Specification of spatiotemporal interactions in Bayesian hierarchical models

Project thesis for the course TMA4500 at NTNU

Aanes Jostein Aastebøl

Institutt for matematiske fag — NTNU 2023

Abstract

The Bayesian hierarchical framework is frequently used to model spatiotemporal processes. In this framework, spatiotemporal interactions are employed to account for variations in spatiotemporal processes that cannot be separated into effects of time or space. The best way to specify spatiotemporal interactions has not yet been determined. My project thesis serves as a prelude to my master thesis, in which I aim to provide a conclusive answer to this question. Specifically, the goal of my project thesis is to gain a thorough understanding of Bayesian hierarchical spatiotemporal models, review the relevant literature, and compare common and relevant models using an appropriate real-world data set. The data set analyzed is widely used and popular, containing the deaths caused by lung cancer in the counties of Ohio during the years 1968 to 1988. Two distinct approaches for the specification of spatiotemporal interactions are considered, and a total of thirteen models are constructed. The models were compared based on their ability to estimate the death rate from lung cancer using different model choice criteria. Additionally, one-year ahead predictions of lung cancer deaths were also evaluated. The best model for estimating the rate of deaths caused by lung cancer was a common model that used improper spatial and temporal effects and improper spatiotemporal interactions smoothing over space and time. The best model for one-year ahead predictions consisted of a fixed effect for time and proper spatiotemporal interactions smoothing over space and time.

Contents

1 Introduction: Why do we want to specify spatiotemporal interactions?

Many fields of study deal with processes that develop in space and time such as the study of diseases in epidemiology and weather phenomena in climatology. A spatiotemporal process is a process developing in both space and time, where both space and time are interesting and relevant aspects of the process in question. The interest in the analysis of spatiotemporal processes has grown over the last decades as a result of the ever-increasing availability of data and improved computational capacity. It is common to use a Bayesian hierarchical framework to model spatiotemporal processes in many fields, such as epidemiology [Richardson *and others*, 2006; Wakefield *and others*, 2019; Schrödle and Held, 2011a; Coly *and others*, 2021], climatology [J. A. García and Acero, 2018], and economy [Zhang *and others*, 2018]. Bayesian hierarchical models are often used because certain priors utilize the temporal and spatial structure of the problem to 'borrow strength' between areas and times, which enables the model to smooth estimates in space and time. Research on spatiotemporal processes is often interested in uncovering the spatiotemporal patterns of the process, such as temporal trends, persistent patterns in space, clusters in space, and regions developing differently over time compared to other regions. These patterns are interesting because they give insight into how space and time affect the process.

UNICEF's report on under-five child mortality

An example of research on a spatiotemporal process is UNICEF's report on sub-national estimates of under-five child mortality rates (U5MR) for developing countries from 1990 to 2019 [UNICEF, 2021]. The report focuses on sub-national estimates because it would be limiting to only look at national mortality rates. A nationwide reduction in child mortality can hide sub-national regions with disproportionately high U5MR. An example of this is Kenya which according to UNICEF [2021] has experienced a 57% nationwide reduction in U5MR from 1990 until 2019, while for administrative regions within Kenya changes in U5MR ranged from 81% reduction to a 32% increase in U5MR. In UNICEF [2021] they are interested in discovering the temporal trends of U5MR for the different countries and the variation in U5MR within the countries themselves. The recorded under-five child death counts vary greatly, and smoothing is necessary to provide accurate sub-national estimates of U5MR. Therefore a spatiotemporal Bayesian hierarchical model is employed that has smooth effects in time and space, making the estimates of the rate more accurate as the effects 'borrow strength' between times and areas. Finding the sub-national estimates of U5MR enables UNICEF to direct aid towards hot spots for child mortality. Additionally, authorities and researchers can determine if a measure in a region is effective at lowering U5MR based on the temporal trend of U5MR after the enactment of the measure. Thus, a country in the study can see in what areas there is a need for further aid, and also learn which measures are efficient at reducing U5MR. In Figure 1, the change of U5MR within Kenya between 1990 and 2019 is illustrated [UNICEF, 2021]. By looking at Figure 1, we can see that measures in the northeast of Kenya have had a positive impact on reducing child mortality rates. For the central regions, whatever

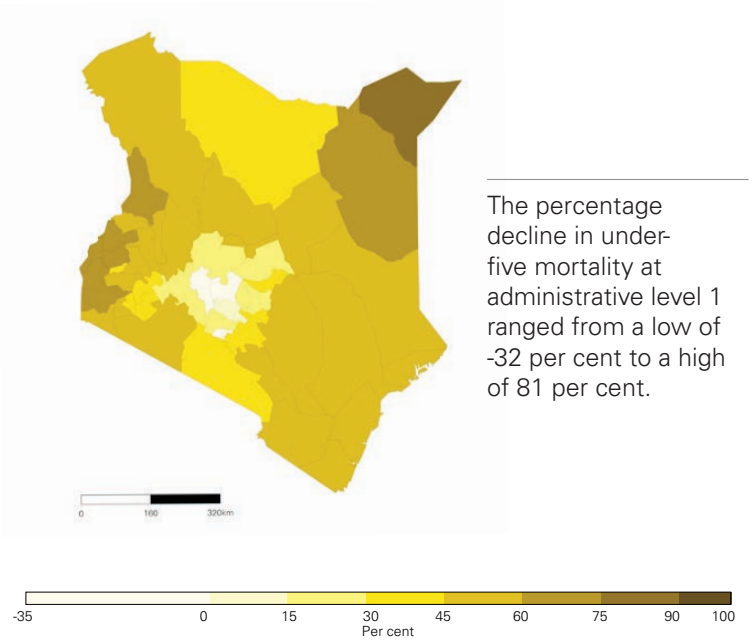


Figure 1: Estimated change in U5MR in percentage within Kenya from 1990 to 2019. The Figure was extracted from UNICEF [2021] and was part of UNICEF's work to estimate sub-national under-five child mortality rates in developing countries.

measures or lack thereof, have increased U5MR. Local authorities for the central regions in Kenya can use these results to adjust their policies and enact measures having reduced U5MR in other regions of Kenya. All of this relies upon accurate estimates of U5MR, something that spatiotemporal Bayesian hierarchical models can provide. UNICEF's report on sub-national estimates of U5MR for developing countries is a concrete example of how Bayesian hierarchical spatiotemporal models enable researchers, authorities, and organizations to better deal with the greatest challenges of the 21st century. The United Nations committed itself to 17 sustainable development goals in 2015, which are aimed at ending poverty, protecting our planet, and that everyone lives in peace and health by 2030 [UN, 2015]. The example of sub-national U5MR estimates in UNICEF [2021] shows how Bayesian hierarchical spatiotemporal models can play an important part in reaching the third sustainable development goal: "good health and well-being for all" [UN, 2015].

Motivation for the project thesis

A common way to model a spatiotemporal process is through spatial and temporal random effects and spatiotemporal interactions [Richardson *and others*, 2006; Schrödle and Held, 2011a; Ugarte *and others*, 2014]. The random effects and interactions encapsulate unobserved properties of the process related to the areas or time points of the considered process. Only using spatial and temporal random effects would be limiting as it would assume that the variation of the process is separable into effects of space and time.

Therefore spatiotemporal interactions are crucial for Bayesian hierarchical spatiotemporal models. This leads to the question: How should spatiotemporal interactions be specified?
In Knorr-Held [2000] four different specifications of spatiotemporal interactions were introduced. These four specifications have become a common way to specify spatiotemporal interactions [Schmid and Held, 2004; Schrödle and Held, 2011b; Ugarte *and others*, 2014]. There are other alternative specifications of spatiotemporal interactions, such as those introduced in Vivar and Ferreira [2009] and Utazi *and others* [2018]. Currently, there is no publication having compared the different spatiotemporal interactions, which have come to a decisive conclusion on how to best specify the spatiotemporal interactions. This lack of a conclusive study is the motivation behind this project thesis and subsequently my master thesis.

Overview of the project thesis
This project thesis prepares and begins the work aiming at providing a conclusive study on how to best specify spatiotemporal interactions. In this project thesis, the goal is to gain background knowledge, get familiar with relevant models, using integrated nested Laplace approximations for inference, and finally compare common and relevant spatiotemporal models on an appropriate real-world data set.

The thesis starts by introducing relevant theory on Gaussian Markov Random Fields (GMRF) in Section 2 and intrinsic Gaussian Markov Random Fields (IGMRF) in Section 3. Relevant theory on GMRFs and IGMRFs is necessary as the components in the considered spatiotemporal models have either GMRFs or IGMRFs as priors. Common temporal models are presented in Section 4, followed by Section 5 presenting common spatial models. The temporal and spatial models presented build upon the theory of GMRFs and IGMRFs, and they are used to model temporal and spatial random effects in the considered spatiotemporal models. In Section 6 the need for scaling IGMRFs is shown and the penalized complexity priors, which are used extensively in this project thesis, are presented. Having detailed relevant theory, common temporal and spatial models, scaling of IGMRFs, and penalized complexity priors used for the hyperparameters, the necessary preliminaries are in place to introduce the spatiotemporal models in Section 7. In Section 7 common spatiotemporal models using spatiotemporal interactions as introduced by Knorr-Held [2000] are presented and motivated. Additionally in Section 7 alternative spatiotemporal models which specify the spatiotemporal interactions in a different way than Knorr-Held [2000] are also presented. After having presented all spatiotemporal models that will be considered in this project, Section 8 introduces a popular and widely used data set of lung cancer death counts in the counties of Ohio from 1968 until 1988. This data set will be used to compare the performance of the different models. Section 8 also presents how the models are compared using model choice criteria and the methods used to compare their one-year ahead predictive performance. Lastly in Section 8 the results of the different models applied to the Ohio lung cancer data set are reported and discussed. This project thesis is concluded in Section 9.

2 Gaussian Markov Random Fields

GMRFs and IGMRFs are frequently used in state-of-the-art spatiotemporal modeling [Richardson *and others*, 2006; Wakefield *and others*, 2019; Schrödle and Held, 2011a]. Therefore it is a necessary preliminary to define GMRFs and IGMRFs and present some of their properties before introducing spatiotemporal models. This section presents GMRFs and the next section presents IGMRFs. Both GMRFs and IGMRFs are defined on graphs, so this section begins by defining an undirected graph. Then the definition of a GMRF is provided followed by a presentation of relevant properties of a GMRF.

2.1 Undirected graphs

The tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is called an undirected graph when $\mathcal{V} = \{1, ..., n\}$ is a set of nodes and $\mathcal{E}$ is a set of undirected edges between the nodes $\{i, j\}$ where $i, j \in \mathcal{V}$ and $i \neq j$.
An example of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, 3, 4\}$ and $\mathcal{E} = \{\{1, 2\}, \{2, 3\}, \{2, 4\}\}$, is provided in Figure 2. Having defined an undirected graph, the definition of a GMRF on an undirected graph is now provided.

2.2 Definition of a GMRF

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with nodes $\mathcal{V} = \{1, ..., n\}$. Following the definition provided by Rue and Held [2005, Section 2]: A random vector $\boldsymbol{x}^\top = (x_1, ..., x_n) \in \mathbb{R}^n$, with mean $\boldsymbol{\mu}$ and symmetric positive definite precision matrix $\boldsymbol{Q}$, is a GMRF with respect to $\mathcal{G}$ if and only if

$$\pi(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{Q}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})\right) \tag{1}$$

and

$$Q_{ij} \neq 0 \iff \{i, j\} \in \mathcal{E} \text{ for } i \neq j \tag{2}$$

Remarks:
The precision matrix $\boldsymbol{Q}$ is the inverse of the covariance matrix $\boldsymbol{Q}^{-1}$.
Equation (1) implies that $\boldsymbol{x}$ follows a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q}$. If this were not the case, it would not be a *Gaussian* Markov

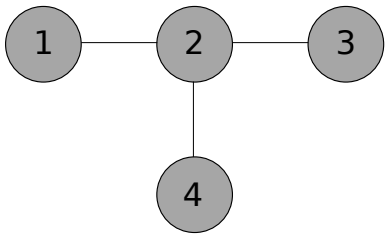


Figure 2: Illustration of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of nodes $\mathcal{V} = \{1, 2, 3, 4\}$ and undirected edges $\mathcal{E} = \{\{1, 2\}, \{2, 3\}, \{2, 4\}\}$.

Random Field.
Equation (2), states than an entry, Q_{ij} where $i \neq j$, in $\boldsymbol{Q}$ is non-zero if and only if the undirected graph $\mathcal{G}$ has an edge between the corresponding nodes i and j. This means that a GMRF with respect to the undirected graph in Figure 2 has a precision matrix in which $Q_{1,2} \neq 0$ as there is an edge between nodes 1 and 2, whereas $Q_{1,4} = 0$ as there is no edge between nodes 1 and 4.

2.3 Properties of GMRFs

For the remainder of this section let $\boldsymbol{x}^\top = (x_1, ..., x_n)$ be a GMRF, with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q}$, with respect to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
From Equation (1), it follows that

$$Q_{ij} \neq 0 \iff x_i \not\perp x_j | \boldsymbol{x}_{-ij} \tag{3}$$

which implies that an entry Q_{ij} in $\boldsymbol{Q}$ is non-zero if and only if x_i and x_j are conditionally dependent given the other elements $\boldsymbol{x}_{-ij} = \boldsymbol{x} \setminus \{x_i, x_j\}$. If x_i and x_j are conditionally independent, then

$$\pi(x_i | x_j, \boldsymbol{x}_{-ij}) = \pi(x_i | \boldsymbol{x}_{-ij})$$

One way to view conditional independence is that x_j does not provide any new information about x_i if $\boldsymbol{x}_{-ij}$ is known. From Equations (2) and (3) we get that for $i \neq j$, x_i and x_j are conditionally dependent given $\boldsymbol{x}_{-ij}$ if there is an edge between nodes i and j, and conditionally independent if there is no edge between the nodes. This means that for $\boldsymbol{x}$ a GMRF with respect to the graph in Figure 2, $x_1 \perp x_4 | x_2, x_3$ and $x_1 \not\perp x_2 | x_3, x_4$.

If each node in an undirected graph has relatively few edges to other nodes, a GMRF with respect to this graph will have a precision matrix with many zero entries. As a result $\boldsymbol{Q}$ is sparse, enabling the use of sparse matrix algorithms. This makes the computation of the determinant $|\boldsymbol{Q}|$ or solving linear systems involving $\boldsymbol{Q}$ computationally efficient. In Equation (1) there are two major bottlenecks to computation and those are the computation of $|\boldsymbol{Q}|$ and the calculation of $(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{\mu})$. With a sparse $\boldsymbol{Q}$, these computations can be done a lot faster than for dense matrices.

Another reason GMRFs are popular is that we can incorporate prior beliefs about the dependency structure of $\boldsymbol{x}$ in an intuitive manner. Since GMRFs are defined with respect to graphs where edges represent conditional dependence, it is fairly easy to understand the dependency structure of a GMRF given the graph. One only has to look at the edges between the nodes. Since it is quite easy to define graphs for temporal and spatial structures, it's easy to find relevant dependency structures in these cases. In Rue and Held [2005, Section 2] it is shown that the full conditional expectation of x_i is given by

$$\mathbb{E}[x_i | \boldsymbol{x}_{-i}] = \mu_i - \frac{1}{Q_{ii}} \sum_{j : j \sim i} Q_{ij} (x_j - \mu_j) \tag{4}$$

where $j \sim i$ means that $Q_{ji} \neq 0$ for $j \neq i$. This means that the full conditional mean of x_i is influenced by the value and mean of the entries x_j which x_i is conditionally dependent on. In addition, the full conditional precision is given as

$$\text{Prec}[x_i|\boldsymbol{x}_{-i}] = Q_{ii} \tag{5}$$

According to Rue and Held [2005, Section 2] it is possible to define a GMRF implicitly through the full conditionals given in Equations (4) and (5). Hence, it is fairly simple to define a GMRF intuitively.

Lastly, it is worth noting that for a GMRF the precision matrix $\boldsymbol{Q}$ is full-rank due to it being positive definite. As will become clear, this is not the case for every model in this project.

3 Intrinsic Gaussian Markov Random Fields

An IGMRF is closely related to a GMRF, but it has some unique properties that can be desirable. This section starts by deriving the density of a GMRF under linear constraints. Then the definition of an IGMRF is provided. Afterward, relevant properties of an IGMRF are presented and used to motivate why an IGMRF may be desirable over a GMRF.

3.1 GMRF under linear constraints

Let $\boldsymbol{x}^\top = (x_1, ..., x_n)$ be a GMRF with precision matrix $\boldsymbol{Q}$ and zero mean for simplicity. Furthermore, let $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$ be a diagonal matrix where $\lambda_1, ..., \lambda_n$ is the eigenvalues of $\boldsymbol{Q}$, and let $\boldsymbol{V} = (\boldsymbol{e}_1, ..., \boldsymbol{e}_n)$ be a matrix where $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$ are the eigenvectors of $\boldsymbol{Q}$. The density of $\boldsymbol{x}$ subject to linear constraints, $\boldsymbol{Ax} = \boldsymbol{a}$, is written as

$$\pi(\boldsymbol{x}|\boldsymbol{Ax} = \boldsymbol{a})$$

where $\boldsymbol{A} = (\boldsymbol{e}_1, ...\boldsymbol{e}_k)^\top$, $k \in \{1, ..., n\}$, and $\boldsymbol{a}^\top = (a_1, ..., a_k) \in \mathbb{R}^n$. Hence $\boldsymbol{x}$ is constrained as $\pi(\boldsymbol{x}|\boldsymbol{Ax} = \boldsymbol{a}) = 0$ if $\boldsymbol{Ax} \neq \boldsymbol{a}$, and therefore $\boldsymbol{x}$ can only take values satisfying $\boldsymbol{Ax} = \boldsymbol{a}$. Expanding on this it is possible to find an expression for the log-density of $\boldsymbol{x}$ subject to these constraints. From Rue and Held [2005, Section 3], the expected value of $\boldsymbol{x}$ subject to linear constraints is

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{Ax} = \boldsymbol{a}] = \boldsymbol{V}\begin{bmatrix}\boldsymbol{a} \\ \boldsymbol{0}\end{bmatrix}$$

and the precision is

$$\text{Prec}(\boldsymbol{x}|\boldsymbol{Ax} = \boldsymbol{a}) = \tilde{\boldsymbol{Q}} = \boldsymbol{V}\tilde{\Lambda}\boldsymbol{V}^T$$

where $\tilde{\Lambda} = \text{diag}(0, ..., 0, \lambda_{k+1}, ..., \lambda_n)$. Clearly $\tilde{\boldsymbol{Q}}$ is not full-rank as the first k diagonal elements of $\tilde{\Lambda}$ are zero. The constrained log-density can then be derived as

$$\log(\pi(\boldsymbol{x}|\boldsymbol{Ax} = \boldsymbol{a})) = -\frac{n-k}{2}\log(2\pi) + \frac{1}{2}\sum_{i=k+1}^{n}\log(\lambda_i) - \frac{1}{2}\boldsymbol{x}^T\tilde{\boldsymbol{Q}}\boldsymbol{x}$$

This means that $\boldsymbol{x}|\boldsymbol{Ax} = \boldsymbol{a}$ has a density on a lower dimensional space.

3.2 Definition of an IGMRF

Firstly, let $\boldsymbol{Q}$ be an $n \times n$ symmetric positive semi-definite matrix, with $\text{rank}(\boldsymbol{Q}) = n-k$, $\boldsymbol{\mu}$ be an arbitrary vector with n elements, and $\boldsymbol{S}_{k-1}$ be the polynomial design matrix of order $k-1$. In accordance with Rue and Held [2005, Section 3]: $\boldsymbol{x}^\top = (x_1, ..., x_n)$ is an IGMRF of order k with respect to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ having parameters $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ if its density is of the form

$$\pi(\boldsymbol{x}) = (2\pi)^{-\frac{n-k}{2}} (|\boldsymbol{Q}|^*)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{6}$$

and $\boldsymbol{Q}$ is defined by

$$Q_{ij} \neq 0 \iff \{i,j\} \in \mathcal{E}, \text{ for } i \neq j \tag{7}$$

and

$$\boldsymbol{Q}\boldsymbol{S_{k-1}} = \boldsymbol{0} \tag{8}$$

Remarks:
The density of an IGMRF $\boldsymbol{x}$ as seen in Equation (6) is similar to the density of a GMRF subject to linear constraints. The difference is that an IGMRF has an improper density, whereas a GMRF subject to linear constraints is a proper density on a lower dimensional space. It is worth noting that in reality $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ do not represent mean and precision, as these do not exist under an improper distribution. However, it is still useful to think of them as mean and precision, and for the remainder of this project, they will be referred to as mean and precision respectively. The operator $|\cdot|^*$ is the generalized determinant, and it is defined as the product of the non-zero eigenvalues of the matrix.
Equation (7) implies that an entry, Q_{ij} where $i \neq j$, in $\boldsymbol{Q}$ is non-zero if and only if the undirected graph has an edge between the corresponding nodes i and j. This is the same as for a GMRF.
Lastly, Equation (8) states that the product of the precision matrix, $\boldsymbol{Q}$, and the polynomial design matrix of order $k-1$, $\boldsymbol{S}_{k-1}$, is zero for an IGMRF. Hence, the columns in $\boldsymbol{S}_{k-1}$ make up the basis of the null space of $\boldsymbol{Q}$, something which will be discussed further when detailing the properties of an IGMRF.

