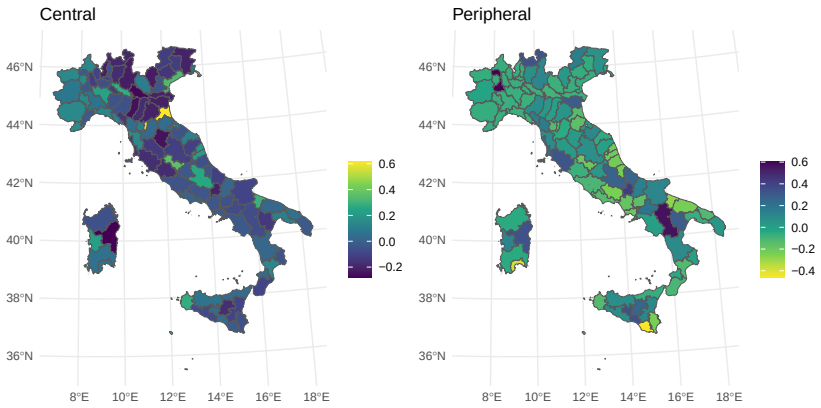

Figure 3.5: Proportion of central (left) and peripheral (right) municipalities per province once the spatial structure is removed by applying the S+(2) correction (see Table 3.3).

(2009) 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 Watanabe (2013), 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 (Gelman et al. (2014)). We use the formulation provided by Gelman et al. (2014).

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

Lastly we compute the Mean Square Error (MSE) of the expected value of linear predictors, a purely goodness-of-fit metric, only informative for point estimates. All selection criteria, except for the MSE, are internally computed by `R-INLA` via the option `control.compute = list(...,dic = TRUE, waic = TRUE, cpo = TRUE, ...)`

Models defined with ICAR random effects are compared in Table 3.4. "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 3.2.4), 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 (Khan and Calder, 2022).

In Appendix .1.1 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.

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 3.4: 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.

3.4 Results

In Table 3.5 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 2.5-th and 97.5-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 3.5: 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 3.6 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 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 3.7 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 4.3).

The highest scores are estimated in the municipalities of Lecco (Lombardia),

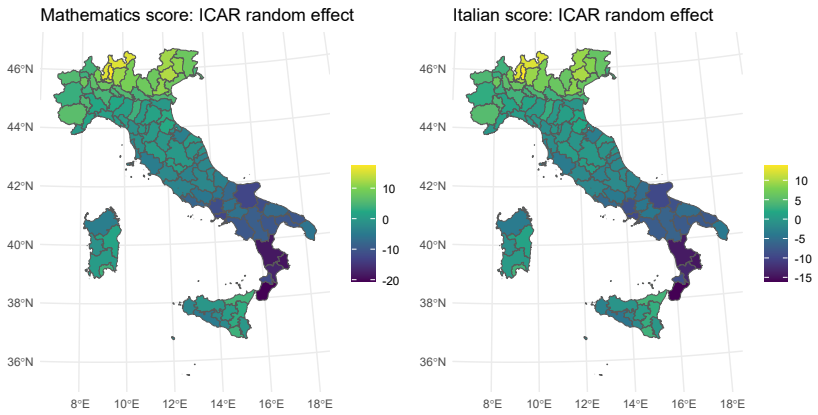

Figure 3.6: Expected values of z modelled as a province-level ICAR and applying S+(2) correction

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 4.3 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 3.6. Lastly, the choice of modelling Italian 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, defined as the difference between observed and predicted scores, is shown in Figure 3.8. Negative skewness is easily noticeable. Density is estimated by the Gaussian kernel, using the Silverman's thumb rule to define the bandwidth (Silverman, 1986, Section 3.4.2)

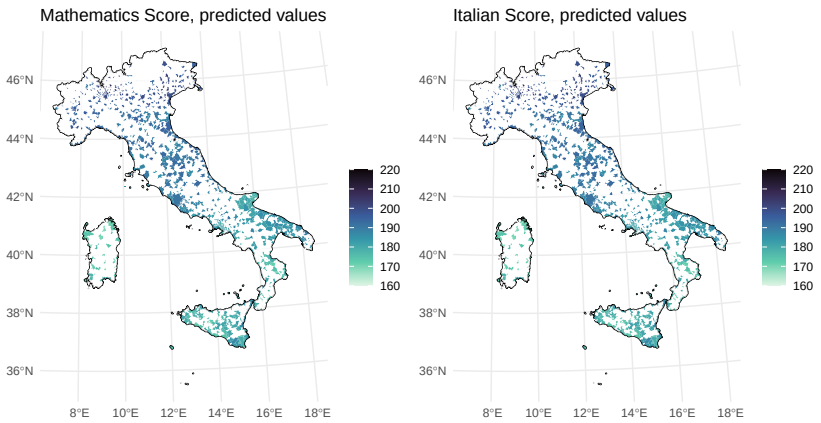

Figure 3.7: Predicted values of Invalsi scores using as a province-level ICAR latent effect and applying the S+(2) correction

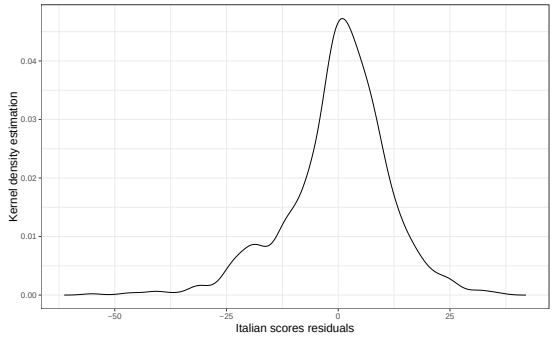

Figure 3.8: Kernel density estimation of the residuals of Italian scores under the S+(2) model

		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 3.6: 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)

3.5 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 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.

Choosing the appropriate number of spatial trends to apply the Spatial+2.0 method required to fit a high number of models, which would have been hardly feasible without an efficient computational method, as the INLA proves to be.

The mapping of infrastructure access and social vulnerability requires adequate statistical computational methods, and `R-INLA` appears an excellent 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 by default in the current `R-INLA` framework and has been used to estimate the models whose results are summarised in Sections **??** and 3.4. 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 Section 3.3.2. Alongside the selection criteria mentioned in Section **??**, 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 (Held et al., 2010), 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

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.

VB approximation even if we have a skewed likelihood for one of the two target variables.

.1.1 Estimates of regression coefficients 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

Chapter 4

Investigating access to support centers for Violence Against Women in Apulia: A Spatial analysis over multiple years

4.1 Introduction

Violence against women (VAW) is a widespread violation of human rights and a form of gender-based discrimination, encompassing acts that cause or risk causing physical, sexual, psychological, or economic harm. As recognised by the Istanbul Convention (Council of Europe, 2011), this structural issue is deeply tied to persistent gender inequalities worldwide, making accurate data essential for understanding its scale and developing effective policies. Despite this need, tracking gender-based violence remains challenging. In Italy, legislative initiatives such as Law No. 53/2022 (Italian Official Journal n.120, 2022) underscore the importance of systematic data collection for prevention, yet data sources remain fragmented, often outdated, and not easily accessible. Stigma and underreporting further hinder accurate quantification, as many survivors fear social repercussions, blame, or disbelief. Local Anti-Violence Centers (AVCs), established under Italian Law No. 119/2013 (Italian Official Journal n. 242, 2013), play a crucial role in supporting victims through a holistic array of services, starting from consultancy and including psychological support, legal assistance, and sheltering, while also documenting cases of abuse.

As stressed by Toffanin et al. (2020), a distinctive feature of AVCs is establishing intimate trust and promoting the subjectivity and the active role of

women seeking help, instead of substituting to them in their decisional processes. Taking into account also the exclusive participation of women as AVC workers and the outworks activities carried out, such as training activities for health and social operators, law enforcement and lawyers or cultural initiatives (Italian National Institute of Statistics, 2024), qualifies AVCs not as mere service providers, but as core political actors in the process of combating systematic gender-based violence. In addition, it is worthwhile to notice that both aims and operating modes of AVCs belong to a different conceptual level than judicial measures; hence, we deem that the access to AVCs deserves a dedicated, standalone analysis, not to be confused with the well-developed debate on violence reporting.

In 2017, the Italian National Institute of Statistics (ISTAT) started monitoring the activity of AVCs (Italian National Institute of Statistics, 2024), and support participation in the ISTAT experimental survey, conducted nationally since 2020 in collaboration with the DPO (Dipartimento delle Pari Opportunità, Presidenza del Consiglio). The ISTAT survey aims at capturing the scope and characteristics of violence experienced by women seeking help. From 2017 until 2023, the number of AVCs active over the national territory increased by 43.8%, with 404 AVCs operative in 2023. Similarly, the number of women accessing them increased by 41.5% (ibidem). This information may suggest that the expansion of AVC activity encourages women victims of violence to seek help and find a way out of violence and abuse. Records from AVCs provide valuable information on victim demographics and violence patterns (Toffanin et al., 2020), enabling investigation of the distribution of the phenomenon at the territorial level and allowing us to assess how local socioeconomic factors affect the occurrence and reporting of VAW.

As the process of finding a way out of violence may start with seeking help from AVCs (Statistical Office of the Apulia Region, 2024) and, more generally, from sheltering structures, the importance of shelter accessibility emerges immediately. On the one hand, a solution is provided by the availability of local help desks, which operate under the control of the AVCs and are widely distributed across the territory. On the other hand, the isolation of peripheral areas and the resulting lower accessibility to shelters may be hindering help-seeking, as noticed by Peek-Asa et al. (2011). In addition, Denti and Iammarino (2022) found strong evidence that local support services in the United Kingdom have a positive impact on gender-based violence reporting, even when controlling for the implementation of nationwide measures. Similarly, in Italy, Bettin and Ciaschini (2025) examined the *Codice Rosa* initiative (a triage code dedicated to hospitalising VAW victims) and anti-stalking help desks, highlighting the significant impact of the former on violence reporting nationwide.

Building on this evidence, we investigate how local disparities in the accessibility of AVCs may shape women's capacity to seek help and escape situations of violence. To this end, we propose a territorial analysis of accesses to AVCs within the Apulia region (Southern Italy) in years 2021–2024, detailed by the origin municipality of women seeking help, by means of a Bayesian hierarchical spatio-temporal Poisson regression model (Martínez-Beneito et al., 2008). We

investigate the association of AVC accesses with both the distance from AVCs and help desks, and a set of socio-economic indicators employed to define the Municipality Frailty Index provided by the ISTAT [1].

The inherent spatial dependence in the data suggests us to consider a latent spatial process, explicitly modelled through four competing spatial structures: the Intrinsic Conditional Autoregressive model (ICAR), the Proper Conditional Autoregressive model (PCAR), the Leroux Conditional Autoregressive model (LCAR) and the scaled Besag, York and Mollié convolution model (BYM2), all of which have been described in Chapter 1. These spatial models depend in turn on an additional layer of hyperparameters, which require an appropriate prior distribution setup. Here, following the principle of parsimony, we set hyperpriors in such a way to penalise departures from the corresponding non-spatial models. This can be done intuitively and flexibly through the Penalised Complexity prior (PC-prior) approach, developed by Simpson et al. (2017). Since, to the best of our knowledge, for the hyperparameters controlling the strength of spatial association in the PCAR and LCAR models the PC-prior has not been employed yet, we compare the results of the full PC-prior specification with a hybrid setup in which only this parameter is assigned a Uniform distribution.

Explicitly accounting for unmeasured spatial variability may cause confounding with some explanatory variables, when the latter show a spatial trend, as it implies some confounding bias in the estimation of the effects of such covariates. For those covariates whose effects mostly differ between a nonspatial and the corresponding spatial regression model, we mitigate this bias by reducing the effect of the spatial trend through the Spatial+2.0 method (Urdangarin et al., 2024) that captures the spatial trends through the lowest-frequency eigenvectors of the graph Laplacian matrix. We have introduced this methodology in Section 1.5.4.

Model estimation is carried out within INLA (Rue et al., 2009; Van Niekerk et al., 2023, see also chapter **??**); model selection is based on the Watanabe-Akaike Information Criterion (WAIC Gelman et al., 2014, see also 3.3.3), computed internally to `R-INLA` as well.

The spatial patterns of reported violence related to access to AVCs in Apulia allow us to identify differences in prevalence rates and evaluate the adequacy of available services. Specifically, the study region is characterised by significant socioeconomic disparities (Ottomano Palmisano et al., 2022), including variations in economic development, infrastructure, and service accessibility. This circumstance, together with the high level of territorial detail in available data provided by the Regional Statistical Office, makes Apulia an ideal case study to investigate how such factors affect the search for AVC assistance.

The remainder of this paper is structured as follows: section **??** introduces the data set of accesses to AVCs in Apulia and relevant explanatory variables, section 4.3 describes the Bayesian space-time statistical modelling framework and provides some generalities on the proposed PC-priors, section 4.4 accounts

[1] Available on the ISTAT website https://www.istat.it/comunicato-stampa/aggiornato-indice-di-fragilita-comunale/

for some practical implementation issues and summarises the strategy to prevent spatial confounding, section 4.5 provides some results of posterior inferences and interpretative comments.

4.2 The Apulia Region data

This study focusses on the number of accesses to AVCs in the municipalities of the Apulia region over the years 2021–2024. Data are collected by the Statistical Section and the Welfare Department of the Apulia Region as part of a systematic monitoring program established under Regional Law No. 29/2014. Since 2013, all AVCs have been required to submit annual reports detailing women's access to support services to monitor the effectiveness of the service and identify local problems. The resulting data collection and processing activities contribute to the Regional annual reports (Statistical Office of the Apulia Region, 2024).

To avoid considering duplicate data records, we only retain reports of violence that are directly managed by each center, excluding those that are redirected to other services or centers. In Apulia, the total number of women supported by AVCs was 1477, 1516, 1822, 1778 over the four years in scope. Fig. 4.1 shows the incidence of AVC accesses, obtained as the count of accesses over the female population aged > 14 years, across Apulian municipalities in the 4 years on the logarithmic scale, alongside with the municipalities with at least one operating AVC. White areas correspond to zero counts (in 2022, the center in San Severo (FG) was active, but no access results by women residing there). This map suggests a heterogeneous spatial pattern of access, with slightly higher values for coastal municipalities, indicating potential differences in population, infrastructure, and service availability between coastal and inland areas.

To investigate the association of access rates with potential drivers at the municipal level, a set of candidate explanatory variables is considered. First, we employ the distance of women residence municipalities from both the closest municipality hosting an AVC and the closest municipality with a help desk; between-municipalities road distances are provided by ISTAT at this address: https://www.istat.it/notizia/matrici-di-contiguita-distanza-e-pendolarismo/. Considering the deep link between socio-economic vulnerability and both the occurrence and reporting of gender-based violence, investigated in depth by Bettio and Ticci (2017), an additional set of explanatory variables has been selected amid the components of the Municipality Index of Fragility (IFC) referring to year 2021. These variables capture multiple dimensions of local vulnerability, such as economic deprivation, demographic imbalance, and social marginalisation. Including them in the analysis allows us to account for the contextual conditions that may influence both the risk of experiencing violence and the likelihood of seeking help. For instance, areas characterised by higher unemployment, ageing populations, or weaker social support networks may exhibit lower reporting rates or different help-seeking patterns.