3.3 Properties of an IGMRF

For the remainder of this section let $\boldsymbol{x}^\top = (x_1, ..., x_n)$ be an IGMRF of order k, with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q}$, with respect to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

An IGMRF has the same conditional independence properties as a GMRF, see Equation (3). It follows that if every x_i is only conditionally dependent on a few x_j then $\boldsymbol{Q}$ is sparse. Hence, using an IGMRF also allows for sparse matrix calculations. Additionally, an IGMRF can also be defined implicitly by its full conditionals in a similar way as a GMRF.

A reason why IGMRFs may be desirable over GMRFs is Equation (8). This condition

states that the columns of the polynomial design matrix of order $k-1$, $\boldsymbol{S}_{k-1}$, span the null space of $\boldsymbol{Q}$. Since $\boldsymbol{Q}$ has rank $n-k$ it means that the null space is k-dimensional. Furthermore, $\boldsymbol{S}_{k-1}$ is a k-dimensional matrix with orthogonal columns, hence the columns of $\boldsymbol{S}_{k-1}$ must span the null-space of $\boldsymbol{Q}$ for Equation (8) to hold. Therefore the columns of $\boldsymbol{S}_{k-1}$ form a basis for the null space of $\boldsymbol{Q}$. This is important as $\boldsymbol{x}$ is invariant to the addition of any element lying in the null space of $\boldsymbol{Q}$. To see this more clearly, $\boldsymbol{x}$ can be decomposed into two parts:

- $\boldsymbol{x}^{||}$ which lies within the null space of $\boldsymbol{Q}$
- $\boldsymbol{x}^{\perp}$ orthogonal to null space of $\boldsymbol{Q}$

The density of $\boldsymbol{x}$ only depends upon the value of $\boldsymbol{x}^{\perp}$, as

$$\begin{aligned}\pi(\boldsymbol{x}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{x}^{||}+\boldsymbol{x}^{\perp}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}^{||}+\boldsymbol{x}^{\perp}-\boldsymbol{\mu})\right)\\ &= \exp\left(-\frac{1}{2}(\boldsymbol{x}^{\perp}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}^{\perp}-\boldsymbol{\mu})\right)\\ \pi(\boldsymbol{x}) &= \pi(\boldsymbol{x}^{\perp})\end{aligned}$$

Therefore, an IGMRF is invariant to trends in the null space of the precision matrix. Since the polynomial design matrix $\boldsymbol{S}_{k-1}$ spans the null space of $\boldsymbol{Q}$ it means that $\boldsymbol{x}$ is invariant to the addition of any polynomial trend of degree $k-1$. This implies that an IGMRF can fit these trends *intrinsically*. What an IGMRF then models is the deviation from this unspecified trend.

This invariance comes at a cost. An IGMRF leads to an improper distribution as $\boldsymbol{Q}$ is rank-deficient. Improper distributions can lead to issues if a problem is not well-behaved. Additionally, it will be necessary to impose constraints on the IGMRFs used later in this project for identifiability purposes. The number of constraints needed may increase as the rank-deficiency of the precision matrix increases. The more rank-deficient $\boldsymbol{Q}$ is, the more constraints may be necessary. This can become a computational issue when the precision matrix has a large rank-deficiency. Looking at the improper density of an IGMRF in Equation (6), it is evident that it is similar to the density of a GMRF subject to linear constraints. According to Rue and Held [2005, Section 3], the density of $\boldsymbol{x}^{\perp}$ is equal to that of a GMRF subject to linear constraints. This implies that imposing sufficient constraints on $\boldsymbol{x}$, that constrains the values of $\boldsymbol{x}^{||}$, yields a proper distribution. In a model with other components than $\boldsymbol{x}$, imposing constraints on $\boldsymbol{x}$ then enables the identifiability of the components confounded with $\boldsymbol{x}^{||}$.

4 Common temporal models

This section presents common temporal models which are used to model temporal random effects. Firstly a proper temporal model, the autoregressive model, is presented along with some motivation behind it. Then an improper model, the random walk, is presented along with a motivation of why the random walk model may be preferred over the autoregressive.

4.1 The autoregressive model of order 1

As stated earlier, a GMRF can be represented by a graph. For an autoregressive model, this graph is a straight line and is therefore often used to model time. Let $\boldsymbol{x}^\top = (x_1, ..., x_T)$ follow an autoregressive model of order n (ARn), then x_t is conditionally dependent on the n closest points on the graph to either side. In Figure 3 there is an illustration of how a graph for an AR2 model looks for $\boldsymbol{x}^\top = (x_1, x_2, x_3, x_4, x_5)$. From Figure 3 it is easily seen that any x_i is only conditionally dependent on its second-order neighbours, as the graph only has edges between nodes either one or two instances apart.

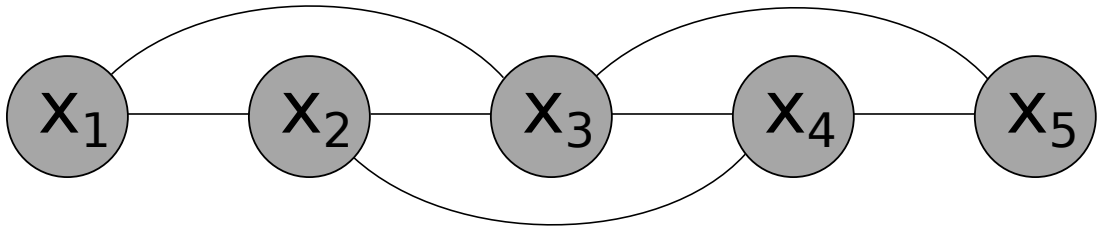


Figure 3: The graph of an AR2 model with five nodes.

Consider $\boldsymbol{x}^\top = (x_1, ..., x_T)$, and assume it has zero mean. For the ARn model, x_t for $t \in [1 + n, ..., T]$ is defined by

$$x_t = \rho_1 x_{t-1} + ... + \rho_n x_{t-n} + \varepsilon_t, \ \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau^{-1}\right) \tag{9}$$

where $\rho_1, ..., \rho_n \in \mathbb{R}$ are correlation parameters such that the process is second-order stationary, $x_1, ..., x_n$ is assumed to have marginal distribution equal to the stationary distribution, and $\tau > 0$ is the precision (inverse variance) of the increments [Rue and Held, 2005]. The process being second-order stationary means that the mean is constant and the covariance between elements of the process only depends on the time difference between them. For x_t having a non-zero mean μ for all $t \in [1, ..., T]$, we can work with the centered version $\tilde{x}_t = x_t - \mu$. Then Equation (9) would be $x_t = \rho_1(x_{t-1} - \mu) + ... + \rho_n(x_{t-n} - \mu) + \mu + \varepsilon_t$. Thus this section will only consider $\boldsymbol{x}$ having zero mean, as the extension to a non-zero mean is easy.

Restricting our attention to the AR1 model, let $\rho \in [0, 1)$, $\tau > 0$, and $t \in [2, ..., T]$, then

$$x_t = \rho \cdot x_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau^{-1}\right) \tag{10}$$

In Equation (10) ρ is specified to be within $[0, 1)$ to ensure both that the process is second-order stationary and that realizations are smooth over time if $\rho > 0$ [Rue and Held, 2005, Section 4]. It is possible to let $\rho \in (-1, 1)$ and this would still be an AR1 model. However, for this project, it is not interesting with $\rho < 0$ as negative correlation would produce dissimilar responses in temporal neighbours.
Since $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau^{-1}\right)$ one gets that the increments are iid Gaussian with mean zero and precision τ

$$x_t - \rho x_{t-1} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau^{-1}\right)$$

As a result, x_t is conditionally independent all $x_1, ..., x_{t-2}$ given x_{t-1}. From this, and assuming $x_1 \sim \mathcal{N}\left(0, (\tau(1-\rho^2))^{-1}\right)$, the joint density of $\boldsymbol{x}$ is

$$\begin{aligned} \pi(\boldsymbol{x}) &= \pi(x_1)\pi(x_2|x_1)...\pi(x_T|x_{T-1}) \\ &\propto \tau^{\frac{T}{2}} \exp\left(-\frac{\tau}{2}\left((1-\rho^2)x_1^2 + \sum_{t=2}^{T}(x_t - \rho x_{t-1})^2\right)\right) \\ &= \tau^{\frac{T}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}\right) \end{aligned} \tag{11}$$

where $\boldsymbol{Q}$ is a $T \times T$ dimensional matrix equal to

$$\boldsymbol{Q} = \tau \begin{bmatrix} 1 & -\rho & & & \\ -\rho & 1+\rho^2 & -\rho & & \\ & \ddots & \ddots & \ddots & \\ & & -\rho & 1+\rho^2 & -\rho \\ & & & -\rho & 1 \end{bmatrix} \tag{12}$$

Since $\rho \in [0, 1)$ it is clear that $\boldsymbol{Q}$ is full rank. Equation (11) is the kernel of a Gaussian distribution with zero mean and precision matrix $\boldsymbol{Q}$. Furthermore, Equation (12) shows that for $i, j \in [1, ..., T]$ and $i \neq j$, $Q_{ij} \neq 0 \iff i = \pm j$. Therefore the AR1 model is a GMRF with respect to an undirected graph that is a straight line with edges only between subsequent nodes.
From (11) it is easy to derive the full conditionals

$$x_t|\boldsymbol{x_{-t}}, \tau \sim \begin{cases} \mathcal{N}\left(\rho x_2,\ \tau^{-1}\right),\ t = 1 \\ \mathcal{N}\left(\frac{\rho}{1+\rho^2}(x_{t+1} + x_{t-1}),\ (\tau(1+\rho^2))^{-1}\right),\ 1 < t < T \\ \mathcal{N}\left(\rho x_{T-1}, \tau^{-1}\right),\ t = T \end{cases} \tag{13}$$

Equation (13) implies that for any t, x_t has a full conditional mean proportional to the sum of its temporal neighbours' values. If $\rho > 0$, this dependency will produce smoothed results over time, as x_t cannot deviate too much from its temporal neighbours. This can be desirable in a modeling context, as many temporal processes are believed to have smooth behaviour.
Since $\rho \in [0, 1)$, the AR1 model is stationary. Stationarity implies that the process will

tend to an overall mean. Hence for $\boldsymbol{x}$ an AR1, all x_i are dependent upon its temporal neighbours and the overall mean of $\boldsymbol{x}$. This dependence upon the overall mean can be undesirable. This motivates the use of IGMRFs in temporal modeling, as an IGMRF of order k is invariant to polynomial trends of order $k-1$.

4.2 Random walk models

If $\boldsymbol{x}^\top = (x_1, ..., x_T)$ has a completely flat prior for x_1, $f(x_1) \propto 1$, and x_t for $t \in [2, ..., T]$ is defined as

$$x_t = x_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau^{-1}\right), \quad \tau > 0$$

then $\boldsymbol{x}$ is an IGMRF of order 1 called a random walk of order 1 (RW1) [Rue and Held, 2005, Section 3]. This is the same as setting $\rho = 1$ in Equation (10). The $T \times T$ precision matrix $\boldsymbol{Q}$ of a RW1 is

$$\boldsymbol{Q} = \tau \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix} \tag{14}$$

Clearly, $\boldsymbol{Q}$ has a rank-deficiency of 1. Additionally, each row sums to zero in Equation (14). Therefore $\boldsymbol{Q}\mathbf{1} = 0$ where $\mathbf{1}$ is a vector of 1's, and $\boldsymbol{x}$ is an IGMRF of order 1. The full conditional of x_t is

$$x_t | \boldsymbol{x_{-t}}, \tau \sim \begin{cases} \mathcal{N}\left(x_2, \tau^{-1}\right), & t = 1 \\ \mathcal{N}\left(\frac{1}{2}(x_{t-1} + x_{t+1}), \ (2\tau)^{-1}\right), & 1 < t < T \\ \mathcal{N}\left(x_{T-1}, \tau^{-1}\right), & t = T \end{cases} \tag{15}$$

The full conditional in Equation (15) shows that a RW1 model exhibits a purely local behaviour. Similarly to an AR1 model, the RW1 model leads to smoothed realizations in time, as x_t is centered around the average of its temporal neighbours. What separates a RW1 from an AR1, is that an AR1 is stationary and tends to an overall mean. A RW1 model is invariant to constant trends and x_t only depends upon its temporal neighbours for all $t \in [1, ..., T]$. Thus, a RW1 models local deviations from an unspecified level. The local behaviour is shown in Figure 4a where the expected value of $x_t | x_1, ..., x_{t-1}$ is shown, and it is centered at x_{t-1} without regard to an overall mean. This completely local behaviour allows the RW1 to capture a wide range of temporal processes, including processes that have non-constant trends. An AR1 process cannot capture any non-constant temporal trends on its own, as this would make it non-stationary.

Going back to Equation (9), let $n = 2$, $\rho_1 = 2$, $\rho_2 = -1$, and $\tau > 0$. Then for $\boldsymbol{x}^\top = (x_1, ..., x_T)$ with x_1 and x_2 having flat priors, for $t \in [3, ..., T]$ x_t defined as

$$x_t = 2x_{t-1} - x_{t-2} + \varepsilon_t,\ \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^{-1})$$

leads to an IGMRF of order 2 called the random walk of order 2 (RW2) [Rue and Held, 2005, Section 3]. The $T \times T$ dimensional precision matrix $\boldsymbol{Q}$ of $\boldsymbol{x}$ is equal to

$$\boldsymbol{Q} = \tau \begin{bmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & 2 \\ & & & & 1 & -2 & 1 \end{bmatrix} \tag{16}$$

For $\boldsymbol{x}^\top = (x_1, ..., x_T)$ following a RW2, x_t depends upon its first and second order neighbours for all $t \in [1, ..., T]$. Figure 3 shows the graph structure of a RW2 with five nodes. This second-order dependency leads to a RW2 having even smoother realizations than a RW1, as it borrows information from its second-order neighbours as well. Following from the fact that a RW2 is an IGMRF of order 2, it means that an RW2 is invariant to the addition of linear trends. Same as for a RW1, a RW2 exhibits purely local behaviour as x_t is only conditionally dependent upon its second-order neighbours. A RW2 then models local deviations from an unspecified line. Figure 4b illustrates this local behaviour, as it shows the density of x_t conditional on $x_1, ..., x_{t-1}$.

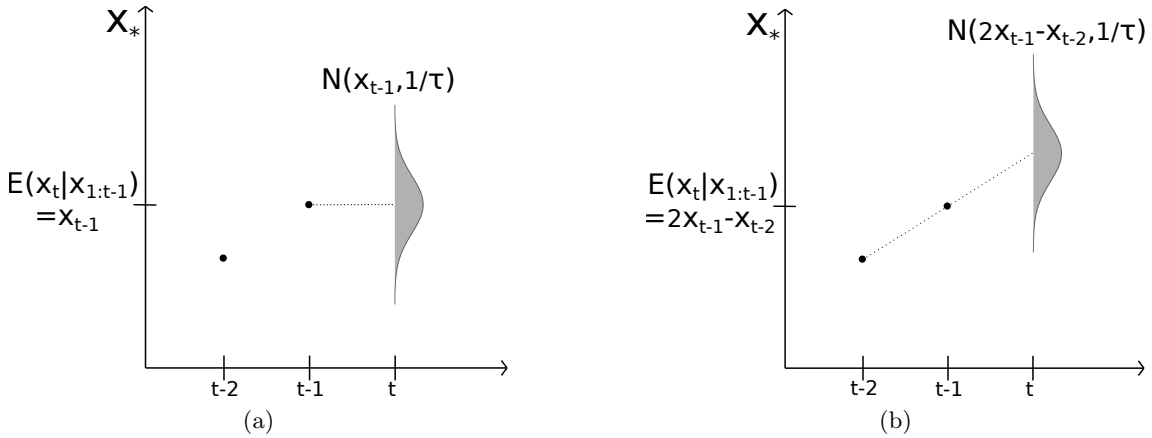


Figure 4: The density of x_t conditional on $\boldsymbol{x_{1:t-1}} = (x_1, ..., x_{t-1})$ for $\boldsymbol{x}$ following a **(a)** RW1 and **(b)** RW2.

5 Common spatial models

In this section, we extend our view from GMRFs and IGMRFs on straight graphs with natural orderings of the nodes, to more complicated graph-structures requiring a different approach. Firstly, this section introduces how to define dependency structures for areal data using the concept of adjacency. Then a proper spatial model following a conditional autoregressive model (CAR) called the Leroux model is introduced [Leroux *and others*, 2000]. Lastly, the Besag model is introduced as an improper spatial model, which is an intrinsic conditional autoregressive model (ICAR) [Besag *and others*, 1991].

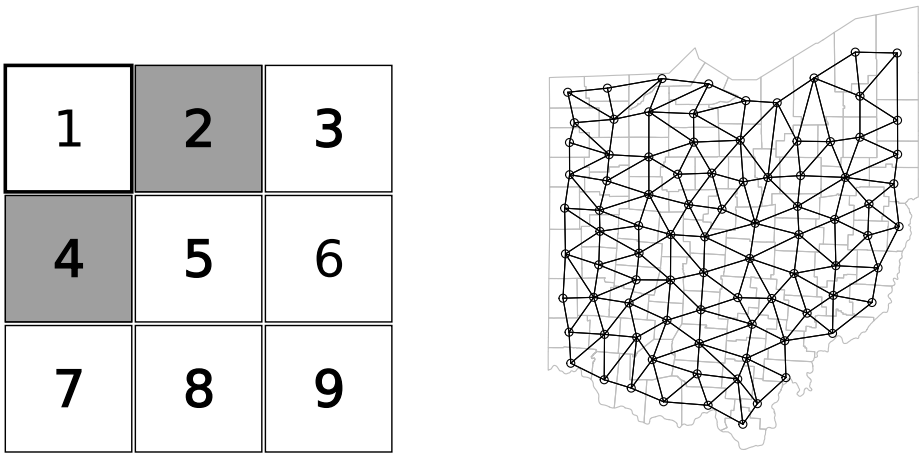
5.1 Defining adjacency

When dealing with areal data, data labeled with an associated area, it is a necessary preliminary to define how to construct the dependency structure. In this project thesis adjacency will be used to construct the spatial dependency structure.
One relevant definition of adjacency is that two non-overlapping areas i and j are adjacent, written $i \sim j$, if they share a common border. In this project, this is the definition of adjacency that is used. Adjacency is illustrated in Figure 5a, where area 1 is adjacent to areas 2 and 4 as these share borders. A more complex illustration of adjacency is given in Figure 5b, showing the counties in Ohio. Adjacency is illustrated by a black line drawn between the adjacent areas. The notion of adjacency ties into graphs. Each area is a node in a graph, and two adjacent areas have an edge between their corresponding nodes in that graph. Using this relation between adjacency and graphs, it is possible to construct the dependency structure in space for GMRFs and IGMRFs using adjacency. Later in this project there will be spatial effects over the counties of Ohio, and therefore the spatial effects are defined as GMRFs and IGMRFs with respect to the graph in Figure 5b.
One important thing to note is that this project works with connected graphs. A graph is connected if there exist edges between nodes such that one can travel from any one node to any other node by traversing edges, i.e. no islands.

5.2 The Leroux model

A CAR model is analogous to an AR model, but it is not restricted to a graph that is a line [Rue and Held, 2005, Section 1]. Hence, there is no natural ordering of the nodes. Assume we have areas $\{1, ..., n\}$ represented by a graph, such as in Figure 5b. Consider $\boldsymbol{x}^\top = (x_1, ..., x_n)$ where each x_i for $i \in [1, ..., n]$ is associated with an observation or a property of area i. It is useful to implicitly construct a model for $\boldsymbol{x}$ by using the full conditionals in Equations (4) and (5). It is because the model is defined by its full conditionals that it has gotten its name conditional autoregressive. Let $\boldsymbol{x}$ have zero mean. We say that $\boldsymbol{x}$ is modelled as a CAR if the full conditional of x_i is defined as

$$x_i|\boldsymbol{x}_{-i} \sim \mathcal{N}\left(\sum_{j:j\neq i} \beta_{ij} x_j,\ \kappa_i^{-1}\right) \tag{17}$$



(a) (b)

Figure 5: (a) shows nine areas $\{1, ..., 9\}$. The adjacent areas $\{2, 4\}$ of area 1 are shaded grey. (b) counties in Ohio. Each county is represented by a node, and edges between nodes represent adjacency between the corresponding counties.

where β_{ij} are defined through the notion of adjacency, $\beta_{ij} \neq 0 \iff i \sim j$, and $\kappa_i > 0 \; \forall i \in \{1, ..., n\}$. Therefore the mean of the full conditional of x_i is a weighted sum of the values in the regions adjacent to region i. This is an extension of the AR1 model, but on an irregular lattice. A region i is only conditionally dependent on its first-order spatial neighbours. If the lattice had been a line, the CAR and AR1 models would have the same conditional dependency structure. Equation (17) makes it clear that $\boldsymbol{x}$ will be smooth in space if $\forall i, j \in [1, ..., n], \;\; \beta_{ij} \geq 0$.
From Rue and Held [2005, Section 1], assuming that we have a connected graph, the joint density becomes

$$\pi(\boldsymbol{x}) \propto |\boldsymbol{Q}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}\right)$$

where $\boldsymbol{Q}$ is the precision matrix with entries

$$Q_{ij} = \begin{cases} \kappa_i, & i = j \\ -\kappa_i\beta_{ij}, & i \neq j \end{cases}$$

There is no clear answer for what exact values β_{ij} and κ_i should take to ensure that $\boldsymbol{x}$ follows a proper distribution. One common approach is to force $\boldsymbol{Q}$ to be diagonal dominant. How we construct $\boldsymbol{Q}$ for the CAR model is by

$$\boldsymbol{Q} = \tau((1 - \lambda)I + \lambda\boldsymbol{K})$$

where $\boldsymbol{K}$ is defined the same way as for the ICAR which is discussed next, $\tau > 0$ is a precision parameter, and $\lambda \in [0, 1)$ is a spatial correlation parameter that ensures that $\boldsymbol{Q}$ is diagonal dominant. This model is known as a Leroux model [Leroux *and others*, 2000].