These aspects will be further discussed with the results of the proposed model estimation. However, among the available indicators, we excluded the

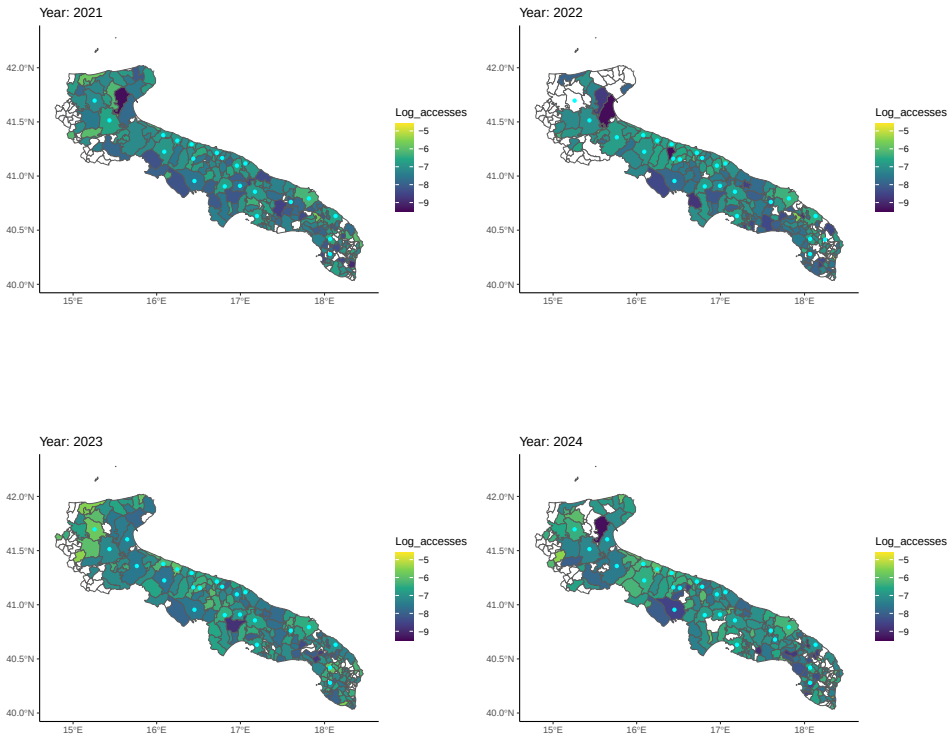

Figure 4.1: Log-incidence of AVC accesses. White areas correspond to zero counts. Points correspond to municipalities hosting at least one AVC.

distance of each municipality from the closest infrastructural pole (ISTAT - Italian National Institute of Statistics, 2022) due to the high correlation with the distance from AVCs (71%), which in the scope of this analysis appears to be a more informative variable. This high correlation is intuitive, as AVCs are typically located in larger municipalities with the potential to gather a higher number of victims and typically corresponding to infrastructure hubs. The set of explanatory variables considered in the present analysis is summarised as follows.

1. **AVC_Dist** Road travel time from each municipality to the closest one hosting an AVC, expressed in minutes. As the record of active AVCs is available for each year, this is the only time-changing covariate. As it can be observed from figure 4.1, changes in the spatial distribution of AVCs are minimal. Specifically, with data at hand the number of AVCs active across the region is 21, 24, 24 and 25 for the relevant years.

2. **Desk_Dist** Road travel time from each municipality to the closest hosting a help desk, expressed in minutes. We do not have the complete record of active desks in each municipality, hence we needed to impute the record from 2024 to all previous years.

3. **ELI** Ventile of the distribution of the municipality-level share of employees working in low productivity units in the fabric and services sectors. Labour productivity is defined as the ratio of value added per employee and is classified "low" if smaller than the first quartile of its distribution.

4. **PGR** Population growth rate in years 2011–2021, i.e. total net migration in 2011–2021 over resident population at the end of 2011.

5. **UIS** Ventile of the density of economic production units, defined as the ratio between the stock of active business companies over the total population.

6. **ELI** Percentage of people aged 25–64 with a low education level, i.e. no more than middle school license or vocational schools diploma.

7. **PDI** Structural dependency index, i.e. population aged $[0-19]$ or >64 over population aged 20–64.

8. **ER** Employment rate among population aged 20–64. The main indicator of economic development.

4.3 Space-time Poisson models

The counts of women residing in each Apulian municipality accessing AVCs within a given year are modelled via Poisson ecological regression. For a generic year $t \in \{2021, \dots, 2024\}$, we denote the count of women from municipality $i \in \{1, \dots, 256\}$ accessing an AVC as $y_{i,t}$, and model it as follows:

$$y_{i,t} \sim \text{Poisson}(P_{i,t}\, e^{\eta_{i,t}})$$

where $P_{i,t}$ is the female population aged >14 years[2] and $\eta_{i,t} := \beta_{0,t} + X_{i,t}^\top \beta + z_{i,t}$ is the linear predictor; $X_{i,t}$ is a vector of covariates described in Table **??**. The only covariate changing over time is the distance from AVCs, as the components of the MFI are not computed yearly, and only the list of help desks in 2024 was available. Moreover, β is a vector of constant covariate effects, $\beta_{0,t}$ are time-dependent intercepts and $z_{i,t}$ is a realisation of a latent spatio-temporal field $Z = (Z_1^\top, \dots, Z_T^\top)^\top$, modelled as a type IV interaction term, namely a latent field both spatially and temporally structured (Knorr-Held, 2000). Specifically,

[2]Data retrieved from ISTAT at: http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1

we assume a lag-1 temporal dependence, and model Z as follows:

$$\begin{cases} Z_t = rZ_{t-1} + \nu_t \quad \text{for} \quad t = 2, \ldots, T \\ \nu_t \sim \mathcal{N}_n\left(0, \sigma_t^2 \Omega_t\right) \quad \text{for} \quad t = 2, \ldots, T \\ Z_1 \sim \mathcal{N}_n\left(0, \frac{1}{1-r^2}\sigma_1^2 \Omega_1\right) \end{cases} \tag{4.1}$$

where ν_t is a spatially structured Gaussian random field taking values over a graph of order n, described by the symmetric and positive semi-definite variance-covariance matrix Ω_t, σ_t^2 is a marginal variance parameter and r is a temporal correlation parameter. Here, the nodes of the reference graph are the $n = 256$ Apulian mainland municipalities, and its edges are the links between pairs of neighbouring municipalities. This rather general model formulation allows the Ω_t matrices to depend on time-varying parameters, thus providing a special case of the more general multivariate M-models (Botella-Rocamora et al., 2015). If instead we assume $\sigma_1^2 = \cdots = \sigma_T^2 = \sigma^2$ and $\Omega_1 = \cdots = \Omega_T = \Omega$, the model reduces to the stationary AR(1) spatio-temporal model outlined by Martínez-Beneito et al. (2008). In this case, the joint distribution of Z then reduces to

$$Z \sim \mathcal{N}_{T \times n}\left(0, \frac{\sigma^2}{1-r^2} R \otimes \Omega\right)$$

where R is the AR(1) correlation matrix of size $T \times T$ with elements $R_{st} = r^{\text{abs}(s-t)}$ for $s, t \in [1, T]$. This constitutes a simplification of a more general multivariate spatial model in which k marginal variances and $k(k-1)/2$ correlations are employed. Since we have no prior information justifying a temporal variation in the parameters of the spatial component of the latent field, we follow this latter approach and assume the distribution of ν_t in equation 4.3 is stationary over time. This spatio-temporal model employs only one scale parameter σ^2 and one correlation parameter r, resulting in a simplified between-years variance-covariance matrix and a sparse between-years precision.

4.3.1 Spatial structures

The regression model outlined in the previous section features a spatial component described by the matrix Ω. In this study, we compare four alternative spatial models, corresponding to as many specifications of the matrix Ω:

1. Intrinsic Conditional Autoregressive (ICAR)

   $$\Omega := L^+ = (D - W)^+$$

   where L^+ is the Moore-Penrose pseudoinverse of the $n \times n$ Laplacian matrix of the graph $D - W$, W is the binary neighbourhood or adjacency matrix. To make this model identifiable, we constrain Z_t to sum to zero for each t.

2. Proper Conditional Autoregressive (PCAR)

   $$\Omega := (D - \rho W)^{-1}$$

Where $\rho \in [0, 1]$ is the spatial autocorrelation parameter, introduced to control the strength of spatial association; the limit case for $\rho = 1$ is the ICAR, while $\rho = 0$ implies an independent but not identically distributed Gaussian random field. Even though this model does not require the sum-to-zero constraint, we apply it in analogy with other models to make results comparable.

3. Leroux Proper Conditional Autoregressive (LCAR)

$$\Omega := (\lambda L + (1 - \lambda) I_n)^{-1}$$

Where $\lambda \in [0, 1]$ is the precision mixing parameter; the limit case for $\lambda = 1$ is again the ICAR, while for $\lambda = 0$ the model reduces to an IID Standard Normal field. Even though this is not a singular model, we follow Goicoa et al. (2018) and apply the sum-zero constraint to ensure that the model is identifiable and that no confounding occurs with the intercept.

4. Scaled Besag-York-Mollié model (BYM2)

$$Z_t = \sigma \sqrt{\phi} U_t + \sigma \sqrt{1 - \phi} V_t$$

Here U_t is an ICAR process with precision matrix given by the scaled Laplacian matrix, say L_s Sørbye and Rue (2014); Riebler et al. (2016), obtained multiplying L by a constant factor. This is required to ensure that L_s^+, representing the variance-covariance matrix of the ICAR component, is such that the geometric mean of the diagonal is equal to one; scaling allows σ^2 to be correctly identifiable and interpretable. Precision scaling has been described in Section 1.3.1. V_t is a standard IID process, such that $V_t \sim \mathcal{N}_n(0, I_n)$. As we have seen in section 1.4.3, $\Omega = \phi L_s^+ + (1 - \phi) I_n$. ϕ is the variance mixing parameter, and is interpreted as the share of variability explained by the spatial stochastic trend, while $1 - \phi$ is the share of variability explained by random noise. This interpretation is made possible by the scaling of the Laplacian matrix. As for the LCAR, the limit cases for $\phi = 1$ and $\phi = 0$ are the ICAR and IID processes respectively. Although the spatial precision matrix Ω^{-1} is dense, in section 1.4.3 we have seen that the joint distribution of (Z, U) has a sparse and singular precision matrix (Riebler et al., 2016) hence the sum-to-zero constraint is required on the ICAR components U_t for all t.

4.3.2 The PC-prior framework

The regression model outlined so far features a latent spatio-temporal effect, which implies a potentially high degree of complexity. In the context of Bayesian hierarchical models, an organic approach to penalising model complexity is provided by PC-priors (Simpson et al., 2017, a short overview of PC-priors is in section **??**); let us recall that to to penalise *a priori* the departure of a flexible model $\mathcal{M}_1$ against a base model $\mathcal{M}_0$, we assign an exponential prior to the

distance function between $\mathcal{M}_1$ and $\mathcal{M}_2$. The exponential prior depends on a rate parameter γ that can be tuned to assume *a priori* that θ is smaller than a boundary value U with a given probability α.

In the spatio-temporal application, we can define a joint PC-prior setup for the whole hyperparameter set. For the spatio-temporal ICAR model, hyperparameters are only the temporal autocorrelation r and the standard deviation σ. The latter three models employ an additional hyperparameter, either ρ, λ or ϕ. To the best of our knowledge the PC-priors for the LCAR precision mixing parameter λ and for the PCAR spatial autocorrelation ρ constitute a novelty; we then test the Uniform prior on γ as a more familiar alternative. Overall, the joint prior on the whole hyperparameter set reads as:

$$\pi(r, \sigma, \gamma) = \pi(r|\sigma, \gamma)\pi(\gamma|\sigma)\pi(\sigma)$$

As these hyperparameters prove to be independent *a priori*, we have shown that using the PC-priors allows us to factor the joint prior into the three marginals. The PC-prior for σ has been described by Simpson et al. (2017) and proved to coincide with an Exponential prior. In the following sections, we provide some details on the computation of the KLD for r, ρ, and λ in the AR(1), PCAR, and LCAR cases, respectively. In the BYM2 case, where $\gamma = \phi$, the PC-prior has already been described by both Riebler et al. (2016) and Simpson et al. (2017).

We obtain the KLD for the temporal autocorrelation parameter r in the AR(1) model (4.1) penalising $r \in [-1, 1]$ against $r = 0$. Hence the KLD becomes:

$$\text{KLD}(r||r = 0) = \frac{n}{2}\left(\ln(1 - r^2) + T\frac{r^2}{1 - r^2}\right) \tag{4.2}$$

The KLD is a U-shaped even function of r, with a minimum at 0 for $r = 0$ and tending to infinity for abs(r) approaching 1. These findings generalise the KLD of the purely temporal autocorrelation parameter developed and discussed by Sørbye and Rue (2017). Additionally, in the spatio-temporal case, the PC-prior can be proved to be independent of n. This is indeed a desirable property, as it would make no sense for the temporal autocorrelation parameter r to depend on the number of areas n. For the scope, consider $\text{KLD}(r||r = 0) = \frac{n}{2}h(r)$ where $h(r)$ does not depend on n; then, assuming an exp(γ) distribution on $\sqrt{2\text{KLD}(r||r = 0)}$ and parametrising the PC-prior through the left-tail probability α and the upper boundary U,we obtain $\gamma = \ln\frac{1}{1-\alpha}\frac{1}{\sqrt{nh(U)}}$. It follows that:

$$\pi(r) = \ln\frac{1}{1 - \alpha}\frac{1}{2\sqrt{h(U)h(r)}}e^{\ln(1-\alpha)\frac{\sqrt{h(r)}}{\sqrt{h(U)}}}\frac{\partial h}{\partial r}$$

where $h(r) = \ln\left(1 - r^2\right) + T\frac{r^2}{1-r^2}$ depends on the number of time periods T but not on any information provided by the spatial domain. In figure 4.2 we display this function for a fixed $U = 0.4$ and $\alpha \in \{0.6, 0.7, 0.8, 0.9\}$

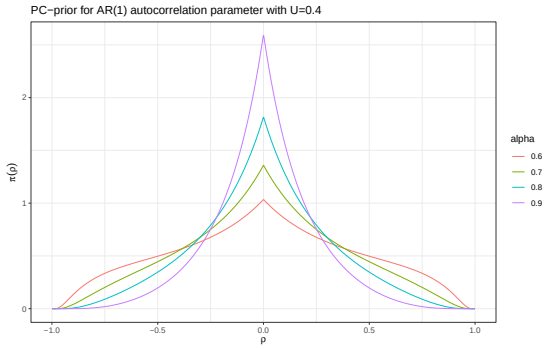

Figure 4.2: PC-prior for the AR(1) temporal autocorrelation in spatio-temporal areal models, with 4 time periods.

4.4 Model implementation

In this section, we first summarise the settings of the prior distributions of all model parameters, then we describe the practical procedure to estimate the space-time Poisson regression model described in section 4.3.