The correlation parameter λ decides to what degree realizations are spatially dependent, and therefore how smooth $\boldsymbol{x}$ is in space. If $\lambda = 0$, then $\boldsymbol{x}$ has iid realizations for each area. If λ close to 1 then each x_i is correlated to x_j for all $j \sim i$, making $\boldsymbol{x}$ smooth in space.

5.3 The Besag model

Following Besag *and others* [1991], $\boldsymbol{x} = (x_1, ..., x_n)$ is an ICAR if the full conditionals are defined as

$$x_i|\boldsymbol{x}_{-i} \sim \mathcal{N}\left(\frac{1}{n_i}\sum_{i\sim j} x_j,\ \frac{1}{n_i\tau}\right) \tag{18}$$

where n_i is the number of regions adjacent to region i. If $\boldsymbol{x}$ is defined by Equation (18), then $\boldsymbol{x}$ is said to follow a Besag model [Besag *and others*, 1991]. Equation (18) implies that x_i is centered around the mean of the adjacent regions, with precision proportional to the number of adjacent regions. In this way x_i cannot deviate too much from its neighbours. Any process following a Besag model will therefore be smooth in space. This is often desirable as many processes in space are believed to have similar characteristics for nearby areas, and as such the realizations should be smooth in space. The joint density is given by

$$\pi(\boldsymbol{x}) \propto \tau^{-\frac{n-1}{2}} \exp\left(-\frac{\tau}{2}\sum_{i\sim j}(x_i - x_j)^2\right) = \tau^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}\right)$$

where $\boldsymbol{Q} = \tau\boldsymbol{K}$ and $\boldsymbol{K}$ is the structure matrix. The structure matrix is defined by

$$K_{ij} = \begin{cases} n_i,\ i = j \\ -1,\ i \sim j \\ 0,\ \text{otherwise} \end{cases} \tag{19}$$

From Figure 5a the structure matrix of the Besag model for the nine areas, $\boldsymbol{K}$, would be

$$\boldsymbol{K} = \begin{bmatrix} 2 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 3 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 3 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2 \end{bmatrix}$$

From $\boldsymbol{K}$ one can see that the rows and columns all sum to zero $\boldsymbol{K}I_n = \boldsymbol{0}$. This clearly shows that it is an IGMRF of order 1. This means that $\boldsymbol{K}$ is not full rank. It has a rank

$n - 1 = 8$. Since the Besag model is an IGMRF of order 1, it is invariant to the addition of any constant trend. This invariance leads to the Besag model, similar to the RW1, having a purely local behaviour.

6 Scaling IGMRFs and penalized complexity priors

The previous two Sections 4 and 5 outlined common models for temporal and spatial random effects. Some of these models are IGMRFs. Like other GMRFs, they have a precision parameter τ to which in a Bayesian framework a prior distribution is assigned. This section motivates the need for scaling IGMRFs to ease prior elicitation for τ. Afterwards, the penalized complexity prior (PC-prior) for the precision parameter proposed by Simpson *and others* [2017] is presented.

Scaling IGMRFs

The precision matrix of a GMRF or an IGMRF can be expressed as

$$\boldsymbol{Q} = \tau \boldsymbol{K}$$

where $\tau > 0$ is the unknown precision parameter and $\boldsymbol{K}$ is the corresponding structure matrix for the respective graph of the model. For $\boldsymbol{x}$ an IGMRF of order k, $\boldsymbol{S}_{k-1}$ is the null space of $\boldsymbol{K}$, and $\boldsymbol{K}$ defines the penalty of deviating from $\boldsymbol{S}_{k-1}$, see Rue and Held [2005, Section 3]. For a first-order IGMRF, the null space is spanned by a constant polynomial, while for a second-order IGMRF, the null space is spanned by a first-order polynomial. This implies that assigning the same prior to the precision parameter of different IGMRFs may lead to different interpretations, see Sørbye and Rue [2014] for details. In this project, it is desired that equal priors on the precision of two different IGMRFs have the same interpretation.

In Sørbye and Rue [2014] the marginal standard deviation of all the components of an IGMRF $\boldsymbol{x}^\top = (x_1, ..., x_n)$ are approximated as

$$\sigma(x_i) \approx \frac{\sigma_{\text{ref}}(\boldsymbol{x})}{\sqrt{\tau}}, \quad \text{for } i = 1, ..., n$$

where $\sigma_{\text{ref}}(\boldsymbol{x}) = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log(\sigma_{\tau=1}(x_i))\right)$ is a reference standard deviation for fixed precision $\tau = 1$. The value of $\sigma_{\text{ref}}(\boldsymbol{x})$ is different for different IGMRFs. As an example, according to Sørbye and Rue [2014] for $\boldsymbol{x}_1$ a RW1 and $\boldsymbol{x}_2$ a RW2 on corresponding graphs with 100 nodes, the reference standard deviations are $\sigma_{\text{ref}}(\boldsymbol{x}_1) = 3.89$ and $\sigma_{\text{ref}}(\boldsymbol{x}_2) = 41.39$, respectively. Since $\sigma_{\text{ref}}(\boldsymbol{x}_1)$ and $\sigma_{\text{ref}}(\boldsymbol{x}_2)$ are different, assigning the same prior to τ for both $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ allows for greater local deviations of $\boldsymbol{x}_2$ compared to $\boldsymbol{x}_1$. Therefore Sørbye and Rue [2014] proposes that we assign a hyperprior for $\frac{\tau}{\sigma_{\text{ref}}(\boldsymbol{x})^2}$. Then the marginal standard deviation is approximately

$$\sigma(x_i) \approx \frac{1}{\sqrt{\tau}}, \quad \text{for } i = 1, ..., n$$

and hence the choice of hyperprior for $\frac{\tau}{\sigma_{\text{ref}}(\boldsymbol{x})^2}$ has the same marginal interpretation. This is equivalent to choosing a prior for τ for a scaled version of $\boldsymbol{x}$, $\boldsymbol{x}^* = \frac{\boldsymbol{x}}{\sigma_{\text{ref}}(\boldsymbol{x})}$. Later in this project, we scale our improper models in this fashion so that the choice of hyperpriors on τ has the same interpretation with respect to the marginal standard deviation for different IGMRFs.

Penalized complexity priors

This project uses PC-priors that favour simplicity over complexity to avoid overfitting [Simpson *and others*, 2017]. Unless there is sufficient support in the data for increased complexity, the PC-prior should penalize complexity to the degree that the resulting posterior is as simple as warranted. PC-priors have decreasing density for parameter values further and further from a base model. In this way, the PC-prior favours a simple model over a more complex model. For a model with one random effect $\boldsymbol{x}$ with precision matrix $\tau\boldsymbol{K}$, the base model is the absence of the random effect. This implies that $\tau = \infty$. Any prior on τ such that $\mathbb{E}[\tau] < \infty$ leads to overfitting according to Simpson *and others* [2017]. This means that a Gamma prior is a poor choice as prior for the precision τ, as it will lead to overfitting. For $\boldsymbol{x}$ having a precision hyperparameter τ, the PC-prior on τ is defined as

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-\frac{3}{2}} \exp\left(-\lambda\tau^{-\frac{1}{2}}\right), \;\; \tau > 0, \; \lambda > 0$$

The PC-prior has $\mathbb{E}[\tau] = \infty$, so unlike the Gamma prior, it is not guaranteed to overfit. The PC-prior provides us with the possibility of imposing a weak upper bound, U, on the marginal standard deviation a priori. By setting $\lambda = -\frac{\ln(\alpha)}{U}$, we get that

$$P\left(\frac{1}{\sqrt{\tau}} > U\right) = \alpha$$

where α is a small probability. By using scaling and a PC-prior on the precision parameter, the interpretation of the prior on the precision is that there is a small probability α that the marginal standard deviation of x_i is greater than U.

7 Bayesian hierarchical spatiotemporal modelling

So far temporal and spatial models that can be used for modelling temporal and spatial random effects, respectively, have been introduced. When the process of interest is developing in both space and time the modelling framework must be extended to deal with both temporal and spatial effects, and possibly also spatiotemporal interactions.

This section starts by presenting one of the current state-of-the-art Bayesian hierarchical spatiotemporal modelling approaches, as introduced in Knorr-Held [2000]. Firstly, the general linear predictor for the spatiotemporal model containing both spatial and temporal random effects and a spatiotemporal interaction term is introduced. Then a brief overview of integrated nested Laplace approximations (INLA) [Rue *and others*, 2009] is

given as the INLA methodology will be used for inference in this project. Afterward, a motivation for the spatial and temporal random effects and how we model them is provided. Then each type of interaction as detailed by Knorr-Held [2000] is specified, along with the necessary constraints to ensure identifiability of all model components. Lastly, some alternative model specifications relying on proper models are introduced along with a motivation of these models. The implementation of all the different models and their components in the INLA framework is shown along the way.

7.1 A typical linear predictor for Bayesian spatiotemporal models

Let $\boldsymbol{y} = (y_{11}, ..., y_{1n}, ..., y_{Tn})$ be the response measured in each area $i \in [1, ..., n]$ at each time point $t \in [1, ..., T]$. Assume that y_{ti} follows a Poisson distribution as

$$y_{ti}|n_{ti}, \lambda_{ti} \sim \text{Poisson}(n_{ti}\lambda_{ti})$$

where n_{ti} is a area and time specific offset, and λ_{ti} the rate in area i at time t. We want to model λ_{ti} using a linear predictor $\eta_{ti} = \log(\lambda_{ti})$ for all i and t. A common linear predictor η_{ti}, as introduced by Knorr-Held [2000], is of the form

$$\eta_{ti} = \mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{ti} \tag{20}$$

In Equation (20) μ is the overall risk and is the same for all areas i and all times t, α_t and γ_t are temporal random effects, while θ_i and ϕ_i are both spatial random effects, lastly δ_{ti} is the spatiotemporal interaction term. All of the effects and interactions are assumed multivariate Gaussian with zero mean and precision matrix $\tau_* \boldsymbol{K}_*$ a priori.
Firstly, the overall risk μ is assumed Gaussian with zero mean and a fixed low precision (inverse variance) usually around 0.001. The overall risk is the same for all areas i at all times t, and thus it can be thought of as a 'best guess' in the absence of any information about time or region. It typically has a fixed low precision to encapsulate a lack of knowledge about the overall risk. We could exclude μ, in which case it would be a part of either the spatial random effects or the temporal random effects. However, it is included here for clarity and the constraints imposed on the random effects are used to identify μ.

Furthermore, in Equation (20) there are two temporal random effects, α_t and γ_t, and two spatial random effects, θ_i and ϕ_i. There are two spatial and two temporal random effects to encapsulate two important and opposing properties of the process being modelled. Namely, that of structured and unstructured effects. Now a brief overview of INLA is given before we motivate why there is a need for both a structured and unstructured effect and how we model them together through the BYM2 model [Riebler *and others*, 2016].

7.2 Bayesian inference using INLA

To do Bayesian inference it is necessary to compute the posterior distribution. In this project the INLA framework is used [Rue *and others*, 2009]. For theoretical details on

how INLA works, see Appendix B.
INLA is a tool for full Bayesian inference that directly approximates the posterior marginals using nested Laplace approximations. It works on a class of models called latent Gaussian models. In this project, the latent fields consist of GMRFs or IGMRFs, making them latent Gaussian models. Additionally, the different latent fields in this project are sparse. INLA exploits the sparse structure of the latent field to do fast computations. As a result, INLA is fast and deterministic. If we were to have used Markov Chain Monte Carlo (MCMC) techniques, we would rely on sampling. Spatiotemporal modeling has high dimensional latent fields making MCMC slow as the chain may take a long time to converge and samples may be very correlated with each other. This makes INLA desirable over MCMC in this context. For the actual implementation, the R-INLA package is used, which can be found at `https://www.r-inla.org/`. R-INLA is a free-to-use package for the programming language R, making it fairly easy to use the INLA framework in practice.

7.3 The BYM2 model for spatial or temporal effects

Consider a process developing in time. The characteristics of the process will likely not change completely from time $t-1$ to t. The characteristics of a process can be anything affecting realizations, such as demographics, climate, public health measures, etc. As a result, it is desirable with an effect that is smoothed in time so that the value at t is not too dissimilar to the value at $t-1$ and $t+1$. This motivates the structured temporal random effect $\boldsymbol{\alpha}^\top = (\alpha_1, ..., \alpha_T)$. From Section 4.2 it is known that the random walk models lead to realizations that are smoothed in time. It is common to let $\boldsymbol{\alpha}$ follow either a RW1 or RW2 [Richardson *and others*, 2006; Wakefield *and others*, 2019]. Then $\boldsymbol{\alpha}$ is invariant to any overall mean, and will exhibit purely local behaviour. As stated earlier, $\boldsymbol{\alpha}$ is assumed multivariate Gaussian with mean zero and precision matrix $\boldsymbol{Q}_\alpha = \tau_\alpha \boldsymbol{K}_\alpha,\ \tau_\alpha > 0$ a priori. The precision matrix of $\boldsymbol{\alpha}$ following a RW1 is defined in Equation (14) and for a RW2 in Equation (16).

Most temporal processes cannot entirely be modelled as only having structured effects in time. A process may have characteristics that deviate from structure, i.e. unstructured effects. There may be time points that have an unlikely response if only looking at structured effects. Take the spread of a contagious disease as an example. There could be one-off events such as demonstrations and concerts, or sporadic events such as air pollution episodes, that affect the response in a singular time point or several that cannot be explained by a smooth effect. Therefore, it is necessary with an unstructured effect. This motivates our second temporal random effect, $\boldsymbol{\gamma}^\top = (\gamma_1, ..., \gamma_T)$, which is assumed multivariate Gaussian with mean zero and precision matrix $\tau_\gamma \boldsymbol{K}_\gamma$ a priori. Since $\boldsymbol{\gamma}$ is supposed to be unstructured, it should be that every γ_t is independent of all other $\gamma_{t'}$ where $t \neq t'$ and thus $\boldsymbol{K}_\gamma = I_T$ where I_T is the identity matrix in $\mathbb{R}^{T \times T}$. This makes γ_t conditionally independent of all other time points. As stated in Section 2.3 where the properties of a GMRF are shown, that γ_t is conditionally dependent on $\gamma_{t'} \iff \boldsymbol{Q}_{\gamma:(t,t')} \neq 0,\ t \neq t'$, and since $\boldsymbol{Q}_\gamma = \tau_\gamma I_T$ this is not the case for any pair of time points.

For processes in space, we have a similar line of arguments as for a temporal process. Commonly, areas that are close have similar characteristics such as climate, demographics, and public health programs like cancer screenings and vaccination programs. For this project, areas are considered close if they are adjacent, and otherwise not. This is a simplification as it does not consider where in an area people live or how large the different areas are, but it is a simplification we are willing to make. Hence, the characteristics of a process are likely somewhat similar in adjacent areas. Therefore, it is desirable with an effect that is smooth in space, such that adjacent areas have similar values. This motivates the spatial random effect $\boldsymbol{\theta}^\top = (\theta_1, ..., \theta_n)$, which is assumed multivariate Gaussian with zero mean and precision matrix $\tau_\theta \boldsymbol{K}_\theta$. From Section 5.3 we know that the Besag model has smooth realizations in space. Therefore, we will model $\boldsymbol{\theta}$ as a Besag.

Similarly, as for a temporal process, a spatial process cannot entirely be modelled by structured effects. Certain areas may differ from their neighbours in a way that is not explainable by a structured effect. This could be due to a characteristic of this area, such as geography, climate, or demographics, being different for this area compared to its neighbours. Take for example the Freetown Christiania in Copenhagen. Freetown is a neighborhood in Copenhagen in which few laws are enforced. This lawlessness attracts different residents than Copenhagen as a whole. It is not reasonable to believe that any process having to do with demographics taking place in the neighbourhoods of Copenhagen, will have similar responses between its normal neighbourhoods and the Freetown Christiania. This is even though these areas are close. Therefore, it is also desirable with a non-structured spatial effect to capture the effects of areas that may deviate from structure. Hence, $\boldsymbol{\phi}^\top = (\phi_1, ..., \phi_n)$ is needed in the linear predictor. Again, $\boldsymbol{\phi}$ is also assumed multivariate Gaussian with mean zero. Since it is unstructured each ϕ_i should be independent all other ϕ_j, $i \neq j$. Therefore the structure matrix is $K_\phi = I_n$ where I_n is the identity matrix in $\mathbb{R}^{n \times n}$.

Having motivated the temporal and spatial random effects, it is necessary to define how we model both spatial or temporal structured and unstructured effects together. Consider the temporal effects $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ for now, the arguments will be the same for the spatial effects $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ so that will not be discussed. Firstly, from Section 6 we have seen that it is useful to use a scaled version of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^*$. Secondly, The structured term $\boldsymbol{\alpha}^*$ cannot be viewed independently from the unstructured term $\boldsymbol{\gamma}$. More precisely, the precision parameters τ_α and τ_γ cannot be seen independently from each other [Wakefield, 2006]. Therefore $\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}$ is modelled as a BYM2 model in this project thesis, see Riebler *and others* [2016] for details on the BYM2 model. The BYM2 model is a reparameterization of the classic BYM model introduced by Besag *and others* [1991]. Using the BYM2 model the temporal random effects are then included in the linear predictor as

$$\frac{1}{\sqrt{\tau}} \left(\sqrt{1-\phi}\gamma_i + \sqrt{\phi}\alpha_i^* \right)$$

where $\tau > 0$ is a common precision parameter of $\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}$, and $\phi \in [0, 1]$ is a mixing parameter. The total temporal variation is $\frac{1}{\tau}$, and the proportion of temporal variation that is structured and unstructured is ϕ and $1 - \phi$, respectively. By using this reparameterization, the model splits the temporal variance based on how much structure it exhibits in time. For $\phi = 0$ the process is completely unstructured in time, and all the temporal variation is explained by the unstructured term $\boldsymbol{\gamma}$. Whereas for $\phi = 1$ all the temporal variation is explained by the structured term $\boldsymbol{\alpha}^*$. The higher the value of ϕ the more structure the process exhibits in time, since more of the temporal variation is explained by $\boldsymbol{\alpha}^*$. The common precision parameter τ eases prior elicitation. The interpretation of τ is easier than it would be for two dependent precision parameters, as $\frac{1}{\tau}$ is the total temporal variation. The interpretation of two dependent precision parameters is not so straightforward. We will impose a PC-prior on τ for reasons outlined in Section 6.

Implementation of spatial and temporal effects in INLA

We now show how to implement the linear predictor in Equation (20) without the interaction term δ_{ti} in R-INLA. This implementation will consist of modelling the spatial and temporal effects as two BYM2 models. In R-INLA, a spatial or temporal model can be implemented through the f() function outlined as

```
f(x, <model>, scale.model = T, constr = T, <graph>, <hyper>)
```

Firstly, the random effect is specified in `x`. For a spatial component, this is usually an indicator of area, and for a temporal component, it is an indicator of time.
Then, a specification of which model the random effects follows is required. As detailed above, both our spatial and temporal random effects will follow the BYM2 model, therefore we replace `<model>` with `model='bym2'`.
The term `scale.model=T` states that we scale the structured component of the BYM2 model. For the temporal effects this means that we use the scaled version of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^*$, and for the spatial effects we use the scales version of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$.
The structured terms follow IGMRFs, therefore we impose a sum-to-zero constraint on the structured components to ensure identifiability of μ. This is done by setting `constr=T`. If we do not restrict both $\boldsymbol{\alpha}^*$ and $\boldsymbol{\theta}^*$ when they follow IGMRFs, then it is difficult to identify μ as it is confounded with the unspecified levels of $\boldsymbol{\alpha}^*$ and $\boldsymbol{\theta}^*$.
In `<graph>` the structure matrix of the structured random effect is provided.
Lastly, in `<hyper>` the priors on the hyperparameters are specified. As stated in Section 6 this project aims to use PC-priors. For the BYM2 model, there are two hyperparameters: the precision τ (called 'prec' in R-INLA) and the mixing parameter ϕ (called 'phi' in R-INLA). For both the spatial and the temporal BYM2 model the term `<hyper>` is replaced by `bym2_hyper` which we defined as

```
bym2_hyper =list(prec= list(prior = 'pc.prec',
                            param = c(1, 0.01)),
```

```
    phi = list(prior = 'pc',
               param = c(0.5, 0.5))
```

This means that for the precision, the upper bound on $\frac{1}{\sqrt{\tau}}$ is 1 with a probability of 0.01 a priori. For the mixing parameter the upper bound is 0.5 with a probability of 0.5. The prior on ϕ has probability 0.5 since we do not wish to state whether the process exhibits more structure or is more unstructured a priori, for further details see Riebler *and others* [2016].

Having defined how to implement a BYM2 model in R-INLA, we can now show how to implement the linear predictor in Equation (20) without the interaction term in R-INLA. We will consider two different versions of this model, one having $\boldsymbol{\alpha}$ as a RW1 and the other as a RW2. We will call the model having no interaction and using a RW1 for `Improper_1_noInt` and the one using a RW2 for `Improper_2_noInt`. Let $\boldsymbol{\alpha}$ follow a RW1 model with precision matrix called `rw1_prec` and $\boldsymbol{\theta}$ follow a Besag model with precision matrix called `icar_prec`. The `Improper_1_noInt` is implemented in R-INLA as

```
Improper_1_noInt <- response ~ 1 + f(time, 'bym2',
                                     scale.model = T, constr = T,
                                     graph = rw1_prec,
                                     hyper = bym2_hyper) +
                                   f(area, 'bym2',
                                     scale.model = T, constr = T,
                                     graph = icar_prec,
                                     hyper = bym2_hyper)
```

If $\boldsymbol{\alpha}$ were to have followed a RW2 instead we would have replaced `rw1_prec` with the precision matrix of the RW2 and called the model `Improper_2_noInt`.
In this case `response` is what we wish to model. The term 1 is the intercept μ in Equation (20). The first f() term is the temporal BYM2 model where `time` is the indicator for time. The second f() term is the spatial BYM2 model where `area` is the indicator for which area.
Having specified how we model the temporal and spatial random effects and how to implement them together, we will now introduce spatiotemporal interactions.

7.4 Spatiotemporal interactions

So far we have only discussed spatial and temporal random effects, and not considered spatiotemporal interactions. While it is possible to only use spatial and temporal random effects for spatiotemporal modelling, it is restrictive. It would assume that the variation of the response is separable into effects of time or effects of space. However, there may be variation in the response that is not separable into either time or space. Just as an example: If an area in the study is developing differently over time compared to the other areas, then the variation of the process is different over time for this area

compared to the others. A model using only spatial and temporal random effects would be incapable of capturing this. Hence, it is necessary with spatiotemporal interactions. In Knorr-Held [2000] four different specifications of the spatiotemporal interaction term were introduced, and these specifications have become common [Schmid and Held, 2004; Schrödle and Held, 2011b; Ugarte *and others*, 2014]. We now present the spatiotemporal interactions $\delta_{ti},\ t = 1, ..., T,\ i = 1, ..., n$, as introduced in Knorr-Held [2000].
The spatiotemporal interaction $\boldsymbol{\delta}^\top = (\delta_{11}, ..., \delta_{1n}, ..., \delta_{Tn})$ is assumed to be the result of one spatial and one temporal random effect interacting. The interaction $\boldsymbol{\delta}$ is assumed zero mean Gaussian with precision matrix $\tau_\delta \boldsymbol{K}_\delta$. In Knorr-Held [2000] $\boldsymbol{K}_\delta$ is defined as the Kronecker product of the precision matrices of the random effects that are assumed to interact. For the definition of the Kronecker product and relevant properties see Appendix A.1. The structure of the interactions is therefore assumed known a priori. Since there are two spatial and two temporal effects, there is a total of four different possible spatiotemporal interactions. Each one assumes a different structure for the interaction a priori. The four different specifications of $\boldsymbol{\delta}$ are called type I, type II, type III, and type IV, and each one will be presented now.

Type I interaction: Overdispersion adjustment for unobserved heterogeneity

If one assumes that it is the two unstructured effects $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ interacting, then one gets that $\boldsymbol{\delta}$ is completely unstructured. Each δ_{ti} is independent of any $\delta_{tj},\ j \neq i$ and $\delta_{t'i},\ t' \neq t$. They are independent since $\boldsymbol{K}_\delta$ is defined as

$$\boldsymbol{K}_\delta = \boldsymbol{K}_\gamma \otimes \boldsymbol{K}_\phi = \boldsymbol{I}_T \otimes \boldsymbol{I}_n = \boldsymbol{I}_{nT}$$

An illustration of the graph structure of a type I interaction is in the upper-left panel in Figure 6. This is a useful interaction to include if we believe that there is something unobserved that causes the process to vary more over space and time than what is captured by the Poisson variance. This could be due to incidents that occur independently in space and time, that have no structure. For example superspreader events for contagious diseases [Sneppen *and others*, 2021].
It is worth noting, that since both $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are proper, their precision matrices have full rank. Hence $\boldsymbol{\delta}$ has a precision matrix with full rank. This is quite obvious as the structure matrix for all three terms is the corresponding identity matrices, which all have full rank. Since $\boldsymbol{\delta}$ has a full-rank precision matrix it is proper.
We implement the linear predictor in Equation (20) with type I interaction in R-INLA by updating the formula for the linear predictor without the interaction term. If we want $\boldsymbol{\alpha}$ to be modelled as a RW1, we call the model `Improper_1_typeI` and it is implemented as

```
Improper_1_typeI <- update(Improper_1_noInt,
                           ~. + f(space.time,
                                  model="iid",
                                  hyper = interaction_hyper ))
```

Here the random effect is specified as `space.time`, which is a unique ID for each area and time pair. The `model` is specified to be `'iid'` producing independent and identically distributed effects as is the definition of the type I interaction. There is only one hyper-parameter for the interaction term, being the precision, τ_δ, of the interaction. The prior for τ_δ is a PC-prior specified in R-INLA as

```
interaction_hyper = list(theta=list(prior="pc.prec",
                                    param=c(1,0.01)))
```

This prior for the precision of the interactions will be identical for the different specifications of the interactions. If we want $\boldsymbol{\alpha}$ to follow a RW2, we call that model `Improper_2_typeI`, and implement it in a similar way to `Improper_1_typeI`. Only replace `Improper_1_noInt` with `Improper_2_noInt` in `update`, and call the formula `Improper_2_typeI`.

Type II interaction: Independent random walks for each area

If there is a reason to believe that the variation of the responses changes differently and independently over the areas with time then we let $\boldsymbol{\alpha}$ interact with $\boldsymbol{\phi}$. As a result $\boldsymbol{\delta}_i^\top = (\delta_{1i}, ..., \delta_{Ti})$ follows a random walk for every area i independently of the other areas. This could be a result of different policies or developments in the studied areas, such as vaccination programs or cancer screening programs. For example, consider deaths from prostate cancer. If some of the considered areas have prostate cancer screening programs, it could very well be that these areas experience changes over time for prostate cancer deaths that are different from the areas without screening programs [Potosky *and others*, 2001].

Since $\boldsymbol{\alpha}$ is improper and follows a random walk, the interaction is also improper. The rank of the structure matrix of $\boldsymbol{\alpha}$ is $T-k$ where k is the order of the Random Walk. The interaction term has a structure matrix defined by

$$\boldsymbol{K}_\delta = \boldsymbol{K}_{\alpha^*} \otimes \boldsymbol{K}_\phi = \boldsymbol{K}_\alpha \otimes \boldsymbol{I}_n \tag{21}$$

where $\boldsymbol{K}_{\alpha^*}$ is the structure matrix of the scaled version of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^*$. An illustration of the graph structure of a type II interaction is provided in the upper-right panel in Figure 6. It is worth noting that this illustration is dependent upon the graph structure provided by $\boldsymbol{\alpha}$ which can be seen below. Figure 6 does illustrate the graph structure of $\boldsymbol{\delta}$ correctly if $\boldsymbol{\alpha}$ follows a RW1. If it were to follow a RW2, the graph structure would be different. We can see from Figure 6 that the interactions in one area follow a structured temporal process, independently of the other areas as there are no edges between nodes of different areas.
Equation (21) implies that $\text{rank}(\boldsymbol{K}_\delta) = \text{rank}(\boldsymbol{K}_{\alpha^*})\text{rank}(\boldsymbol{I}_n) = (T-k) \times n$. This means we have to impose kn constraints on $\boldsymbol{\delta}$ to ensure identifiability of the overall mean μ and the spatial random effects. If $\boldsymbol{\alpha}$ follows a RW1, we impose a sum-to-zero constraint over

each Random Walk. I.e. for each area i, we require that $\boldsymbol{\delta}_i^\top = (\delta_{1i}, ..., \delta_{Ti})$ sums to zero

$$\sum_{t=1}^{T} \delta_{ti} = 0$$

A compact way of writing the constraints is

$$\boldsymbol{A\delta} = \mathbf{0} \tag{22}$$

where $\boldsymbol{A}$ is a $n \times nT$ matrix with entries

$$A_{ij} = \begin{cases} 1, & j \bmod(n) = i \bmod(n) \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

If $\boldsymbol{\alpha}$ follows a RW2 we could use sum-to-zero constraints for the RW2 of each area $\boldsymbol{\delta}_i$, and also constrain the slope of $\boldsymbol{\delta}_i$ by requiring

$$\sum_{t=1}^{T} t \cdot \delta_{ti} = 0$$

For this project, if $\boldsymbol{\alpha}$ follows a RW2 we define the rows of $\boldsymbol{A}$ in Equation (22) to be equal to the eigenvectors corresponding to the eigenvalues equal to zero of $\boldsymbol{K}_\delta$ in Equation (21).
Now we detail how to implement the linear predictor in Equation (20) with type II interaction in R-INLA. If $\boldsymbol{\alpha}$ follows a RW1 we call the model `Improper_1_typeII` and it is implemented as

```
Improper_1_typeII <- update(Improper_1_noInt,
                            ~. + f(space.time,
                                   model = "generic0",
                                   Cmatrix = typeII_prec,
                                   extraconstr = typeII_constraints,
                                   hyper = interaction_hyper))
```

Here as for the type I interaction, we have specified the effect as `space.time`. Furthermore, the `model` is specified as `'generic0'` which implements the precision matrix specified in `Cmatrix`. The precision matrix, `typeII_prec`, is constructed by taking the Kronecker product in Equation (21). The precision matrix of the structured temporal effect must be scaled before the Kronecker product, as we do not scale the precision matrix in the INLA call. The term `extraconstr` is where we specify our constraints. This constraint consists of a constraint matrix $\boldsymbol{A}$ in Equation (23), and a n dimensional zero-vector. If $\boldsymbol{\alpha}$ follows a RW2 we call the model `Improper_2_typeII`. It is implemented in the same way as `Improper_1_typeII`. Replace `Improper_1_noInt` with `Improper_2_noInt` in `upgrade`. The precision matrix must be that coming from Equation (21) where $\boldsymbol{K}_{\alpha^*}$ is the scaled structure matrix of a RW2, and the constraints are as detailed for $\boldsymbol{\alpha}$ following a RW2.

Type III interaction: Independent Besag for each time

Assuming that the spatial effects change from time to time independently of each time point, then one assumes that $\boldsymbol{\gamma}$ interacts with $\boldsymbol{\theta}$. For each time point $t \in \{1, ..., T\}$, $\boldsymbol{\delta}_t^\top = (\delta_{t1}, ..., \delta_{tn})$ then follows a Besag independently of all the other times. Hence, the spatial dependence changes independently from time to time. This could be a result of policies enacted only for a specific time for all or most of the regions, such as the national policies enacted during Covid-19 in Norway which were only in effect for some years. These policies may have changed the effect of space for the years they were enacted, for processes being affected by the lockdowns and quarantines.

Since $\boldsymbol{\gamma}$ is interacting with $\boldsymbol{\theta}$ we define the structure matrix of $\boldsymbol{\delta}$ as

$$\boldsymbol{K}_\delta = \boldsymbol{K}_\gamma \otimes \boldsymbol{K}_{\theta^*} = \boldsymbol{I}_T \otimes \boldsymbol{K}_{\theta^*} \tag{24}$$

where $\boldsymbol{K}_{\theta^*}$ is the structure matrix of the scaled version of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$. Equation (24) results in a diagonal block matrix that is $T \times T$ dimensional, where the entries on the diagonal are the structure matrix of an ICAR. Hence, we get that $\boldsymbol{\delta}$ follows an ICAR for each time t independently of the other times. An illustration of the graph structure of a type III interaction is shown in the lower-left panel in Figure 6. It is important to note here this graph structure is given the graph structure of $\boldsymbol{\theta}$. In Figure 6 the graph structure of $\boldsymbol{\theta}$ is simplified to just a straight line. In reality, it is often an irregular grid as in Figure 5b. However, this would be difficult to draw. What can easily be seen is that there are ICARs at each time point independent of the other time points as there are no edges between nodes of different times.

From Equation (24) we have that the $\text{rank}(\boldsymbol{K}_{\delta^*}) = \text{rank}(\boldsymbol{I}_T)\text{rank}(\boldsymbol{K}_\theta) = T \times (n-1)$. I.e. $\boldsymbol{\delta}$ is improper as it has a rank-deficiency of T. Therefore we need T constraints to ensure identifiability of μ and the temporal random effects. Since there is an ICAR for each time point independently of the other time points, we sum-to-zero over each of the ICARs. This is for each $t \in [1, ..., T]$ impose on $\boldsymbol{\delta}_t^\top = (\delta_{t1}, ...\delta_{tn})$ the constraint

$$\sum_{i=1}^{n} \delta_{ti} = 0$$

These constraints can be written as a matrix product as in Equation (22), by having $\boldsymbol{A}$ be a $T \times nT$ matrix defined as

$$A_{tj} = \begin{cases} 1, \ j \in \{(t-1) \cdot n + 1, ..., t \cdot n\} \\ 0, \ \text{otherwise} \end{cases} \tag{25}$$

and the right hand side in Equation (22) be a T dimensional vector of zeros.
Now we show how to implement the linear predictor in Equation (20) with a type III interaction in R-INLA. If $\boldsymbol{\alpha}$ is to follow a RW1, then call the model `Improper_1_typeIII` and it is implemented as

```
Improper_1_typeIII <- update(Improper_1_noInt,
                             ~. + f(space.time,
                                    model = "generic0",
                                    Cmatrix = typeIII_prec,
                                    extraconstr = typeIII_constraints,
                                    hyper = interaction_hyper))
```

This is mostly the same as the implementation of type II interactions. The difference is that the provided structure matrix is given in Equation (24), and the constraints in `typeIII_constraints` are a list consisting of a matrix as given in Equation (25) and a T dimensional vector of zeros. If $\boldsymbol{\alpha}$ is to follow a RW2, we call the model `Improper_2_typeIII` and it is implemented in the same way, just replacing `Improper_1_noInt` with `Improper_2_noInt` in the update call.

Type IV interaction: Completely structured

This interaction assumes that $\boldsymbol{\alpha}$ interacts with $\boldsymbol{\theta}$. Let $\boldsymbol{\alpha}$ follow a RW1 for now, as it simplifies the discussion. This interaction makes it so that not only does δ_{ti} depend upon first order neighbours which are either δ_{tj} where $i \sim j$ or $\delta_{t\pm1,i}$, but it also depends upon the second order neighbours $\delta_{t\pm1,j}$ where $i \sim j$. I.e. not only does δ_{ti} depend upon the areas close by at the same time t or the interaction in that area at its temporal neighbours $t \pm 1$, but it also depends upon the interaction in the adjacent areas at times $t \pm 1$. An illustration of the graph structure is provided in the lower-right panel in Figure 6. Again, as for the previous interactions, the illustrated graph structure of $\boldsymbol{\delta}$ is a simplification given the simplified graph structures of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. What it does illustrate nicely is how δ_{ti} is now dependent upon its spatial neighbours, its temporal neighbours, and its second-order spatiotemporal neighbours, as there are edges between these.
This dependence means that the type IV interaction has smoothed terms over both space and time. This structure incorporates how things such as political measures can have a spillover effect between adjacent areas at different times. Take as an example a vaccination program in one area aimed at combating a contagious disease. The vaccination program in that specific area will likely affect the prevalence of the disease in the area going forward in time. Since contagion is often spread between nearby areas, it is likely that this in turn will have an effect lowering the prevalence of the disease in adjacent areas at future times. Hence, the effect of the vaccination program propagates smoothly from the specific area with the program to the adjacent areas at future times. This interaction term can account for quite complex effects, but it is very rank-deficient. Its structure matrix is defined as the Kronecker product

$$\boldsymbol{K}_{\delta} = \boldsymbol{K}_{\alpha^*} \otimes \boldsymbol{K}_{\theta^*} \tag{26}$$

This means that $\text{rank}(\boldsymbol{K}_{\delta}) = \text{rank}(\boldsymbol{K}_{\alpha^*})\text{rank}(\boldsymbol{K}_{\theta^*}) = (T-k) \times (n-1)$. Constraints are now needed for the identifiability of μ and the temporal and spatial random effects. The constraints we impose on $\boldsymbol{\delta}$ is a combination of those used for type II interactions and type III interactions.

The implementation of the linear predictor in Equation (20) with type IV interactions in R-INLA is almost the same as the implementation of type II and type III interactions. If $\boldsymbol{\alpha}$ follows a RW1, we call the model `Improper_1_typeIV` and it is implemented as

```
Improper_1_typeIV <- update(Improper_1_noInt,
                            ~. + f(space.time,
                                   model = "generic0",
                                   Cmatrix = typeIV_prec,
                                   extraconstr = typeIV_constraints,
                                   hyper = interaction_hyper))
```

Here the provided precision matrix `typeIV_prec` is as defined in Equation (26). The constraints in `typeIV_constraints` are a list consisting of a $(n+T-1)\times nT$ matrix $\boldsymbol{A}$ where the first n rows are defined as the matrix in Equation (23) and the last $T-1$ rows are defined as the first $T-1$ rows of the matrix in Equation (25), and a $n+T-1$ dimensional vector of zeros. If $\boldsymbol{\alpha}$ follows a RW2 we call the model `Improper_2_typeIV`, and the constraints are made in the same way, only using the type II constraints used for $\boldsymbol{\alpha}$ following a RW2.

Table 1 contains a summary of the components in the spatiotemporal models as introduced in Knorr-Held [2000]. The table consists of which structure matrix the effects and interactions have, what their rank is, and how rank-deficient the components are.

	Structure matrix $\boldsymbol{K}_*$	Rank	Rank-deficiency
$\boldsymbol{\alpha}_{\text{RW1}}$	RW1 $\boldsymbol{K}_\alpha$: Equation (14)	$T-1$	1
$\boldsymbol{\alpha}_{\text{RW2}}$	RW2 $\boldsymbol{K}_\alpha$: Equation (16)	$T-2$	2
$\boldsymbol{\gamma}$	$\boldsymbol{I}_T$	T	0
$\boldsymbol{\theta}$	Besag $\boldsymbol{K}_\theta$: Equation (19)	$n-1$	1
$\boldsymbol{\phi}$	$\boldsymbol{I}_n$	n	0
Type I $\boldsymbol{\delta}$	$\boldsymbol{I}_T \otimes \boldsymbol{I}_n = \boldsymbol{I}_{nT}$	nT	0
Type II $\boldsymbol{\delta}$: $\boldsymbol{\alpha}_{\text{RW1}}$	$\boldsymbol{K}_{\alpha^*} \otimes \boldsymbol{I}_n$	$(T-1)\times n$	n
Type II $\boldsymbol{\delta}$: $\boldsymbol{\alpha}_{\text{RW2}}$	$\boldsymbol{K}_{\alpha^*} \otimes \boldsymbol{I}_n$	$(T-2)\times n$	$2n$
Type III $\boldsymbol{\delta}$	$\boldsymbol{I}_T \otimes \boldsymbol{K}_{\theta^*}$	$(n-1)\times T$	T
Type IV $\boldsymbol{\delta}$: $\boldsymbol{\alpha}_{\text{RW1}}$	$\boldsymbol{K}_{\alpha^*} \otimes \boldsymbol{K}_{\theta^*}$	$(n-1)\times(T-1)$	$n+T-1$
Type IV $\boldsymbol{\delta}$: $\boldsymbol{\alpha}_{\text{RW2}}$	$\boldsymbol{K}_{\alpha^*} \otimes \boldsymbol{K}_{\theta^*}$	$(n-1)\times(T-2)$	$2n+T-2$

Table 1: All the components of the linear predictor in Equation (20) and their different specifications. Shows their structure matrices, the rank of the structure matrices, and how rank-deficient the structure matrices are.

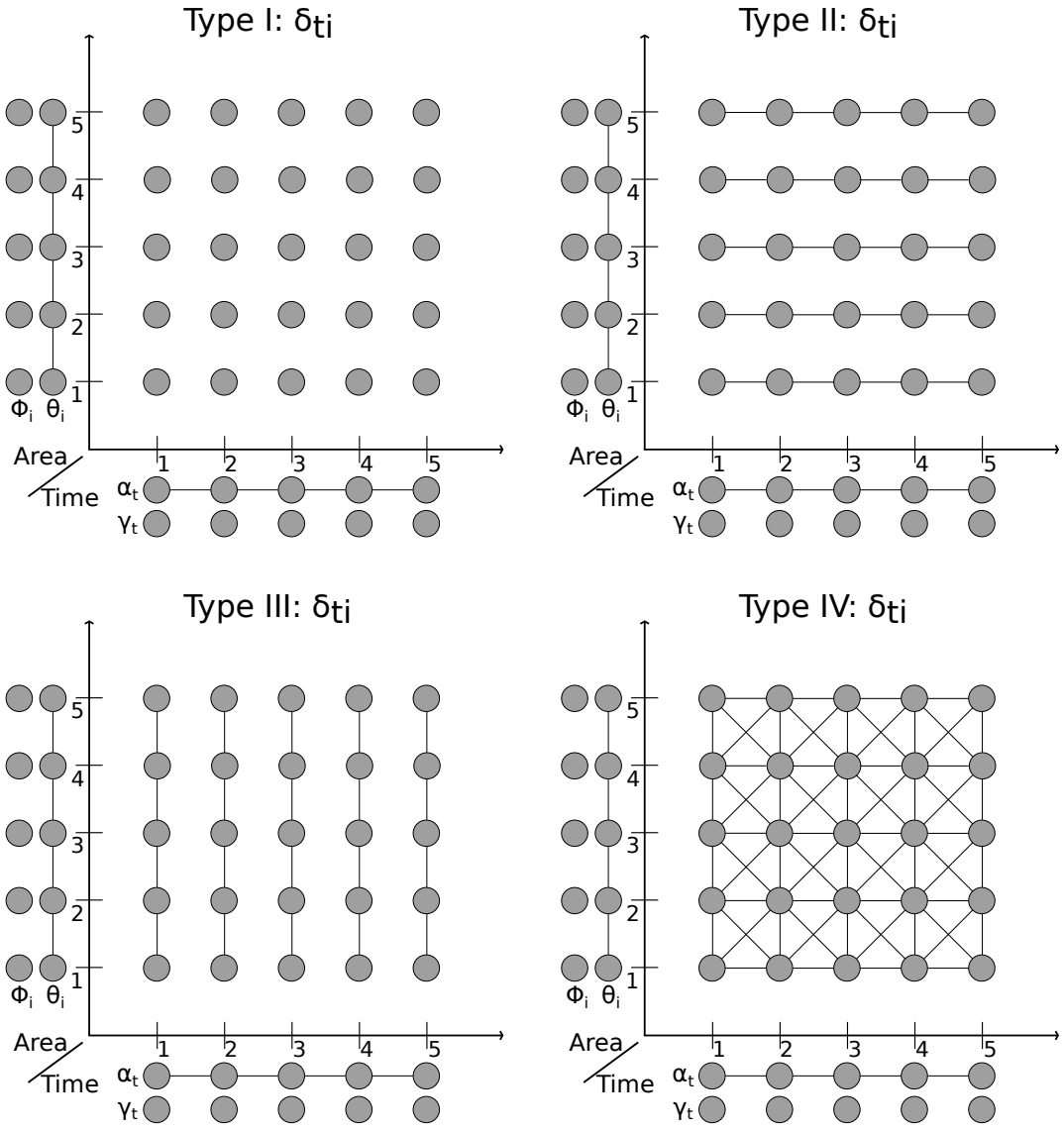


Figure 6: Illustration of the graph of $\boldsymbol{\delta}$ for the different specifications: (top-left) type I, (top-right) type II, (bottom-left) type III, and (bottom-right) type IV. For all the interactions, the graph structures of the temporal and spatial random effects are simplified.