Constant covariate effects in β and time-dependent intercepts $\beta_{0,t}$ are assigned independent $\mathcal{N}(0, 10^3)$ priors. For the marginal standard deviation σ and the temporal correlation parameter r, we follow the PC-prior approach outlined in section 4.3.2. Specifically, for σ we assume $\text{Prob}(\sigma^2 \leq 1/2) = 0.9$, implying $\mathbb{E}[\sigma] \approx 0.307$ *a priori*. For the AR(1) autocorrelation r, we penalise the departure from the time-independent model while attempting not to be excessively restrictive, and assume $\text{Prob}(r \leq 0.4) = 0.8$. This parametrisation was also used, though in a different context, by Fioravanti et al. (2021). We acknowledge, however, that in our context the posteriors of σ^2 and r result to be scarcely sensitive to hyperparameter tuning. For the parameters ρ, λ and ϕ we tested three alternative prior specifications: the PC-prior with $\text{Prob}\{\gamma \leq 1/2\} = 2/3$, for $\gamma \in \{\rho, \lambda, \phi\}$, considered to be a reasonable standard for the BYM2 parameter ϕ and thus implemented by default in `R-INLA` (only for the BYM2), the more restrictive PC-prior $\text{Prob}\{\gamma \leq 0.6\} = 0.9$, and the Uniform prior in $[0, 1]$.

As is frequently the case in Bayesian applications, posterior inference cannot be obtained by analytical calculations. Therefore, we need to approximate the posterior distributions of all layers of model parameters. Considering that with the model at hand, covariate effects and the spatio-temporal latent field have a joint Gaussian prior and the predictor is linear, it is particularly convenient to employ the Integrated Nested Laplace Approximation (INLA Rue et al., 2009) for the purpose. All spatio-temporal models are implemented within the `rgeneric` environment (Gómez-Rubio, 2020, chapter 11) using package version `2025-06-22.1`.

Spatial patterns are observed for some of the explanatory variables. Moreover, we notice some major changes in the posterior distribution of the effect

of two covariates, i.e., ELI (share of low productivity workers) and ELL (low education incidence), as shown in the forest plot in Figure 4.3 and discussed in more detail in the next section. This suggests the presence of spatial confounding (Reich et al., 2006) between those covariates and unmeasured spatial effects. Therefore, following the simplified Spatial+ method developed by Urdangarin et al. (2024), we attempt to mitigate spatial confounding by filtering out these variables' spatial trends represented by the eigenvectors of the Laplacian matrix associated with the smallest eigenvalues (Von Luxburg, 2007). Part of the explanatory variables, i.e. the MFI components and the distance from help desks, are mapped in Appendix .1.

The four spatial structures for the space-time Poisson model presented in section 4.3, namely ICAR, PCAR, LCAR, and BYM2, are compared using the Watanabe-Akaike information criterion (WAIC hereinafter, Gelman et al., 2014), computed internally in `R-INLA`. The WAIC and its complexity component, i.e., the number of free parameters (P. eff), are shown in Table 4.1. For the α, λ, and ϕ parameters of PCAR, LCAR and BYM2, respectively, we compare the uniform prior and two alternative parametrisations of the PC-prior: PC1 with $\alpha = 2/3 \wedge U = 1/2$, and PC2 with $\alpha = 0.9 \wedge U = 0.6$. Differences in

Models	Prior	Base		S+ 15		S+ 20		S+ 25	
		WAIC	P. eff	WAIC	P. eff	WAIC	P. eff	WAIC	P. eff
ICAR		3928.35	255.65	3928.15	255.58	3927.85	255.42	3928.54	255.88
PCAR	Unif	3915.55	261.17	3914.08	261.02	3913.65	260.89	3914.69	261.33
PCAR	PC1	3915.11	261.54	3913.43	261.40	3913.12	261.23	3914.10	261.71
PCAR	PC2	3914.24	262.44	3912.51	262.19	3912.14	262.02	3913.15	262.52
LCAR	Unif	3910.67	261.08	3908.45	260.83	3908.02	260.69	3909.38	261.13
LCAR	PC1	3910.20	261.58	3907.86	261.28	3907.49	261.12	3908.83	261.57
LCAR	PC2	3910.02	261.76	3907.47	261.47	3907.18	261.31	3908.48	261.76
BYM2	Unif	3907.66	255.83	3906.71	255.84	3906.40	255.67	3908.28	256.26
BYM2	PC1	3908.00	256.04	3907.24	255.97	3906.90	255.82	3908.79	256.41
BYM2	PC2	3908.37	256.23	3907.53	256.14	3907.22	255.97	3909.07	256.59

Table 4.1: Model diagnostics using the Watanabe-Akaike criterion for four spatial models with three possible Spatial+ adjustments, consisting of removing either 15, 20 or 25 Laplacian eigenvectors from the covariates ELI and ELL. For the parameter controlling the spatial association in the PCAR, LCAR, and BYM2 models, we compare the Uniform prior and the PC prior with left-tail probability values of either 2/3 and 0.9, associated with upper boundaries of 0.5 and 0.6, respectively. The P_eff column consists of the number of free parameters, i.e., the complexity component.

model performances do not appear relevant; the WAIC leads us to prefer the BYM2 model over the others. Neither does the treatment of spatial confounding change radically model performance, but a small improvement can be noticed if we filter out the 20 smallest-frequency eigenvectors.

4.5 Results

In this section, we display the main results of the alternative model specifications described in Section 4.4, focusing on the BYM2 with $\phi \sim \text{Unif}(0,1)$ after removing the 20 lowest-frequency Laplacian eigenvectors (applying the simplified Spatial+ correction proposed by Urdangarin et al., 2024) from two covariates: the share of employees in low-productivity firms (ELI) and the low education incidence (ELL). All explanatory variables have been scaled to zero mean and unit variance to facilitate the interpretation of results.

In Figure 4.3 we show a forest plot of the 95% posterior credible intervals of covariate effects using the non-spatial and spatial models before and after removing the 20 lowest-frequency Laplacian eigenvectors from the covariates ELI and ELL. Such a correction is justified by the difference in the posterior distributions of these two variables between the non-spatial and spatial models, which suggests the potential confounding with the latent spatial field. In particular, including the spatial field shrinks the estimated effect of ELI towards zero, while it raises in absolute value the estimated effect of ELL. Generally speaking, the inclusion of a spatial latent field increases uncertainty in the estimation of covariate effects, as can be seen from the wider credible intervals. The most noticeable differences among spatial models can be observed for ELL and the employment rate (ER); specifically, in both cases, controlling for the spatial confounding in ELI and ELL slightly reduces the covariate effect. Overall, the estimation of covariate effects does not appear to be sensitive to the spatial model choice or to spatial confounding treatment. In Table 4.2 we show more in detail the posterior summaries for intercepts and covariate effects using the BYM2 model and filtering out 20 eigenvectors from ELI and EL. To assess

Effect	mean	sd	0.025quant	0.975quant
Year 2021	-7.453	0.051	-7.555	-7.353
Year 2022	-7.583	0.055	-7.691	-7.477
Year 2023	-7.216	0.047	-7.310	-7.124
Year 2024	-7.328	0.051	-7.428	-7.229
AVC_dist	-0.196	0.042	-0.278	-0.113
Desk_dist	-0.061	0.038	-0.135	0.012
ELI	-0.002	0.035	-0.069	0.067
PGR	0.071	0.042	-0.011	0.153
UIS	0.095	0.040	0.018	0.174
ELL	-0.155	0.037	-0.229	-0.083
PDI	-0.024	0.046	-0.114	0.065
ER	-0.229	0.053	-0.333	-0.126

Table 4.2: Posterior summaries of year-specific intercepts and covariate effects when Z is modelled as a BYM2 process with a Uniform prior on the mixing parameter. Simplified spatial+ correction has been applied to ELI and ELL covariates by removing the 20 lowest-frequency Laplacian eigenvectors.