7.5 Alternative specification of spatiotemporal interactions

So far we have used IGMRFs to model the structured random effects as this is common practice. However, using IGMRFs is not without problems. Namely, we have to constrain the improper effects and interactions. In Table 1 the rank-deficiency of the improper effects and constraints are reported. For the interactions the higher the rank-deficiency the more constraints are needed for identifiability of μ and the spatial and temporal effects. As can be seen, for increasingly complex models with increasing rank-deficiency the number of constraints needed grows rapidly. This becomes a computational issue when the number of constraints needed becomes large. We aim to compare these improper specifications with alternative models relying on proper effects and interactions. We now present three different proper spatiotemporal models. The first will be analogue to the `Improper_1_noInt` model in that it will only have spatial and temporal effects, no spatiotemporal interaction. It will act as a baseline to assess the proper spatiotemporal interaction. Then a second model is presented which has no analogue in the improper models, as it will only consist of a temporal fixed effect and a proper spatiotemporal interaction. The proper spatiotemporal interaction will however be similar to the type I-IV interactions as presented earlier, depending upon the values of the hyperparameters of the interaction. Lastly, a third model is presented combining the previous two. This model is analogue to the linear predictor (20).

Proper Spatiotemporal Model without Interactions

This model consists of proper random effects, but no interactions. The linear predictor is of the form

$$\eta_{it} = \mu + \beta \cdot t + \alpha_t + \theta_i$$

where μ is the overall risk, same as for Equation (20). Furthermore, α_t is modelled as an AR1, see Section 4.1, and β is a fixed temporal effect. The reason we have added β is that $\boldsymbol{\alpha}$ is stationary when it is modelled as an AR1, thus the addition of $\beta \cdot t$ allows the model to capture linear temporal trends. We could add more complex temporal fixed effects to enable the model to capture more complicated trends, but we will not concern ourselves with that in this project. The reason for the inclusion of α_t is similar to why a random walk component is included in the improper models, we believe that a process will have similar characteristics at close times. As seen in Section 4.1, the AR1 model can produce smoothed realizations in time. One major difference from the improper models is that we do not have an iid temporal effect. This is because the correlation parameter ρ in the AR1 model controls how correlated α_t is to α_{t-1}. Therefore a lower value of ρ means that there is a lower degree of structure, and $\boldsymbol{\alpha}^\top = (\alpha_1, ..., \alpha_T)$ will be less smooth in time. This means that ρ implicitly captures the process's degree of temporal structure. Lastly, $\boldsymbol{\theta}^\top = (\theta_1, ..., \theta_n)$ follows a Leroux model, see Section 5.2. Thus, $\boldsymbol{\theta}$ will also produce smoothed realizations in space, something which is desirable. Similarly as for $\boldsymbol{\alpha}$ we do not have an iid spatial effect here. This is because the Leroux model has a correlation parameter λ. This correlation parameter determines how much of the precision in space is structured and how much is unstructured. Therefore, we have

not included a spatial iid random effect. Note that for the proper models, the order of the indices has switched from *ti* for the improper models, to *it* here. This is because of how we will implement the proper interactions.

We call this model `Proper_noInt` and implement it in R-INLA as

```
proper_noInt <- response ~ 1 + time +
                           f(time.copy,
                             model = "ar1",
                             hyper = ar1_hyper) +
                           f(area,
                             model = "besagproper2",
                             graph = icar_prec,
                             hyper = spatial_hyper)
```

Same as for the improper models `response` is what we wish to model, `time` is the identifier of time, and `area` is the identifier of which area. The term `time.copy` is just a copy of `time` as R-INLA needs separate identifiers for different effects. In the formula `1` is mean μ, and `time` implements the fixed effect β. The first `f()` term is for $\boldsymbol{\alpha}$ following an AR1. The AR1 model has two hyperparameters, the precision τ and the correlation parameter ρ. As stated earlier we impose a PC-prior on our precision parameters, and for ρ we use a Gaussian prior. These hyperpriors are specified in R-INLA as

```
ar1_hyper = list(prec = list(prior = 'pc.prec',
                             param = c(1, 0.01)),
                 rho = list(prior = 'normal',
                            param = c(0, 0.25)))
```

The second `f()` term is $\boldsymbol{\theta}$, which follows a Leroux model and therefore `model='besagproper2'`. We provide this model with the precision matrix of a corresponding Besag model, as `graph=icar_prec`. The Leroux model has two hyperparameters, the precision τ, and the spatial correlation parameter λ. The chosen priors and how they are implemented in R-INLA are as follows

```
spatial_hyper = list(prec= list(prior = 'pc.prec',
                                param = c(1, 0.01)),
                     lambda = list(prior = 'gaussian',
                                   param = c(0, 1)))
```

For the hyperparameter ρ for the AR1 and λ for the Leroux model, PC-priors such as that for the mixing parameter ϕ in the BYM2 model can also be used but were not for this project. This model is not expected to perform that well on a spatiotemporal problem, as it assumes variation of response is separable in time and space, therefore we introduce a proper spatiotemporal interaction model.

Proper Spatiotemporal Interaction Model

The second proper model considered has a linear predictor of the form

$$\eta_{it} = \mu + \beta \cdot t + \delta_{it}$$

This linear predictor consists only of the fixed effects μ and β and a spatiotemporal interaction term δ_{it}. The term δ_{it} is now supposed to encapsulate both the spatial and the temporal random variation.

To model $\boldsymbol{\delta}^\top = (\delta_{11}, ..., \delta_{1T}, ..., \delta_{nT})$ we will take inspiration from the first order temporal trend model in Vivar and Ferreira [2009]. Let $\boldsymbol{x}^\top = (\boldsymbol{x}_1, ..., \boldsymbol{x}_T)$ follow the first order temporal trend model. Then $\boldsymbol{x}_t^\top = (x_{t1}, ..., x_{tn})$ for $t \in [2, ..., T]$ is defined as

$$\boldsymbol{x}_t = \rho \cdot \boldsymbol{x}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \overset{\text{iid}}{\sim} \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{R}^{-1}) \tag{27}$$

where $|\rho| < 1$, $\boldsymbol{R} = \tau((1-\lambda)I + \lambda \boldsymbol{K})$, and $\boldsymbol{K}$ is the structure matrix of a Besag model Section 5.3. As a result $\boldsymbol{\omega}_t$ is following a Leroux model independently for each time point t, see Section 5.2. Clearly the first order temporal trend model is similar to an AR1 model, see Equation (10). By modelling $\boldsymbol{x}$ as in Equation (27), $\boldsymbol{x}$ can be smooth in both space and time. This way of modelling the spatiotemporal interaction can capture how an effect in one area i at a time t, such as a vaccination program, can have a spillover effect for adjacent areas $j \sim i$ at time $t+1$. The temporal correlation parameter ρ and the spatial correlation parameter λ determine the degree of smoothness. Let $\boldsymbol{Q}_{\text{AR1}}$ be the precision matrix of an AR1 process with T nodes, then the precision matrix of $\boldsymbol{x}$ can be derived as

$$\boldsymbol{Q}_{\text{AR1}} \otimes \boldsymbol{R} = \tau \begin{bmatrix} \boldsymbol{R} & -\rho\boldsymbol{R} & & & \\ -\rho\boldsymbol{R} & (1+\rho^2)\boldsymbol{R} & -\rho\boldsymbol{R} & & \\ & \ddots & \ddots & \ddots & \\ & & -\rho\boldsymbol{R} & (1+\rho^2)\boldsymbol{R} & -\rho\boldsymbol{R} \\ & & & -\rho\boldsymbol{R} & \boldsymbol{R} \end{bmatrix}$$

When we model $\boldsymbol{\delta}$, we will construct the precision matrix by taking the product $\boldsymbol{R} \otimes \boldsymbol{Q}_{\text{AR1}}$. Since the Kronecker product is permutation equivalent, see Appendix A.1, this means that our way of modelling $\boldsymbol{\delta}$ is equivalent to the first order temporal trend model in Vivar and Ferreira [2009]. The degree of correlation between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ is decided by the value of ρ, and the degree of correlation between each node in $\boldsymbol{x}_t$ is decided by λ. This makes this interaction specification very flexible in capturing different levels of correlation. For $\lambda = 0$ and ρ close to 1, then $\boldsymbol{\delta}$ is similar to the type II interaction. For λ close to 1 and $\rho = 0$, then $\boldsymbol{\delta}$ is similar to the type III interaction. If both λ and ρ are close to 0 then $\boldsymbol{\delta}$ is similar to the type I interaction. If neither λ nor ρ is close to 0, then $\boldsymbol{\delta}$ is more similar to the type IV interaction. This proper interaction is very flexible and can produce smoothed realizations in space and time as well as less smooth realizations, depending upon the values of ρ and λ. Since the model uses an AR1, it is stationary.

That is why we include the $\beta \cdot t$ term in the linear predictor so that the model can capture linear trends. We call this model `Proper_onlyInt` and implement it in R-INLA as

```
proper_onlyInt <- response ~ 1 + time +
                  f(area, model = "besagproper2",
                    graph = icar_prec, hyper = spatial_hyper,
                    group = time, control.group = list(model = "ar1"))
```

As usual `response` is what we wish to model, `1` is the intercept μ, and `time` becomes the fixed effect β. The `f()` term is $\boldsymbol{\delta}$. The random effect is specified as `area` and model as `'besagproper2'` with graph given as `icar_prec` and hyperparameter priors are specified as `spatial_hyper` as for the previous model. Additionally we have included `group=time` with `control.group=list(model="ar1")`. This results in $\boldsymbol{\delta}$ being modelled similarly to that in Equation (27).

Proper Spatiotemporal Model with Random Effects and Interactions

The last alternative model we will consider is a combination of the two other ones. I.e.

$$\eta_{it} = \mu + \beta \cdot t + \alpha_t + \theta_i + \delta_{it}$$

For this model, μ, β, α_t, and θ_i are defined as in `Proper_noInt` and δ_{it} is defined as in `Proper_onlyInt`. This is analogous to the improper model with linear predictor as given in Equation (20). We call this model `Proper_full` and it is implemented in R-INLA as

```
proper_full <- response ~ 1 + time +
                          f(time.copy,
                            model = "ar1",
                            hyper = ar1_hyper) +
                          f(area,
                            model = "besagproper2",
                            graph = Besag_prec,
                            hyper = spatial_hyper) +
                          f(area.copy,
                            model = "besagproper2",
                            graph = Besag_prec,
                            hyper = spatial_hyper,
                            group = time,
                            control.group = list(model = "ar1"))
```

The only difference now is that for the third `f()` term, the spatial identifier must be a copy of the area identifier. This is because INLA requires effects to be unique, and therefore we must make a copy.

Having specified quite a few different models, Table 2 contains a full summary of what the improper and proper models are called and how they are all specified.

Model name	Formula specification
Improper_1_noInt	$\eta_{ti} = \mu + \alpha_{t:\text{RW1}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}}$
Improper_1_typeI	$\eta_{ti} = \mu + \alpha_{t:\text{RW1}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}} + \delta_{ti:\text{iid}\otimes\text{iid}}$
Improper_1_typeII	$\eta_{ti} = \mu + \alpha_{t:\text{RW1}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}} + \delta_{ti:\text{RW1}\otimes\text{iid}}$
Improper_1_typeIII	$\eta_{ti} = \mu + \alpha_{t:\text{RW1}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}} + \delta_{ti:\text{iid}\otimes\text{Besag}}$
Improper_1_typeIV	$\eta_{ti} = \mu + \alpha_{t:\text{RW1}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}} + \delta_{ti:\text{RW1}\otimes\text{Besag}}$
Improper_2_noInt	$\eta_{ti} = \mu + \alpha_{t:\text{RW2}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}}$
Improper_2_typeI	$\eta_{ti} = \mu + \alpha_{t:\text{RW2}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}} + \delta_{ti:\text{iid}\otimes\text{iid}}$
Improper_2_typeII	$\eta_{ti} = \mu + \alpha_{t:\text{RW2}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}} + \delta_{ti:\text{RW2}\otimes\text{iid}}$
Improper_2_typeIII	$\eta_{ti} = \mu + \alpha_{t:\text{RW2}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}} + \delta_{ti:\text{iid}\otimes\text{Besag}}$
Improper_2_typeIV	$\eta_{ti} = \mu + \alpha_{t:\text{RW2}} + \gamma_{t:\text{iid}} + \theta_{i:\text{Besag}} + \phi_{i:\text{iid}} + \delta_{ti:\text{RW2}\otimes\text{Besag}}$
Proper_noInt	$\eta_{it} = \mu + \beta \cdot t + \alpha_{t:\text{AR1}} + \theta_{i:\text{Leroux}}$
Proper_onlyInt	$\eta_{it} = \mu + \beta \cdot t + \delta_{it:\text{Leroux}\otimes\text{AR1}}$
Proper_full	$\eta_{it} = \mu + \beta \cdot t + \alpha_{t:\text{AR1}} + \theta_{i:\text{Leroux}} + \delta_{it:\text{Leroux}\otimes\text{AR1}}$

Table 2: Name of all models and their specification of the linear predictor. The components of the linear predictor have their specification in subscript.

8 Application on the Ohio lung cancer data set

In this project, we aim to highlight how different specifications of spatiotemporal interactions compare on an appropriate real-world data set. The previous section detailed the spatiotemporal models we will look at, see Table 2. This section begins by introducing the Ohio lung cancer data set, before detailing how performance is evaluated. The first evaluation is model choice criteria. The second evaluation is predictive performance. After having presented how the models are evaluated, the results of the Ohio lung cancer data set are reported and discussed.

8.1 The Ohio lung cancer data set

To compare the models we have chosen a widely used and popular data set: Yearly lung cancer death counts in the counties of Ohio stratified by gender and race from 1968 until 1988. The data is collected from `https://github.com/Paula-Moraga/SpatialEpiApp/tree/master/inst/SpatialEpiApp/data/Ohio`. The data set will be referred to as the Ohio lung cancer data set.
The raw data consists of 7392 observations, and each observation consists of 7 variables: `county`, `gender`, `race`, `year`, `y`, `n`, and `NAME`. The variable `year`$\in \{1968, ..., 1988\}$ indicates what year the observation was made, `NAME` and `county`$\in \{1, ..., n\}$ where $n = 88$ is the name and ID of the county the observation was made in. The lung cancer death counts are stratified by `gender` and `race`, which indicates which gender (male, female) and race (white, colored) the counts were observed for. The observed lung cancer death counts are `y`, and `n` is the population of the gender and race stratum in the county that year the observation was made.
For our analysis, the data is aggregated over the strata `gender` and `race`. After aggregating there are a total of 1848 observations where `y` and `n` are the total lung cancer death count and population in the observed county that year, respectively. We rename `year` to `Year` and add a index for year called `year`$\in \{1, ..., T\}$, $T = 21$. For clarity `y` is renamed to `deaths` and `n` to `pop_at_risk`. To implement the improper interactions it is necessary with a unique identifier for each observation, so the term `space.time`$\in \{1, ..., n \cdot T\}$, $n \cdot T = 1848$ is added. Lastly the observed death rate is calculated as `rate` $= \frac{\texttt{deaths}}{\texttt{pop\_at\_risk}}$ and added. Note here that when talking about the observed death rate, this is short-hand for the observed death count divided by the population. It is not an observation of the true underlying lung cancer death rate. This distinction is important, because if the true death rate was observed then there would be little use for the subsequent analysis. For the remainder of the project, when talking of observed rate it means the observed death count divided by the population. The first five entries of the transformed data that the analysis is done on can be seen in Table 3.
In addition to the Ohio lung cancer count data, a shapefile containing geographical polygons of the counties of Ohio was also extracted from `https://github.com/Paula-Moraga/SpatialEpiApp/tree/master/inst/SpatialEpiApp/data/Ohio`. These polygons are used to find the adjacent counties, which are used to construct the structure matrices of the spatial effects in the models. The polygons are also used to create heat maps over Ohio.

The geographical polygons can be seen in Figure 5b along with the undirected graph constructed by letting each county be a node and adjacent counties having edges between their respective nodes.

Year	year	name	county	deaths	pop_at_risk	rate	space.time
1968	1	Adams	1	6	17952	0.00033	1
1968	1	Allen	2	32	111028	0.00029	2
1968	1	Ashland	3	15	43504	0.00034	3
1968	1	Ashtabula	4	27	95077	0.00028	4
1968	1	Athens	5	12	54180	0.00022	5

Table 3: First five entries of the Ohio lung cancer data that the analysis is done on.

We are interested in modelling `deaths` as a spatiotemporal process. The count of lung cancer deaths in year t in county i, y_{ti}, is assumed

$$y_{ti}|\lambda_{ti} \sim \text{Poisson}(\texttt{pop\_at\_risk}_{ti} \cdot \lambda_{ti})$$

where $\texttt{pop\_at\_risk}_{ti}$ acts as a population offset, and λ_{ti} is the death rate from lung cancer which is unknown. When modelling we assume that $\log(\lambda_{ti}) = \eta_{ti}$ where η_{ti} is defined in the previous section.

Now, explorative plots of the data are presented. In Figure 7 the average and empirical standard deviation of the observed death rate per 100,000 aggregated over the years is plotted as a heatmap. The average observed rate looks somewhat smooth between the counties, motivating the use of a spatial model like the Leruox or Besag models. The empirical standard deviation of the observed rate is not that high for most counties, it is highest in the south-eastern counties. The differences in the empirical standard deviation of the observed rate indicate that the process varies differently over the counties, motivating the need for spatiotemporal interactions.
Figure 8 shows the observed death rate per 100,000 for each county for each year 1968,...,1988, along with the modified band depth median of the death rate. The modified band depth median of the death rate increases over the years, motivating the need for temporal effects. The observed death rates for each county fluctuate around the depth median, indicating that interactions are needed to capture the differences in rate for each county over the years.
The last explorative plot is Figure 9, which shows heatmaps of the observed lung cancer death rate per 100,000 for each county in Ohio and each year 1968,...,1988. One heatmap is the observed rate for each county in Ohio for one year. From Figure 9 it is clear that the observed rate increases over the years, as the heatmaps become increasingly red. It looks like a persistent pattern that the south-eastern counties have the largest observed death rates over the years.
Figures 7,8, and 9, all combined motivates the necessity for a spatiotemporal model to capture the variation in the problem. Furthermore, there are indications that interactions may prove useful in the modelling.

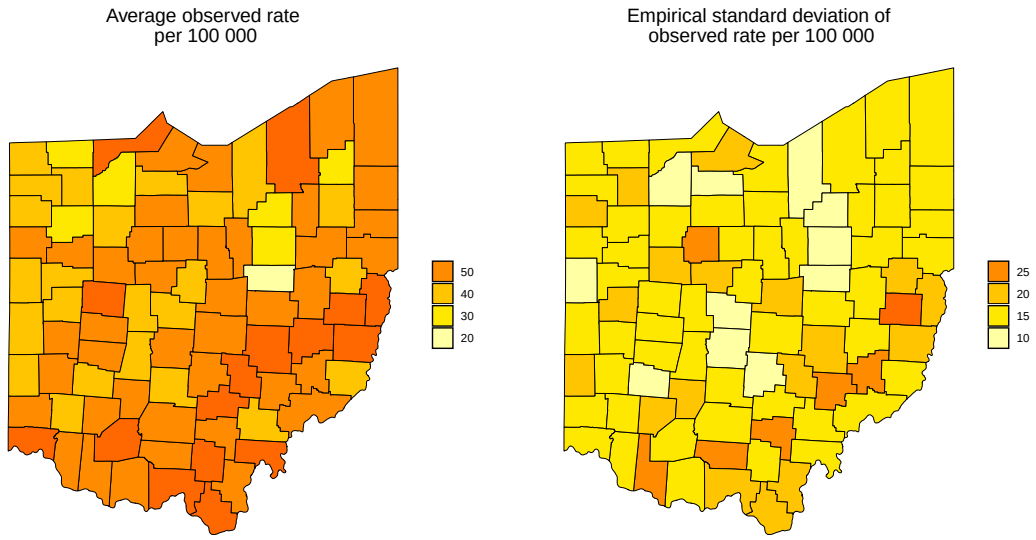


Figure 7: **(left)** The average observed lung cancer death rate per 100,000 aggregated over the years 1968 until 1988 for each county of Ohio **(right)** The empirical standard deviation of the observed lung cancer death rate per 100,000 aggregated over the years 1968 until 1988 for each county of Ohio.