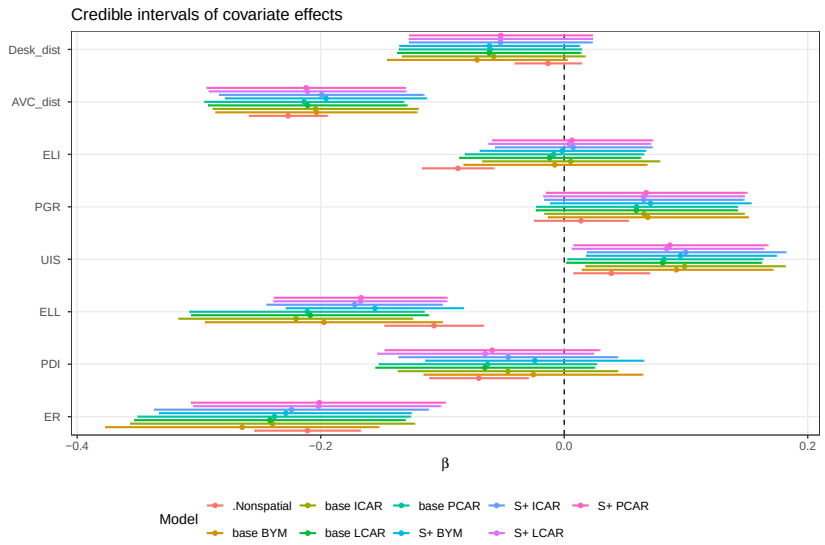

Figure 4.3: Forest-plot of the posterior credible intervals of covariate effects using either the non-spatial model and the four spatio-temporal models, before ("base" models) and after filtering out the 20 lowest-frequency Laplacian eigenvectors from covariates ELI and EL ("S+" models). For the ρ and λ parameters of the PCAR and LCAR models respectively, a PC-prior parametrised by boundary value and left-tail probability of 0.6 and 0.9 has been used, while a Uniform prior has been assigned to ϕ in the BYM2 model.

the effects of explanatory variables on AVC accesses, it is needful to recall the twofold interpretation of the response variable, as it depends on both the frequency at which violence incidents occur and the frequency at which victims seek help when it occurs. The three strongest predictors across the years in scope appear to be the travel time from the closest municipality hosting an AVC (AVC_dist), the percentage of people with a low educational level (ELL), and the employment rate (ER), in all cases with a negative association. Additionally, the density of productive units (UIS) has a positive effect, and the credibility interval of such effect ranges entirely away from zero.

The negative association between AVC accesses and distance from AVCs can be read as a symptom of the persisting physical barrier to seeking help. The magnitude of the coefficient may be interpreted as follows: an increase in the travel time to the closest AVC by one standard deviation, i.e. approximately $13'40''$ is associated, *ceteris paribus*, with a decrease of about 17.7% in the frequency of accesses to AVCs (since $\exp\{E\left[\beta_{\text{AVC\_Dist}}|y\right]\} = 0.823$). The expected association with the distance from the closest help desk (Desk dist) is negative, consistent with the distance from the closest AVC, but the magnitude of such expected association is smaller, and, most importantly, the credible interval in-

cludes zero. This result may suggest that the distance from help desks is less prohibitive than the distance from AVCs in seeking assistance, which is not surprising considering how widespread help desks are. In light of the twofold process generating AVC accesses, a tentative yet intuitive interpretation of the negative association with both low education and employment rate could be that women victims of violence have a lower propensity to access AVCs in poor educational contexts, but the incidence of violence may become lower in contexts of higher economic development. This latter consideration should be interpreted in light of the state-of-the-art evidence that both women's and partners' unemployment are associated with higher violence occurrence, as reviewed by (Bettio and Ticci, 2017, chapter 2). The positive association of accesses with the density of productive units (UIS), instead, is less intuitive, as a higher number of firms does not imply *per se* higher economic development, even though it has a negative weight in determining the frailty index of Italian municipalities.

Lastly, hyperparameter posterior summaries for the BYM2–uniform model after ELI and ELL are corrected for spatial confounding are shown in Table 4.3. Residual variability is balanced between spatial autocorrelation and random noise, as the variance mixing parameter has a posterior expectation of ≈ 0.6, which means that the share of residual variability explained by the spatial structure is slightly larger. The standard deviation of the spatio-temporal latent field has a posterior expectation of ≈ 0.47. This can be interpreted as an increment of 1 standard deviation of the latent field in a municipality implies, *ceteris paribus*, a change of approximately $\exp(0.47) - 1 \approx 59\%$ in the frequency of AVC accesses. Lastly, the temporal autocorrelation is positive, as one could expect by comparing the spatial patterns in AVC accesses in Figure 4.1.

param	mean	sd	Q0.025	Median	Q0.975
ϕ	0.602	0.117	0.358	0.609	0.808
σ	0.468	0.046	0.383	0.466	0.563
r	0.622	0.059	0.499	0.625	0.729

Table 4.3: Posterior summaries of the hyperparameters of the BYM2 latent effect when $\phi \sim \text{Unif}(0, 1)$ *a priori*, where ϕ is the variance mixing parameter, σ is the marginal standard deviation, and r is the temporal autocorrelation.

4.6 Discussion and conclusions

The main limitation of this research is that we employ a unique predictor combining the two processes of violence occurring and help seeking after violence occurs, which cannot be disentangled with the available data. Consequently, the effects we estimate reflect both mechanisms at once. Additional limitations include the lack of individual-level information for privacy reasons, the unavailability of service-level performance indicators, and the constraints posed by aggregated data that can mask within-area heterogeneity. A promising development would be the implementation of a Poisson-logistic regression model

(Stoner et al., 2019), in line with the analysis of Polettini et al. (2024), to allow the explicit quantification of VAW under-reporting. To this end, more integrated datasets are required combining administrative, health, social-service and AVC-level information, thereby enabling the identification of respective contributions of structural incidence, help-seeking propensity and territorial accessibility, and allowing the two processes of occurrence and help-seeking to be modelled separately.

Overall, present results show that seeking support is far from a straightforward process: the negative association between access rates and proximity to the nearest AVC suggests that even modest distances can act as barriers to help-seeking. Moreover, education-related vulnerability and local economic conditions appear to affect the likelihood of accessing AVCs, suggesting that structural factors, cultural attitudes, and service accessibility interact in shaping women's behaviour. These findings are in line with the analysis of Gracia et al. (2020), who gather solid evidence that education is one of the key factors corresponding to a higher propensity to recognise and report intimate partner violence.

Even after adjusting for potential spatial confounding, a residual spatial effect remains, pointing to unobserved factors that may influence women's ability to activate support pathways. Beyond the empirical findings, the methodological contribution of this work lies in the derivation of PC-priors for the LCAR mixing parameter λ and the PCAR spatial autocorrelation parameter ρ. These priors, obtained via explicit KLD penalisation, may contribute to improving interpretability and stabilising the INLA-based inference.