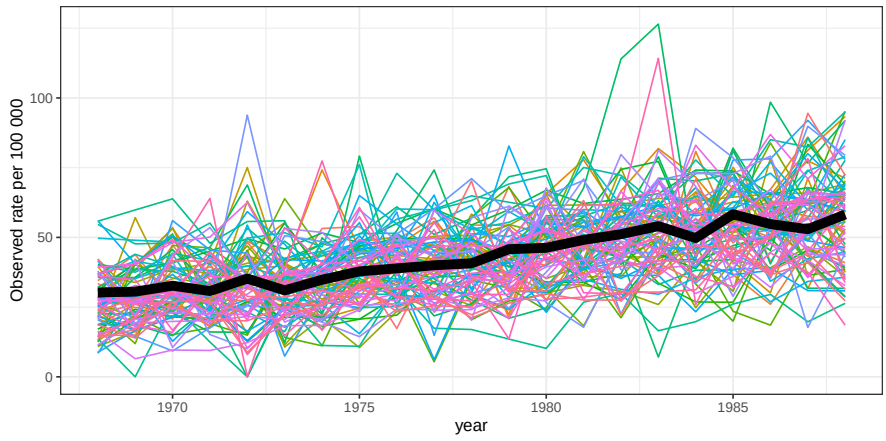


Figure 8: Observed lung cancer death rate per 100,000 for each county in Ohio (colored lines) over the years 1968 until 1988. The solid black line is the modified band depth median of the observed death rate of each county from 1968 until 1988.

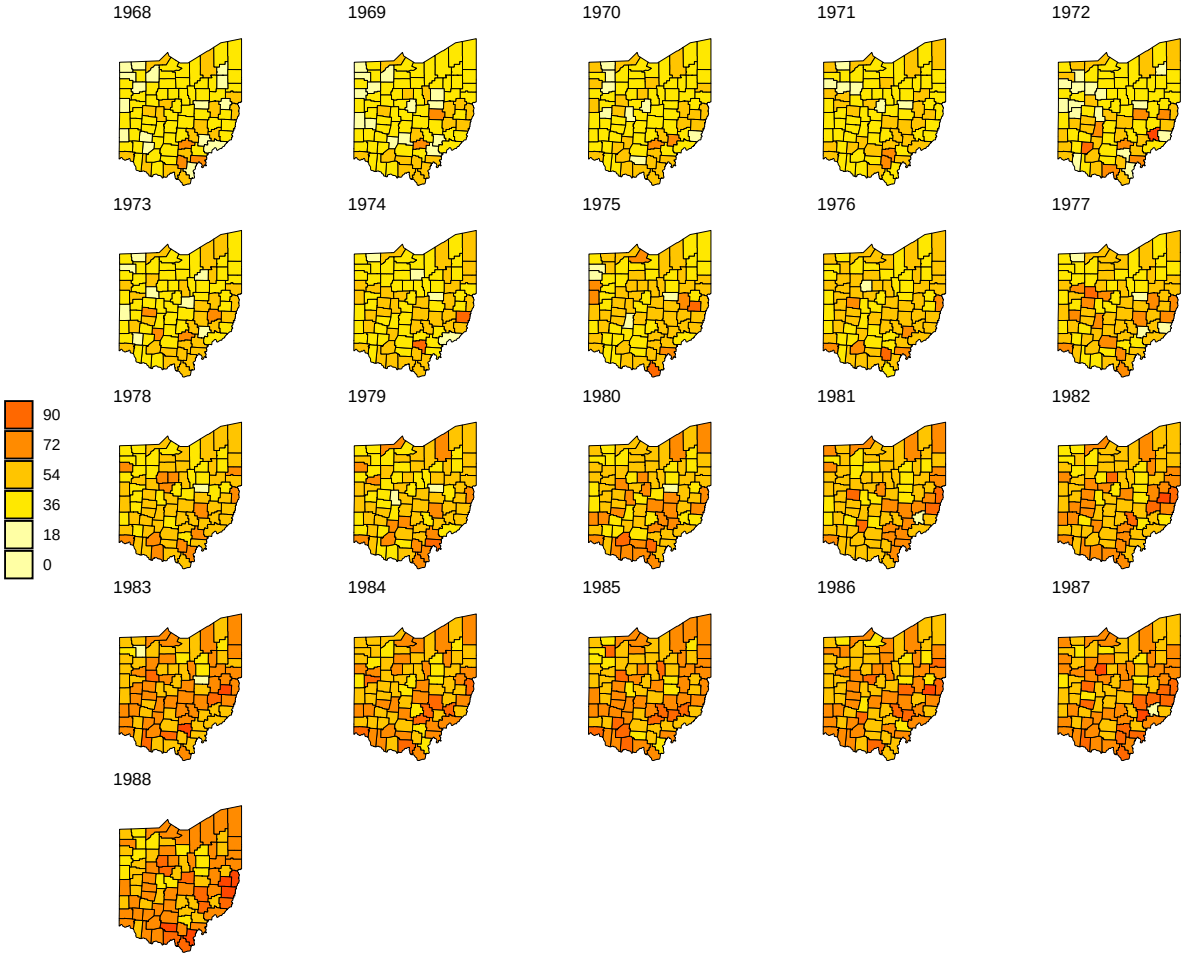


Figure 9: Observed lung cancer death rate per 100,000 in each county of Ohio, for each year 1968,...,1988.

8.2 Model selection

Model selection criteria are necessary to determine which model performs the best on the Ohio lung cancer data set. Two model selection criteria are used to evaluate the models: conditional predictive ordinates (CPO) and Watanabe-Akaike Information Criterion (WAIC) [Vehtari *and others*, 2017].

CPO is deemed the most important model selection criteria in this project as it is a proper scoring rule, see Appendix A.2 for theoretical details of proper scoring rules. Given observations $y_{ti},\ i = 1, ..., 88\ \ t = 1, ..., 21$, the conditional predictive ordinate of y_{ti} is calculated as

$$\text{CPO}_{ti} = \pi(y_{ti}|\boldsymbol{y}_{-ti})$$

where $\pi(y_{ti}|\boldsymbol{y}_{-ti})$ is the posterior predictive distribution given the observations $\boldsymbol{y}_{-ti} = \boldsymbol{y} \setminus \{y_{ti}\}$. CPO is therefore a leave-one-out cross-validation criterion, as CPO_{ti} is computed for all $t \in \{1, ..., 21\}$ and $i \in \{1, ..., 88\}$. The logarithmic score can be linked to CPO as

$$\log(S(P_{ti}, y_{ti})) = \log(\pi(y_{ti}|\boldsymbol{y}_{-ti}))$$

The logarithmic score is a proper and effective scoring rule [Gneiting and Raftery, 2007], see Appendix A.2 for the definition of an effective scoring rule. The logarithmic score is effective as it gives a higher expected score to predictive distributions closer to the true distribution. The following summary statistic of CPO will be calculated

$$S_{Tn} = -\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} log(\pi(y_{ti}|\boldsymbol{y}_{-ti}))$$

Therefore, the model resulting in the lowest S_{Tn} is the best. When the results are presented it is the summary statistic of CPO, S_{Tn}, which is presented.

The second model choice criteria used to evaluate model performance on the Ohio lung cancer data set is WAIC [Vehtari *and others*, 2017]. The WAIC for a model is calculated as

$$-\widehat{\text{elpd}}_{\text{WAIC}} = -\widehat{\text{lpd}} + \hat{p}_{\text{WAIC}}$$

where $\widehat{\text{elpd}}_{\text{WAIC}}$ is the estimated expected log point-wise predictive density, $\widehat{\text{lpd}}$ is estimated log predictive density, and $\hat{p}_{\text{WAIC}}$ is estimated number of effective parameters and acts as a penalization for increased model complexity. A model having low WAIC, $-\widehat{\text{elpd}}_{\text{WAIC}}$, is said to perform better. Information criteria such as WAIC are criticized as they under-penalize complex models. This is because $\hat{p}_{\text{WAIC}}$ can in many cases be too low, and WAIC is therefore not penalizing complexity enough.

8.3 One-year ahead predictions and prediction evaluation

It is interesting to compare the predictive performance of the models. In this project the one-year ahead predictive performance is assessed. To start the data for the years

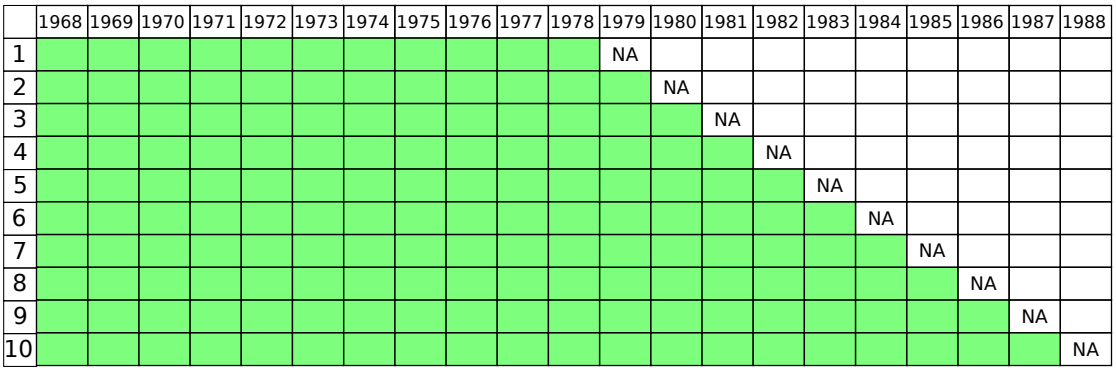


Figure 10: Illustration of this project's one-year ahead prediction setup. Each row corresponds to one one-year ahead prediction. There will be a total of 10 predictions, and hence there are 10 rows. Each column indicates a year of data, so the columns range from 1968 until 1988. For each row the cells colored green represent the years of data used to fit the model, the cell with `NA` represents the year predicted, and the white empty cells are data for years not considered and are therefore discarded for that prediction. For the cells with `NA` the population for each county is known, but the lung cancer death counts are unknown.

$1968, ..., 1978$ is used to predict the lung cancer death count for each county in Ohio in 1979. Then using the data for $1968, ..., 1978, 1979$ the lung cancer death count for each county in 1980 is predicted, and so on until the counts in 1988 are predicted. The one-year ahead prediction setup is illustrated in Figure 10.

Having obtained all the one-year ahead predictions for the models, the one-year ahead predictive performance of each model on the Ohio lung cancer data set is evaluated using mean absolute error (MAE), root mean squared error (RMSE), and interval score (IS) [Orozco-Acosta *and others*, 2023].

After having fitted the models according to the one-year ahead prediction setup in Figure 10, the posterior predictive distribution for the rate of lung cancer deaths $\lambda_{ti}|\boldsymbol{y}_{1:t-1}$, $\boldsymbol{y}_{1:t-1}^{\top} = (y_{11}, ..., y_{1n}, ..., y_{t-1,n})$, in each county $i \in \{1, ..., n\}$ for each year predicted on $t \in \{1979, ..., 1988\}$ is obtained. Following Orozco-Acosta *and others* [2023] the MAE and RMSE of a model can then be calculated as

$$\text{MAE} = \frac{1}{n \cdot 10} \sum_{t=1979}^{1988} \sum_{i=1}^{n} |y_{ti} - \hat{y}_{ti}| \quad \text{RMSE} = \sqrt{\frac{1}{n \cdot 10} \sum_{t=1979}^{1988} \sum_{i=1}^{n} (y_{ti} - \hat{y}_{ti})^2}$$

where $\hat{y}_{ti} = n_{ti} \cdot \mathbb{E}[\lambda_{ti}|\boldsymbol{y}_{1:t-1}]$ is the expected lung cancer death count from the posterior predictive distribution for county i and year t. Note that n_{ti}, the population in county i for year t, is known.

While MAE and RMSE tell how close the point predictions are from the true values, they do not consider how well-calibrated the model is. A point prediction may be close to the true value, but assign a low probability to the true value. Therefore it is desirable with a prediction evaluation criterion that measures both how well-calibrated the models are and their sharpness. This motivates the use of interval scores [Orozco-Acosta *and others*, 2023].

Let l and u be the 2.5% and 97.5% quantiles of the posterior predictive distribution, $y_{ti}|\boldsymbol{y}_{1:t-1}$, of the lung cancer death count for county i at year t, respectively. From the one-year ahead predictions the posterior predictive distribution of the rate $\lambda_{ti}|\boldsymbol{y}_{1:t-1}$ is obtained. The quantiles l and u are estimated. To estimate the quantiles 5000 samples of the rate are sampled from $\lambda_{ti}|\boldsymbol{y}_{1:t-1}$ using the function `inla.rmarginal` from R-INLA. For each sample λ^*_{ti}, one count y^*_{ti} is sampled from a Poisson distribution as $y^*_{ti} \sim \text{Poisson}(n_{ti}\lambda^*_{ti})$. the posterior predictive quantiles l and u can then be estimated as the empirical 2.5% and 97.5% quantiles of the 5000 samples of y^*_{ti}. The interval $[l, u]$ is a estimated 95% credible interval (CI) for the posterior predictive distribution $y_{ti}|\boldsymbol{y}_{1:t-1}$. The IS of a model given observed value y_{ti} is defined in Orozco-Acosta *and others* [2023] as

$$\text{IS}_{0.05}(y_{ti}) = (u-l) + \frac{2}{0.05}(l-y_{ti})\text{I}(y_{ti} < l) + \frac{2}{0.05}(y_{ti}-u)\text{I}(y_{ti} > u)$$

where $\text{I}(\cdot)$ is the indicator function. The sharpness of the distribution is accounted for by the term $u-l$. The larger the 95% CI the larger the value of $u-l$ is. Well-calibration is accounted for by giving higher IS to models where the true observation lies outside the 95% CI. The average IS is calculated for each model using all the predictions. The model resulting in the lowest average IS is the best at predicting.

8.4 Results on the Ohio lung cancer data set

The results obtained on the Ohio lung cancer data set using the models in Table 2 are now presented. First, the model choice results are presented and discussed. Then the results of the one-year ahead prediction scheme as outlined in Figure 10 are presented and discussed. The code used to fit the models was run on NTNUs server Markov. Information about Markov can be found here: `https://wiki.math.ntnu.no/drift/stud/ommarkov`.

Model choice results

In Table 4 the model choice criteria CPO, WAIC, and computational time in seconds on Markov are reported. From Table 4 `Improper_1_typeIV` performs the best with respect to CPO, having the lowest value at 3.134. `Improper_1_typeIV` and `Proper_full` perform the best with respect to WAIC, having the lowest values at 11576. None of the models have a high computational cost, only `Proper_full` has a computational time over one minute at about 500 seconds. It is clear from Table 4 that the inclusion of spatiotemporal interactions enhances the models, as all the models with interac-

tions outperform their respective counterparts without interactions. Interactions improve the ability of the models to capture the variation in lung cancer death counts in Ohio from 1968 to 1988. `Improper_1_typeIV` performs the best with respect to CPO, but `Improper_1_typeII`, `Proper_full`, and `Proper_onlyInt` are very close in performance. The same is true for WAIC, but for WAIC `Proper_full` and `Improper_1_typeIV` are tied. It is notable that `Proper_onlyInt` almost performs equally as well as `Proper_full` and `Improper_1_typeIV`. This indicates that the proper specification of the spatiotemporal interaction is flexible and capable of capturing variation in both space and time. A drawback of the model `Proper_onlyInt` is that it is harder to state what are the effects of space or time. The fact that `Improper_1_typeIV`, `Proper_full`, and `Proper_onlyInt` perform the best with respect to model choice suggests that the graph structure in the bottom right panel in Figure 6 is well suited for the Ohio lung cancer data set. The model `Improper_1_typeII` performs almost equally as well with respect to model choice as the very best models. This suggests that while the added complexity in the proper interaction and the type IV interaction performs better, it does not increase performance that much from the type II interactions. It could be interesting to see how a proper interaction akin to the type II interaction would have performed. The models using a RW2 for the structured temporal effect all performed worse, indicating that the added smoothing and structure imposed by a RW2 are unsuited for this problem. This is especially the case for `Improper_2_typeII` and `Improper_2_typeIV`, which did not experience as much of an improvement from the model `Improper_2_noInt` as the models using a RW1 with type II and type IV interactions did from their baseline. The higher computational time of `Proper_full` is likely because of difficulties in locating the mode of the hyperparameters. Surprisingly, the improper models with type IV interactions were as fast as they were. They were expected to be computationally demanding, but recent updates on the approximation strategy of the R-INLA software are likely the reason why the improper models were not that expensive.

Figure 11 shows the posterior median values of the structured temporal effects, along with 95% credible intervals for the models `Improper_1_noInt`, `Improper_2_noInt`, and `Proper_noInt`. As can be seen, the effect of time increases with the years. This implies that the death rate of lung cancer increased from 1968 until 1988. This is expected as Figure 8 indicates an increasing rate over the years. The trend of increasing temporal effects could be due to an increase in lung cancer incidents over the years, which could be due to an aging population or more people getting diagnosed with cancer [Weir *and others*, 2015]. The posteriors of the structured temporal random effects are quite similar for both the RW1 and RW2. Surprisingly, the RW1 is smoother in time than the RW2. This is surprising as a RW2 should produce smoother realizations of α_t as it borrows information also from its second-order neighbours. This could be one of the reasons why the specifications with a RW2 performed poorly compared to the other specifications. The posterior temporal effects of `Proper_noInt`, $\alpha_t + \beta t$, have a similar shape to that of the RW1 and RW2 specifications. This is expected as they are all meant to capture the same temporal effects. The value of the proper temporal effect starts at a positive

Model	CPO	WAIC	Computational time (s)
Improper_1_noInt	3.151	11646	9.100
Improper_1_typeI	3.147	11606	9.757
Improper_1_typeII	3.137	11584	12.489
Improper_1_typeIII	3.147	11612	15.873
Improper_1_typeIV	3.134	11576	24.538
Improper_2_noInt	3.152	11648	8.699
Improper_2_typeI	3.148	11608	10.820
Improper_2_typeII	3.149	11636	18.390
Improper_2_typeIII	3.147	11613	17.549
Improper_2_typeIV	3.148	11633	44.881
Proper_full	3.135	11576	498.500
Proper_noInt	3.151	11644	3.006
Proper_onlyInt	3.135	11577	38.967

Table 4: Model choice results for the Ohio lung cancer data set. Reported results are CPO, WAIC, and computational time in seconds on NTNUs server Markov.

value, whereas the improper temporal effect, starts at a negative value and becomes positive. This difference is due to the sum-to-zero constraint imposed on the improper effects, while the proper temporal effects are not constrained. The posterior densities of the hyperparameters of the temporal effects for the models can be seen in Figure 17 in Appendix C. The posterior distribution of the temporal mixing parameter ϕ in the BYM2 model for both `Improper_1_noInt` and `Improper_2_noInt` have the bulk of their probability mass close to 1. The posterior density of the temporal correlation parameter ρ for both `Proper_noInt` and `Proper_full` also have the bulk of their probability mass close to 1. This indicates that the effect of time on lung cancer death rates in Ohio exhibits a great degree of structure in time.