.1 Maps of Explanatory Variables

Here we display some additional maps, i.e. the components of the Municipality Frailty Index (MFI) included as explanatory variables in the analytical model described in section 4.3, and the distance from municipalities hosting a help desk, which equals to zero for about 40% of municipalities (105 over 256), as they have at least one. The population growth rate map does not include the municipality of Anzano di Puglia (FG) due to its extreme record (-311%)

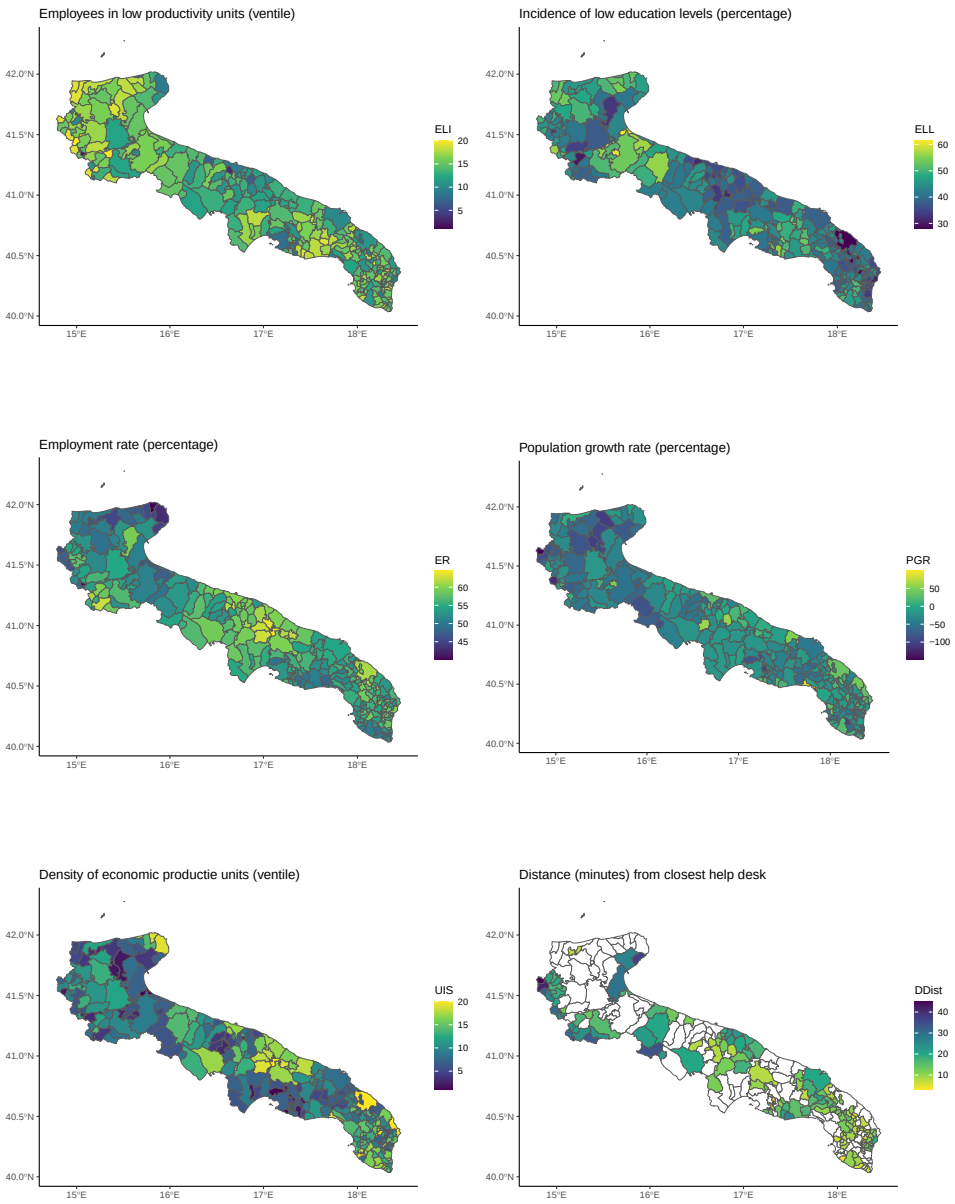

Figure 4: Additional explanatory variables: MFI components and distance from help desks. White areas in the last map correspond to municipalities in which a desk is available (distance is zero).

Bibliography

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

Agasisti, T. and Vittadini, G. (2012). Regional economic disparities as determinants of student's achievement in italy. *Research in Applied Economics*, 4(2):33.

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

Anselin, L. (1988). *Spatial Econometrics: methods and models*, volume 4. Springer Science & Business Media.

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

Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*, volume 3. Cambridge University Press.

Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). inlabru: an r package for bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766.

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

Banerjee, S., Carlin, B., and Gelfand, A. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2nd edition.

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

Barrett, P., Treves, A., Shmis, T., and Ambasz, D. (2019). The impact of school infrastructure on learning: A synthesis of the evidence. *World Bank Publications*.

Battagliese, D. (2020). Extensions of the univariate pc prior. *Annali del Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza (MEMOTEF)*, pages 33–46.

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

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

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20.

Bettin, G. and Ciaschini, C. (2025). The effectiveness of italian measures in facilitating complaints for domestic violence: a regional analysis. *Kybernetes*, pages 1–21.

Bettio, F. and Ticci, E. (2017). *Violence and Economic Independence*. Publication Office of the European Union.

Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., and Blanchet, G. (2017). Package 'spdep': Spatial dependence: Weighting schemes, statistics, r package version. R package version 1.3-7.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blatchford, P., Bassett, P., and Brown, P. (2011). 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.

Botella-Rocamora, P., Martinez-Beneito, M. A., and Banerjee, S. (2015). A unifying modeling framework for highly multivariate disease mapping. *Statistics in Medicine*, 34(9):1548–1559.

Bratti, M., Checchi, D., and Filippin, A. (2007). Territorial differences in italian students' mathematical competencies: evidence from pisa 2003. *Giornale Degli Economisti e Annali Di Economia*, 66(3):299–333.

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

Bucci, M., Gazzano, L., Gennari, E., Grompone, A., Ivaldi, G., Messina, G., and Ziglio, G. (2023). 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.

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

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

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models and Applications*. Pion, London.

Corpas-Burgos, F., Botella-Rocamora, P., and Martinez-Beneito, M. (2019). On the convenience of heteroscedasticity in highly multivariate disease mapping. *Test*, 28(4):1229–1250.

Council of Europe (2011). Convention on preventing and combating violence against women and domestic violence. volume 11.V.2011, Istanbul. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:22023A0602(01).

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

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.

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

Dean, C., Ugarte, M., and Militino, A. (2001). Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics*, 57(1):197–202.

Denti, D. and Iammarino, S. (2022). Coming out of the woods. do local support services influence the propensity to report sexual violence? *Journal of Economic Behavior & Organization*, 193:334–352.

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

Dupont, E., Marques, I., and Kneib, T. (2023). Demystifying spatial confounding. *arXiv preprint*, pages 1–36.

Dupont, E., Wood, S. N., and Augustin, N. H. (2022). Spatial+: a novel approach to spatial confounding. *Biometrics*, 78(4):1279–1290.

Dutta, S., van Niekerk, J., and Rue, H. (2025). Scalable skewed bayesian inference for latent gaussian models. *arXiv preprint*.

Eurostat (2024). Nuts - nomenclature of territorial units for statistics. last access December 17th 2024.

Fattah, E. A., Van Niekerk, J., and Rue, H. (2022). Smart gradient-an adaptive technique for improving gradient estimation. *Foundations of Data Science*, 4(1):123–136. arXiv = https://arxiv.org/abs/2106.07313.

Fioravanti, G., Martino, S., Cameletti, M., and Cattani, G. (2021). Spatio-temporal modelling of pm10 daily concentrations in italy using the spde approach. *Atmospheric Environment*, 248:118192.

Freni-Sterrantino, A., Rustand, D., van Niekerk, J., Teixeira Krainski, E., and Rue, H. (2025). A graphical framework for interpretable correlation matrix models for multivariate regression: A. freni-sterrantino et al. *Statistical Methods & Applications*, pages 1–39.

Freni-Sterrantino, A., Ventrucci, M., and Rue, H. (2018). A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and spatio-temporal epidemiology*, 26:25–34.

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

Gaedke-Merzhäuser, L., van Niekerk, J., Schenk, O., and Rue, H. (2023). Parallelized integrated nested laplace approximations for fast bayesian inference. *Statistics and Computing*, 33(1):25.

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

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

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian statistics 4*, pages 147–168.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016.