The posterior mean of the structured spatial random effects for `Improper_1_noInt` and `Proper_noInt` are plotted in Figure 12a as heatmaps over the counties of Ohio. Note that the coloring of the heatmaps is not on the same scale. The posterior mean of the proper spatial effects has larger absolute values compared to the posterior mean of the improper structured spatial effects. This likely stems from space having a lesser degree of structure in its effects on the lung cancer death rates. The posterior densities of the hyperparameters for the spatial effects of the proper and improper models can be seen in Figure 18 in Appendix C. Both the mixing parameter ϕ in the BYM2 model for the spatial effects of the improper models and the spatial correlation parameter λ in the Leroux model for the proper spatial effects, have posterior densities with the bulk of their probability mass at values smaller than 0.5. This implies that there is less structure to the spatial effects. The proper spatial effect implicitly contains unstructured effects, which explains why the posterior of the proper spatial effects and the posterior of the improper structured spatial effects are so different. The posterior mean of the improper

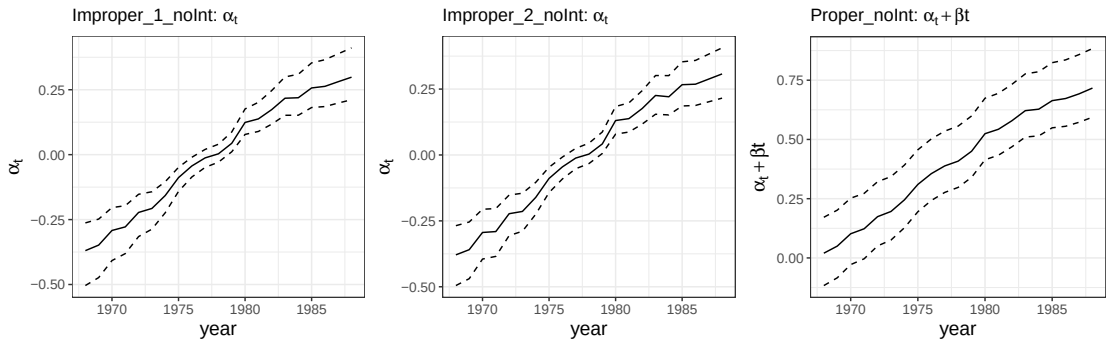


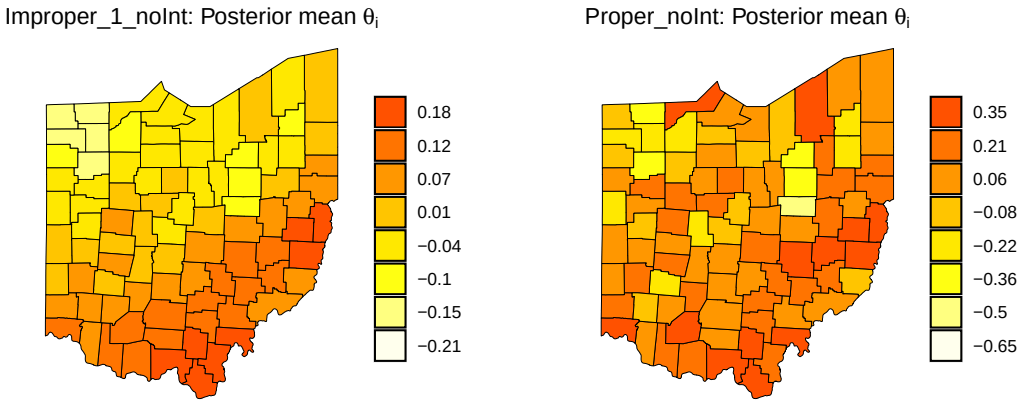
Figure 11: Posterior median values of the structured temporal random effects illustrated as a solid black line along with 97.5% and 2.5% quantiles shown as upper and lower dashed lines respectively for models (left) `Improper_1_noInt`, (middle) `Improper_2_noInt`, and (right) `Proper_noInt` where the temporal effect at time t is $\alpha_t + \beta t$. Note for βt that t starts at 1 when the year is 1968 and goes to 21 when the year is 1988.

structured spatial effects reveals a pattern in Ohio, with negative effects in the northern and western counties and positive effects in the southern and eastern counties. This is also captured by the posterior of the proper spatial effect, but it also has effects with positive posterior mean in the north due to it capturing the unstructured spatial effect as well. This implies that the southern and eastern counties of Ohio have higher rates of lung cancer deaths than the northern and western counties.

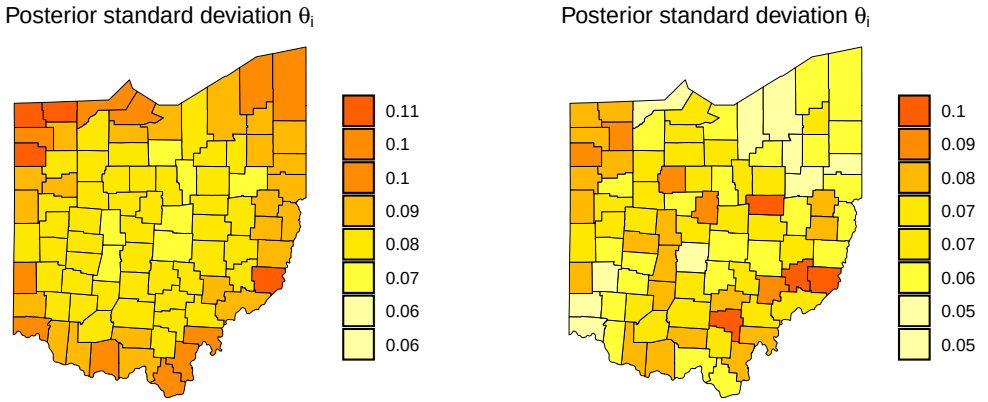
One possible explanation for the positive spatial effects in the south is the Fernald Materials Processing Center, which is located in the southwestern county of Hamilton. This plant started operation in 1951, and radioactive materials were processed there. According to Silver *and others* [2013], there are indicators that workers at the plant experienced higher than normal rates of lung cancer. There were additional worries that filters at the plant did not work properly between 1951 and 1960, spreading radioactive dust particles to nearby regions [Xia and Carlin, 1998]. Since the wind blows from west to east in Ohio, this could be one reason why the posterior mean of the spatial effects is larger in the south of Ohio than in the north.

From the heatmap of the posterior mean of the proper spatial effects in Figure 12a, there is one county in central-eastern Ohio that separates itself by having a large negative effect. This is the county of `Holmes`, which has a large Amish population. Since Amish people smoke less than the general population it could explain the negative effect in `Holmes` [Ferketich *and others*, 2008]. It could also be because `Holmes` has a relatively small population of about 22000.

The posterior standard deviation of the improper structured spatial effects and the proper spatial effects are plotted in Figure 12b. For `Improper_1_noInt` the posterior standard deviation is the largest in the outlying counties. This is consistent with it being modelled as a Besag, as the most outlying counties have fewer neighbours. The posterior standard deviation of the spatial effects of `Proper_noInt` is less structured, as expected. Overall



(a)



(b)

Figure 12: **(a)**: Posterior mean of the structured spatial effects for models (left) `Improper_1_noInt` and (right) `Proper_noInt`. **(b)**: Posterior standard deviation of the structured spatial effects for models (left) `Improper_1_noInt` and (right) `Proper_noInt`.

it seems that both the improper and the proper spatial effects have captured the same structure in space.

In Figure 13 the posterior median estimates of the lung cancer death rate per 100,000 along with 95% credible intervals for the models `Improper_1_noInt`, `Improper_1_typeIV`, and `Proper_onlyInt` are displayed for four select counties over all the years 1968 until 1988. The first county is `Jefferson` with a population of about 95000. `Jefferson` is the county in Ohio that has the highest average observed lung cancer death rate over the years 1968 until 1988. The second county `Holmes` with a population of about 22000, is the county in Ohio that has the lowest average observed rate. The third and fourth counties displayed are `Mahoning` and `Ashtabula`. They were chosen at random as more 'common' counties. `Mahoning` and `Ashtabula` have populations of about 300000

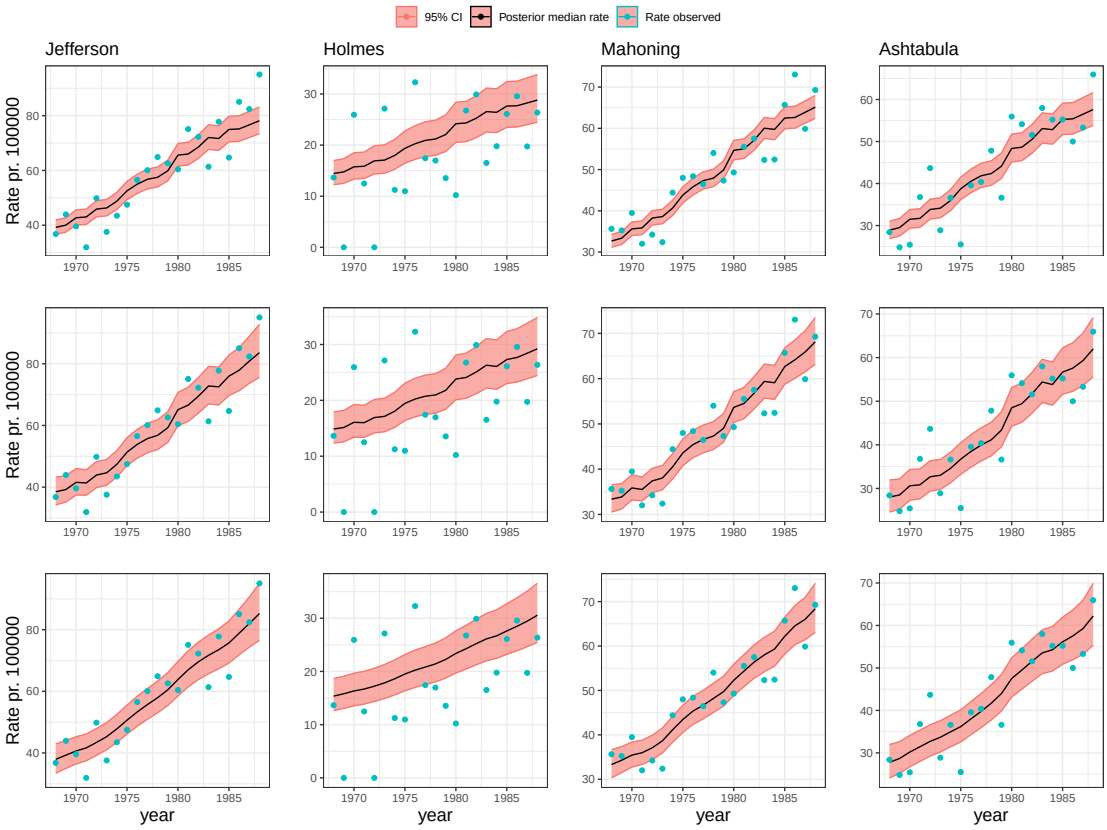


Figure 13: Posterior median of lung cancer death rate per 100,000 in four counties for each year 1968,...,1988 displayed as a black line along with 95% credible intervals shown as salmon red bands. The observed rate in the four counties is plotted as turquoise points for each year. The four counties are (left-most column) Jefferson which is the county in Ohio with the highest average observed rate in the years 1968 until 1988, (center-left column) Holmes which has the lowest average observed rate, (center-right column) Mahoning chosen at random, and (right-most column) Ashtabula chosen at random. The posterior median of rates and 95% CIs are from the models (upper) `Improper_1_noInt`, (middle) `Improper_1_typeIV`, and (lower) `Proper_onlyInt`.

and 95000, respectively. The observed rates for the four counties are also plotted for each year. The four selected counties give a better understanding of how the considered models compare by looking at how they capture the variation in the observed rate for these counties. Note that the observed rate is not the true rate, and is expected to vary greatly around the estimates of the models. All the models act similarly and seem to capture the general trend of death rates, as expected. `Improper_1_typeIV` and `Proper_onyInt` separate themselves from `Improper_1_noInt` in two ways. The first is that the posterior 95% CIs of `Improper_1_typeIV` and `Proper_onyInt` are wider than for `Improper_1_noInt`. This means that the posterior estimates for the rate vary more

for these models. This way, more of the observed rates are within their posterior 95% CIs compared to `Improper_1_noInt`. This implies that the models with interactions state that it is more likely in some of these cases that the observed rate is a reflection of the true rate, compared to the model without interactions. The second way in which `Improper_1_typeIV` and `Proper_onyInt` separate themselves from `Improper_1_noInt`, is in their posterior medians and 95% CIs for the last years. This can especially be seen for the year 1988, which for `Ashtabula`, `Jefferson`, and `Mahoning` was one of the most extreme years of observed lung cancer death rates. Possible high mortality rates are crucial for a model to capture, especially if public health decisions are made based on the estimates from the model.

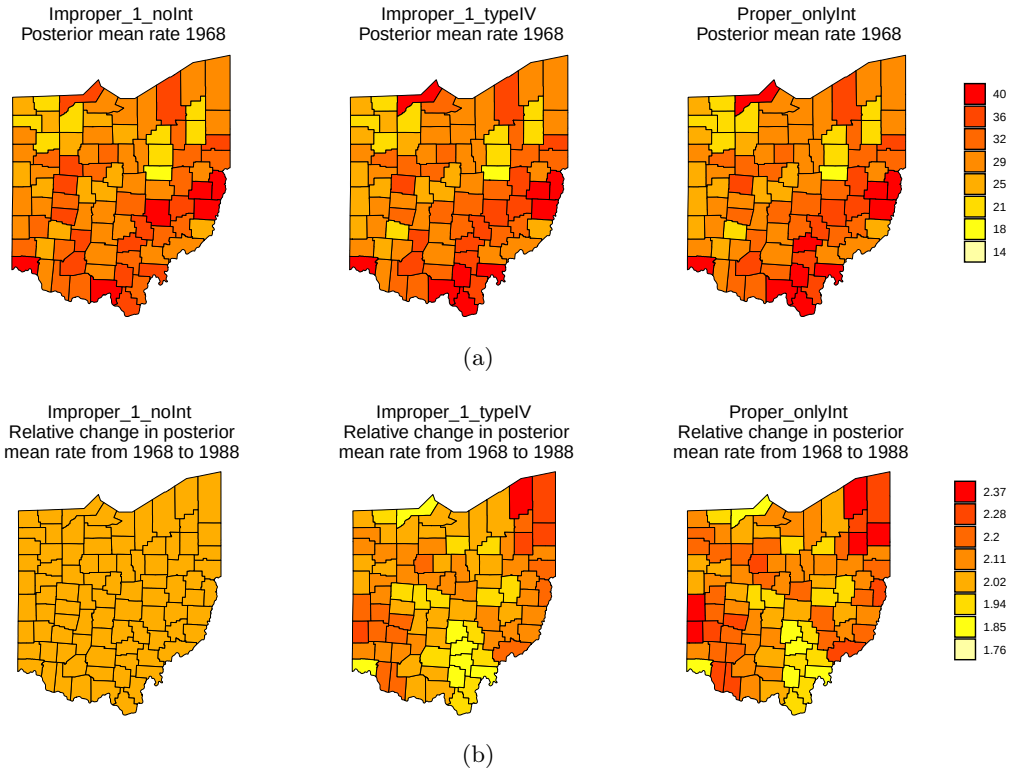


Figure 14: **(a)** The posterior mean lung cancer death rate per 100,000 in 1968 for the counties of Ohio using the models (left) `Improper_1_noInt`, (middle) `Improper_1_typeIV`, and (right) `Proper_onlyInt`. **(b)** The relative change in the posterior mean of the lung cancer death rate per 100,000 from 1968 to 1988 for the models (left) `Improper_1_noInt`, (middle) `Improper_1_typeIV`, and (right) `Proper_onlyInt`.

The county `Holmes` exhibits a relatively low average death rate, and the increase is not as high as for the other counties. None of the models assign as low a posterior median rate for `Holmes` as the observed rates may suggest. This is a benefit of smoothing. Since `Holmes` has a relatively low population, many of the most extremely low observed rates

are likely a reflection of the small population and not the true rate. All the models assign a low rate to `Holmes`, which is appropriate.

In Figure 14a the posterior mean lung cancer death rate per 100,000 in each county of Ohio for the year 1968 is shown for the models `Improper_1_noInt`, `Improper_1_typeIV`, and `Proper_onlyInt`. All the models have similar posterior mean rates, something which is expected. Figure 20a in Appendix C shows the posterior standard deviations of the estimated rates in 1968 for the same models. While the posterior mean rates are similar, the models with interactions have larger posterior standard deviations for their rates. The relative change in posterior mean rate for each county of Ohio from 1968 until 1988 is shown in Figure 14b for the models `Improper_1_noInt`, `Improper_1_typeIV`, and `Proper_onlyInt`. The relative change in posterior mean rate for `Improper_1_noInt` is only a reflection of the change in the temporal effect from 1968 until 1988. For `Improper_1_noInt` the relative change is a doubling of the posterior mean rate. The relative changes in posterior mean rates for `Improper_1_typeIV` and `Proper_onlyInt` are a reflection of both the change in temporal effects and the spatiotemporal interactions. Both of the models show similar relative changes, as expected. There are some differences, as `Proper_onlyInt` experienced greater relative changes in some of the western counties of Ohio. The relative changes in rates highlight which counties of Ohio have had the most negative development over the years. The north-eastern counties and some western counties have experienced greater increases in the posterior mean rates over the years, something which public health decisions should aim to revert. The posterior standard deviation of the rates for each county of Ohio in 1988 for the models `Improper_1_noInt`, `Improper_1_typeIV`, and `Proper_onlyInt` are shown in Figure 20b in Appendix C.

Prediction results

Before delving into the prediction results, it is necessary to mention that the priors for the precision parameters of the proper models were changed for optimization reasons. When predicting for some years, such as 1984, the proper models would not converge while having the hyperpriors as specified earlier. The parameters of the PC-prior for the precision parameters of `Proper_noInt` and `Proper_onlyInt` were changed to

```
prec = list(prior = 'pc.prec',
            param = c(1, 0.008)))
```

and for `Proper_full` they were changed to

```
prec = list(prior = 'pc.prec',
            param = c(1, 0.005)))
```

This is because R-INLA struggled to locate the mode, and a more informative prior helped solve this. By trial and error, values that led to convergence were found. This illustrates a problem with prior sensitivity for these models on the Ohio lung cancer data

Model	IS	MAE	RMSE
Improper_1_noInt	39.39	6.78	11.68
Improper_1_typeI	39.37	6.79	11.82
Improper_1_typeII	38.37	6.65	11.18
Improper_1_typeIII	39.26	6.79	11.76
Improper_1_typeIV	38.71	6.63	11.12
Improper_2_noInt	40.15	6.97	12.41
Improper_2_typeI	41.05	6.97	12.54
Improper_2_typeII	40.50	6.96	12.39
Improper_2_typeIII	40.89	7.00	12.51
Improper_2_typeIV	40.21	6.95	12.34
proper_full	38.13	6.65	11.52
proper_noInt	42.75	6.78	12.60
proper_onlyInt	37.10	6.54	10.93

Table 5: Each models prediction results from the one-year ahead prediction scheme as illustrated in Figure 10 on the Ohio lung cancer data set. Reported results are the IS, MAE, and RMSE.

set. It would have been interesting to assess how different specifications of the hyperpriors affected the results, but this is outside the scope of this project thesis. The result is that the proper models penalize complexity more than the improper models do. Especially `Proper_full` penalize complexity. This may have affected the prediction results.

In Table 5 the results of the one-year ahead predictions are reported. From Table 5 `Proper_onlyInt` performs the best with respect to IS, MAE, and RMSE, having the lowest values for each with 37.1, 6.54, and 10.93, respectively. Therefore `Proper_onlyInt` is the best of the models at predicting the one-year ahead count of lung cancer deaths in the counties of Ohio starting by predicting for 1979 until 1988. This indicates that for prediction, the propagation of the effects is well suited for the Ohio lung cancer data set. It could be that the change in the parameters of the hyperprior for `Proper_onlyInt` has influenced the results. However, when comparing the performance of `Proper_onlyInt` to the other proper models it seems like the proper spatiotemporal interaction has a noticeable effect. Therefore, the change in the hyperpriors for the proper models may not have had that large an impact on the final prediction results. The models `Improper_1_typeII` and `Proper_full` are quite close to `Proper_onlyInt` in predictive performance. For the models using a RW2, the model without interactions `Improper_2_noInt` performed the best with respect to IS. This along with the poor performance of all the models using a temporal effect following a RW2, indicates that the structure of a RW2 is poorly suited for prediction on the Ohio lung cancer data set.

In Figure 15a the observed death rate from lung cancer per 100,000 for the counties of Ohio is displayed for the years 1979, 1982, 1985, and 1988. In Figures 15b and

15c the posterior predicted mean death rates per 100,000 from models `Proper_onlyInt` and `Improper_1_typeII` are displayed for the same years. Both `Proper_onlyInt` and `Improper_1_typeII` have similar posterior predicted mean rates over the years, as expected. Where `Proper_onlyInt` seems to separate itself from `Improper_1_typeII` is by having a slightly closer fit to the observed higher rates. This can be seen in 1988 in some southern and eastern counties. The difference in predictions between `Proper_onlyInt` and `Improper_1_typeII` are minor. The standard deviation of the posterior predicted rates per 100,000 for each county for the years 1979, 1982, 1985, and 1988 are plotted in Figure 21 in Appendix C. The posterior standard deviation is highest for the southeastern counties of Ohio, and is similar for both `Proper_onlyInt` and `Improper_1_typeII`.

Figure 16 shows the posterior median predicted lung cancer death count along with the 95% CI over the years 1979,...,1989 for the counties `Jefferson`, `Holmes`, `Shelby`, and `Ashland`. The true death counts for the years 1979,...,1988 are also plotted. The predictions shown are from the models `Improper_1_noInt`, `Improper_1_typeII`, and `Proper_onlyInt`. The last year predicted on, 1989, has no actual observations in the data, so the offset was copied from the last available observation 1988. The models have similar predictions and 95% CIs. The posterior 95% CIs of all the plotted models are all capturing most of the true counts in the four counties. This indicates that the models `Improper_1_noInt`, `Improper_1_typeII`, and `Proper_onlyInt` are all well-calibrated. In `Shelby` there is one observed death count at year 1983 that falls outside all the models posterior 95% CIs. The distance from the count to the 95% CI is the shortest for `Proper_onlyInt`. For `Ashland` there is one observed death count at year 1985 that falls outside the 95% CIs of `Improper_1_noInt` and `Improper_1_typeII`, but is on the edge of the 95% CI of `Proper_onlyInt`. The counties `Shelby` and `Ashland` illustrate why `Proper_onlyInt` is the best with respect to IS. All the models seem to be well-calibrated, but `Proper_onlyInt` is the best-calibrated.