Giancola, O., Benadusi, L., Fornari, R., et al. (2010). Così vicine, così lontane. la questione dell'equità scolastica nelle regioni italiane. *Scuola democratica*, 1:52–79.

Giancola, O. and Salmieri, L. (2020). Family background, school-track and macro-area: the complex chains of education inequalities in italy. working paper.

Goicoa, T., Adin, A., Ugarte, M., and Hodges, J. (2018). In spatio-temporal disease mapping models, identifiability constraints affect pql and inla results. *Stochastic Environmental Research and Risk Assessment*, 32(3):749–770.

Gracia, E., Lila, M., and Santirso, F. A. (2020). Attitudes toward intimate partner violence against women in the european union: A systematic review. *European Psychologist*.

Griffith, D. A. (2010). The moran coefficient for non-normal data. *Journal of Statistical Planning and Inference*, 140(11):2980–2990.

Gómez-Rubio, V. (2020). *Bayesian Inference with INLA*. Chapman and Hall/CRC.

Hammersley, J. and Clifford, P. (1971). Markov fields on finite graphs and lattices. *Unpublished Manuscript*.

Held, L., Schrödle, B., and Rue, H. (2010). Posterior and cross-validatory predictive checks: a comparison of mcmc and inla. In Kneib, T. and Tutz, G., editors, *Statistical Modelling and Regression Structures*, pages 91–110. Physica-Verlag HD.

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

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

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press, 2nd edition.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(1):139–159.

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

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

ISTAT - Italian National Institute of Statistics (2022). La geografia delle aree interne nel 2020: vasti territori tra potenzialità e debolezze.

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

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

Italian Ministry of Economic Development (2020). Decree of july 7th 2020.

Italian Ministry of Education (2023). Ministerial decree No. 90 of May 19th 2023.

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

Italian National Institute of Statistics (2024). I centri antiviolenza e le donne che hanno avviato il percorso di uscita dalla violenza. anno 2023 (italian).

Italian Official Journal (2007). Law No. 176 of October 25th 2007.

Italian Official Journal (2009). Decree of the President of the Republic No. 81 of March 20th 2009.

Italian Official Journal (2015). Law no. 107 of july 15th 2015.

Italian Official Journal n. 242 (2013). Disposizioni urgenti in materia di sicurezza e per il contrasto della violenza di genere, nonchè in tema di protezione civile e di commissariamento delle province. Gazzetta Ufficiale 24.05.2022. Retrived from https://www.normattiva.it/. Accessed on 31 January 2025.

Italian Official Journal n.120 (2022). Disposizioni in materia di statistiche in tema di violenza di genere. Gazzetta Ufficiale 24.05.2022. Retrived from https://www.normattiva.it/. Accessed on 31 January 2025.

Johnson, R. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, 6th edition.

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

Kabe, D. G. (1964). A note on the bartlett decomposition of a wishart matrix. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):270–273.

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

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, 19(17-18):2555–2567. doi: 10.1002/1097-0258.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.

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

Leroux, B., Lei, X., and Breslow, N. (1999). Exact estimation of disease in small areas: A new mixed model for spatial dependence. In Halloran, M. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment and Clinical Trials*, pages 135–178. Springer-Verlag, New York.

Lindgren, F., Bachl, F., Illian, J., Suen, M. H., Rue, H., and Seaton, A. E. (2024). inlabru: software for fitting latent gaussian models with non-linear predictors. *arXiv preprint*.

Lindgren, F., Rue, H., and Lindström, J. (2011). 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.

Lindley, D. and Smith, A. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34:1–41.

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

Marques, I. and Wiemann, P. F. (2023). Bayesian spatial+: A joint model perspective. *arXiv preprint*, arXiv:2309.05496:1–48.

Martinez-Beneito, M. and Botella-Rocamora, P. (2019). *Disease Mapping: From Foundations to Multidimensional Modeling*. Chapman & Hall/CRC, Boca Raton.

Martínez-Beneito, M. A., López-Quilez, A., and Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in medicine*, 27(15):2874–2889.

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

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

Munzert, S., Rubba, C., Meißner, P., and Nyhuis, D. (2015). *Automated data collection with R: A practical guide to web scraping and text mining*, volume 1. Wiley West Sussex.

Müller, K. and Wickham, H. (2023). tibble: Simple data frames.

OECD (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. OECD Publishing, Paris.

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

Ottomano Palmisano, G., Sardaro, R., and La Sala, P. (2022). Recovery and resilience of the inner areas: identifying collective policy actions through promethee ii. *Land*, 11(8):1181.

Palmí-Perales, F., Gómez-Rubio, V., Bivand, R. S., Cameletti, M., and Rue, H. (2023). Bayesian inference for multivariate spatial models with inla.

Palmí-Perales, F., Gómez-Rubio, V., and Martinez-Beneito, M. A. (2021). Bayesian multivariate spatial models for lattice data with inla. *Journal of Statistical Software*, 98(2):1–29.

Peek-Asa, C., Wallis, A., Harland, K., Beyer, K., Dickey, P., and Saftlas, A. (2011). Rural disparity in domestic violence prevalence and access to resources. *Journal of women's health*, 20(11):1743–1749.

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

Polettini, S., Arima, S., and Martino, S. (2024). An investigation of models for under-reporting in the analysis of violence against women in italy. *Social Indicators Research*, 175(3):1007–1026.

R Core Team (2023). R: A language and environment for statistical computing.

R Studio team (2021). Rstudio: Integrated development environment for r.

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

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165.

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

Rue, H., Martino, S., and Chopin, N. (2009). 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.

Rue, H., Riebler, A., Sørbye, S., Illian, J., Simpson, D., and Lindgren, F. (2017). Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Sanchez, S. M. and Sanchez, P. J. (2005). Very large fractional factorial and central composite designs. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 15(4):362–377.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1 – 28.

Sørbye, S. H. and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38(6):923–935.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Statistical Office of the Apulia Region (2024). Centri antiviolenza e Case rifugio in Puglia. L'accesso e l'accoglienza delle donne. Anno 2023 (EN: Anti-violence centers and shelters in Apulia. Women access and reception. Year 2023).

Stoner, O., Economou, T., and Drummond Marques da Silva, G. (2019). A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association*, 114(528):1481–1492.

Sørbye, S. and Rue, H. (2014). Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51.

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

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.

Toffanin, A., Pietrobelli, M., Gadda, A., and Misiti, M. (2020). VAW Policy Regimes in Italy: An Analysis Across Regional Governments and Women's Centres. *Journal of Mediterranean Knowledge*, 5(1):47–72.

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

Urdangarin, A., Goicoa, T., and Ugarte, M. D. (2023). Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*, 36(2):333–360.

Van Niekerk, J., Krainski, E., Rustand, D., and Rue, H. (2023). A new avenue for bayesian inference with inla. *Computational Statistics and Data Analysis*, 181.

Van Niekerk, J. and Rue, H. (2021). Skewed probit regression - identifiability, contraction and reformulation. *Revstat - Statistical Journal*, 19:1–22.

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

Ver Hoef, J. M., Hanks, E. M., and Hooten, M. B. (2018). On the relationship between conditional (car) and simultaneous (sar) autoregressive models. *Spatial statistics*, 25:68–85.

Vicente, G., Adin, A., Goicoa, T., and Ugarte, M. D. (2023). High-dimensional order-free multivariate spatial disease mapping. *Statistics and Computing*, 33(5):104.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of statistical planning and inference*, 121(2):311–324.

Wang, X., Yue, Y. R., and Faraway, J. J. (2018). *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.

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

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, pages 434–449.

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

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

Wickham, H. and Bryan, J. (2023). *R packages*. "O'Reilly Media, Inc.".

Zellner, A. (1988). Optimal information processing and bayes's theorem. *The American Statistician*, 42(4):278–280.