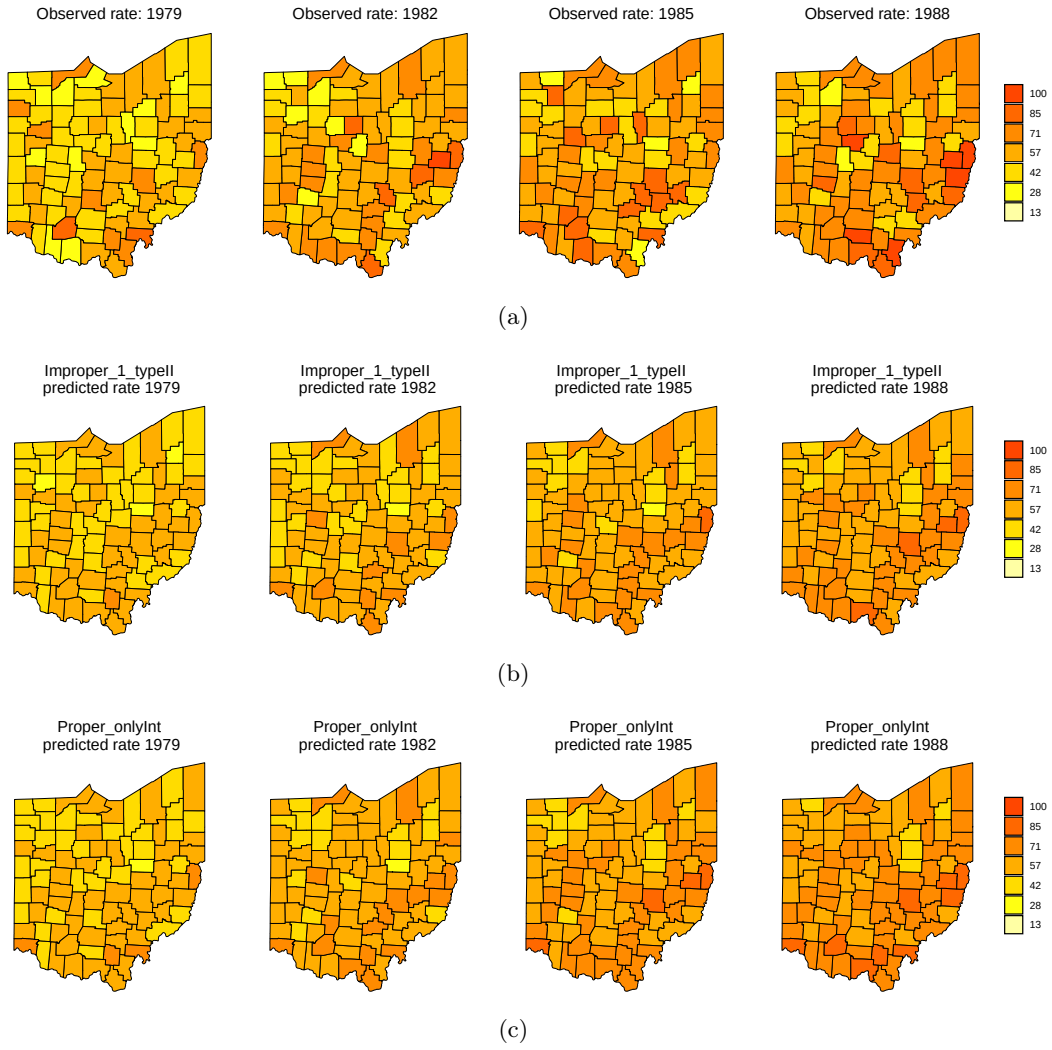


Figure 15: **(a)** Observed lung cancer death rate per 100,000 in the counties of Ohio for the years 1979, 1982, 1985, and 1988. **(b)** and **(c)** is the one-year ahead posterior mean death rate per 100,000 of the predictive distributions in each county of Ohio for the years 1979, 1982, 1985, and 1988 from models `Improper_1_typeII` and `Proper_onlyInt`, respectively.

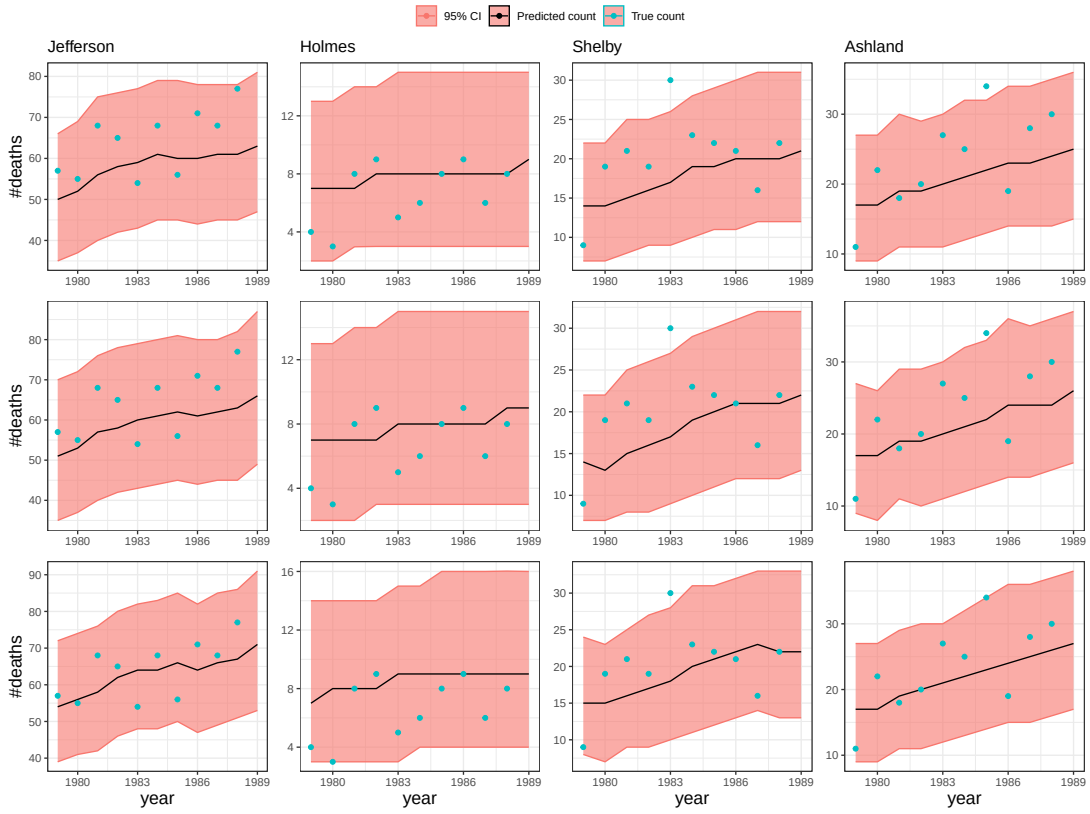


Figure 16: Posterior median predicted lung cancer death counts in four counties of Ohio for years $1979, ..., 1989$ displayed as a black line along with 95% credible intervals shown as salmon red bands. The true counts in the four counties over the years $1979, ..., 1988$ are shown as turquoise points. Note that the prediction is carried one year further than there are observations. In 1989 the population offset `pop_at_risk` is assumed the same as for 1988. The four counties are: (left-most column) Jefferson which has the highest average observed rate, (center-left column) Holmes which has the lowest average observed rate, (center-right column) Shelby a random county, and (right-most column) Ashland a random county. The predictions are from the models: (upper) `Improper_1_noInt`, (middle) `Improper_1_typeII`, and (lower) `Proper_onlyInt`.

9 Conclusions of the project and the way forward

This project thesis is the prelude to a larger work aimed at providing a conclusive answer on how to best specify spatiotemporal interactions. In this project a common way to specify the spatiotemporal interactions as introduced in Knorr-Held [2000] was compared to an alternative specification inspired by Vivar and Ferreira [2009]. The models were implemented on the Ohio lung cancer data set, and both model choice and predictive performance were compared. Of the models, `Improper_1_typeIV` is the best at estimating the rate of lung cancer deaths, and `Proper_onlyInt` is the best at predicting future lung cancer death counts for the Ohio lung cancer data set.

One of the goals of this project thesis was for me to get used to the literature in the field and to get basic experience with spatiotemporal models, before my master thesis. Having implemented 13 different models on the Ohio lung cancer data set has provided me with that experience. It has also highlighted the significance of the specification of the spatiotemporal interactions. The greatest challenges in this project have been computational. In the beginning, there was a challenge to implement the improper models with type IV interaction in R-INLA. On my personal computer, the R code crashed whenever the R-INLA call was made using the type IV interaction. On Markov, NTNUs server, it ran for several days before I decided to stop it. This was using the classic BYM model for the temporal and spatial effects. After switching to the BYM2 model, and updating the R-INLA software, the implementation of the improper models was no longer a problem. The proper models proved hard to get to converge for the one-year ahead predictions. This was due to R-INLA struggling to locate the mode of the hyperparameters. The solution became trial and error to find priors for the hyperparameters that lead to faster convergence.

The follow-up to this project thesis is my master's thesis. It will be interesting to include more alternative specifications of spatiotemporal interactions as those in Vivar and Ferreira [2009] and Utazi *and others* [2018]. With a wider range of specifications, it may help to highlight differences between them. The most important aspect in the master thesis will be a structured method to assess the performance of different specifications. The focus will be on predictive performance. In order to have a more structured comparison, it will be necessary to make simulation data. With simulation data the underlying truth is known and can be changed. Thus, the specifications can be compared in a exhaustive manner. The goal is then to provide a conclusive answer to what is the best way of specifying the spatiotemporal interactions for prediction. Another interesting avenue of further research would be to find what is the best specification for estimation, but this will not be the focus of the master thesis.

References

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): 1–20.

Coly, S., Garrido, M., Abrial, D. and Yao, A.-F. (2021). Bayesian hierarchical models for disease mapping applied to contagious pathologies, *PLOS ONE* **16**(1): 1–28.

Ferketich, A. K., Katz, M. L., Kauffman, R. M., Paskett, E. D., Lemeshow, S., Westman, J. A., Clinton, S. K., Bloomfield, C. D. and Wewers, M. E. (2008). Tobacco use among the Amish in Holmes county, Ohio, *J Rural Health* **24**(1): 84–90.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**(477): 359–378.

J. A. García, J. Martín, L. N. and Acero, F. J. (2018). A Bayesian hierarchical spatio-temporal model for extreme rainfall in Extremadura (Spain), *Hydrological Sciences Journal* **63**(6): 878–894.

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

Leroux, B. G., Lei, X. and Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence, *in* M. E. Halloran and D. Berry (eds), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Springer New York, New York, NY, pp. 179–191.

Orozco-Acosta, E., Riebler, A., Adin, A. and Ugarte, M. D. (2023). A scalable approach for short-term disease forecasting in high spatial resolution areal data, *Biometrical Journal* **65**(8): 2300096.

Potosky, A. L., Feuer, E. J. and Levin, D. L. (2001). Impact of screening on incidence and mortality of prostate cancer in the United States, *Epidemiologic Reviews* **23**(1): 181–186.

Richardson, S., Abellan, J. J. and Best, N. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK), *Statistical Methods in Medical Research* **15**(4): 385–407.

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 (1st ed.), *Monographs on Statistics and Applied Probability, Vol. 104*, Chapman & 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: Statistical Methodology* **71**(2): 319–392.

Schmid, V. and Held, L. (2004). Bayesian extrapolation of space–time trends in cancer registry data, *Biometrics* **60**(4): 1034–1042.

Schrödle, B. and Held, L. (2011a). A primer on disease mapping and ecological regression using INLA, *Computational Statistics* **26**(2): 241–258.

Schrödle, B. and Held, L. (2011b). Spatio-temporal disease mapping using INLA, *Environmetrics* **22**(6): 725–734.

Silver, S. R., Bertke, S. J., Hein, M. J., Daniels, R. D., Fleming, D. A., Anderson, J. L., Pinney, S. M., Hornung, R. W. and Tseng, C.-Y. (2013). Mortality and ionising radiation exposures among workers employed at the Fernald Feed Materials Production Center (1951–1985), *Occupational and Environmental Medicine*.

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.

Sneppen, K., Nielsen, B. F., Taylor, R. J. and Simonsen, L. (2021). Overdispersion in covid-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control, *Proceedings of the National Academy of Sciences* **118**(14): e2016623118.

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

Ugarte, M. D., Adin, A., Goicoa, T. and Militino, A. F. (2014). On fitting spatio-temporal disease mapping models using approximate Bayesian inference, *Statistical Methods in Medical Research* **23**(6): 507–530.

UN (2015). United Nations sustainable development goals. Available from: `https://www.undp.org/sustainable-development-goals?gad_source=1&gclid=CjwKCAiA-bmsBhAGEiwAoaQNmoQIXd3Wk8FRN7sJsXFA7aHsklB942NwTO9cIIbzY_GtqwohN8B8OBoCUbwQAvD_BwE`.

UNICEF (2021). United Nations inter-agency group for child mortality estimation, Subnational under-five mortality estimates, 1990–2019: Estimates developed by the United Nations inter-agency group for child mortality estimation, United Nations children's fund, New York, 2021., Estimates of child mortality.

Utazi, C. E., Afuecheta, E. O. and Nnanatu, C. C. (2018). A Bayesian latent process spatiotemporal regression model for areal count data, *Spatial and Spatio-temporal Epidemiology* **25**: 25–37.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing* **27**(5): 1413–1432.

Vivar, J. C. and Ferreira, M. A. R. (2009). Spatiotemporal models for Gaussian areal data, *Journal of Computational and Graphical Statistics* **18**(3): 658–674.

Wakefield, J. (2006). Disease mapping and spatial regression with count data, *Biostatistics* **8**(2): 158–183.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K. and Clark, S. J. (2019). Estimating under-five mortality in space and time in a developing world context, *Statistical Methods in Medical Research* **28**(9): 2614–2634.

Weir, H. K., Thompson, T. D., Soman, A., Møller, B. and Leadbetter, S. (2015). The past, present, and future of cancer incidence in the United States: 1975 through 2020, *Cancer* **121**(11): 1827–1837.

Xia, H. and Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality, *Statistics in Medicine* **17**(18): 2025–2043.

Zhang, S., Liu, X., Tang, J., Cheng, S., Qi, Y. and Wang, Y. (2018). Spatio-temporal modeling of destination choice behavior through the Bayesian hierarchical approach, *Physica A: Statistical Mechanics and its Applications* **512**: 537–551.

Zhang, T. (n.d.). *MATH 470 Independent Study in Matrix Theory: The Kronecker Product, by Thomas Zhang: https://thomaszh3.github.io/writeups/kronecker.pdf.*

A Additional theory

A.1 Kronecker product

The Kronecker product is used in Knorr-Held [2000] to construct the structure matrix for the interaction. The definition of the Kronecker product is: for two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ which are $n \times m$ and $p \times q$ dimensional respectively, the Kronecker product is defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & a_{13}\boldsymbol{B} & \dots & a_{1m}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & a_{23}\boldsymbol{B} & \dots & a_{2m}\boldsymbol{B} \\ \vdots & & \ddots & & \vdots \\ \\ a_{n1}\boldsymbol{B} & \dots & & \dots & a_{nm}\boldsymbol{B} \end{bmatrix}$$

where a_{ij} is the element in row i, column j of $\boldsymbol{A}$. It is therefore a block matrix, which can be looked at as a $np \times mq$ dimensional matrix.

The Kronecker product is not commutative $\boldsymbol{A} \otimes \boldsymbol{B} \neq \boldsymbol{A} \otimes \boldsymbol{B}$ in general. However, the Kronecker product is permutation equivalent [Zhang, n.d.]. This means that there exists two permutation matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ such that

$$\boldsymbol{A} \otimes \boldsymbol{B} = \boldsymbol{P}(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{Q}$$

Hence, by doing a permutation on the Kronecker product, they are commutative.

Another important property of matrices and the Kronecker product is that for two matrices, the rank of the resulting Kronecker product equals the product of the ranks of the individual matrices

$$rank(\boldsymbol{A} \otimes \boldsymbol{B}) = rank(\boldsymbol{A})rank(\boldsymbol{B})$$

This will be used to find the rank of the precision matrix for the spatiotemporal interactions specified as in [Knorr-Held, 2000].

A.2 Proper scoring rules

We aim to use proper scoring rules as outlined in Gneiting and Raftery [2007]. This will be an informal definition of proper scoring rules and will not concern itself with measurability. Given a predictive distribution P and an event x, the score according to a scoring rule, $S(\cdot, \cdot)$, is $S(P, x)$. Assume that Q is a probability distribution, then the expected score of P under Q is defined as

$$S(P, Q) = \int S(P, \omega) dQ(\omega)$$

Given this definition, S is a proper scoring rule if

$$S(Q, Q) \geq S(P, Q), \quad \forall P, Q \in \mathcal{P}$$

and S is strictly proper if equality only holds when $P = Q$.
This definition simply states that a proper scoring rule should have a maximum value when the predictive distribution equals the true distribution. Strictly proper means that this maximum is unique.

According to Gneiting and Raftery [2007], a scoring rule is effective if

$$S(P_1, Q) \geq S(P_2, Q) \iff d(P_1, Q) \leq d(P_2, Q)$$

where $d()$ is a metric, measuring the 'distance' between a predictive distribution and the true probability distribution. An effective scoring rule will therefore give a higher score to a predictive distribution closer to the true distribution. Hence, using a proper and effective scoring rule will make it so that the predictive distribution closer to the true distribution will be chosen, and that no distribution can have a higher score than the true one.
In practice, the expected score is approximated using the average score as

$$S_n = \frac{1}{n} \sum_{i=1}^{n} S(P_i, x_i) \approx S(P, Q)$$

B Theoretical background on INLA

This appendix provides a theoretical background on INLA. For a thorough explanation see Rue *and others* [2009]. The section begins by detailing latent Gaussian models, and then a Gaussian approximation as done in INLA is presented. Afterwards, how INLA locates evaluation points for the hyperparameters is shown, and lastly the approximation of the latent field is presented.

B.1 Latent Gaussian models

One of the reasons for the efficiency of INLA is that it works on a specific subclass of models, namely Latent Gaussian Models (LGM).
For a LGM, the response, y_i, is assumed to belong to an exponential family with mean, μ_i, linked to a structured, additive predictor η_i through a link function

$$g(\mu_i) = \eta_i$$

The predictor is defined to have the following form

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i$$

where $\{f^{(j)}(\cdot)\}$ unknown functions of covariates $\boldsymbol{u}$, $\{\beta_k\}$ linear effects of covariates $\boldsymbol{z}$ and ϵ_i unstructured term.

A LGM is a model where the priors for α, $\{f^{(j)}(\cdot)\}$, $\{\beta_k\}$, and ϵ_i, are all Gaussian. Hence, the latent field, $\boldsymbol{x} = (\alpha, f^{(1)}(\cdot), ..., f^{(n_f)}(\cdot), \beta_1, ..., \beta_{n_\beta}, \epsilon_1, ..., \epsilon_n)$, has a Gaussian prior, and therefore the name. Additionally, there are hyperparameters, $\boldsymbol{\theta}$. We make no assumptions about the distribution of the hyperparameters, but we do assume that there are few of them.
In our case, one can ignore the $\{\beta_k\}$. If one assumes that $y_i \perp y_j|\boldsymbol{x},\boldsymbol{\theta}$ and that $y_i \perp \boldsymbol{x}_{-i}|x_i,\boldsymbol{\theta}$ one can write the posterior distribution given observations $\boldsymbol{y}$ as

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i\in I} \pi(y_i|x_i,\boldsymbol{\theta})$$

Making no assumptions about the prior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, and using that $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ is assumed Gaussian a priori with zero mean and precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ the posterior becomes

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i\in I} \log(\pi(y_i|x_i,\boldsymbol{\theta}))\right)$$

In our case, the latent field $\boldsymbol{x}$ is assumed to be GMRF or IGMRF, and hence the precision matrix is sparse, something which enables sparse computations. This is an additional benefit of INLA.

B.2 Gaussian approximation

One important part of the approach is Gaussian approximation of densities such as

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in I} g_i(x_i)\right)$$

First, a second-order expansion of $g_i(x_i)$ is calculated around the mode μ_i. The mode is approximated iteratively along with the second-order expansion of $g_i(x_i)$ as

$$g_i(x_i) \approx g_i(\mu_i^{(0)}) + b_i x_i - \frac{1}{2}c_i x_i^2$$

where $\mu_i^{(0)}$ is initial guess of the mode. Then, the Gaussian approximation $\tilde{\pi}_G(\boldsymbol{x})$ of $\pi(\boldsymbol{x})$ becomes a zero mean Gaussian with precision matrix $\boldsymbol{Q} + \text{diag}(\boldsymbol{c})$. Note that if the precision matrix $\boldsymbol{Q}$ is sparse, as it is for GMRFs, then the precision matrix $\boldsymbol{Q} + \text{diag}(\boldsymbol{c})$ is also sparse.

B.3 Locating evaluation points for hyperparameters

The first step of INLA is to compute an approximation to the posterior marginal of the hyperparameters, $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$. It is not necessary to find a parametric expression for this approximation. Rather, one wants to find evaluation points, $\boldsymbol{\theta}_k,\ k \in \{1, ..., n\}$ which

capture the bulk of the probability mass, and corresponding weights Δ_k for each point. As stated in Rue *and others* [2009] this is done by evaluating

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})}$$

where $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ is the Gaussian approximation of $\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$, and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode of the full conditional of $\boldsymbol{x}$ given a specific $\boldsymbol{\theta}$. Exactly how the evaluation points $\boldsymbol{\theta}_k$ and weights Δ_k are calculated is not being discussed here, see Rue *and others* [2009] for details. These evaluation points along with their corresponding weights will be used in the numerical integration to find the posterior marginals for x_i.

B.4 Approximating the latent field

The last part is to find the posterior marginals of the latent field

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

which is approximated using the numeric integration

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}_k,\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})\Delta_k$$

From earlier, we have all ready found the approximation $\tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})$. What remains is to find an approximation to $\tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y})$. There are three different approaches, Gaussian approximation, Laplace approximation, and simplified Laplace approximation. We will only present the Laplace approximation, for the other approximations and more details see Rue *and others* [2009]. The Laplace approximation is

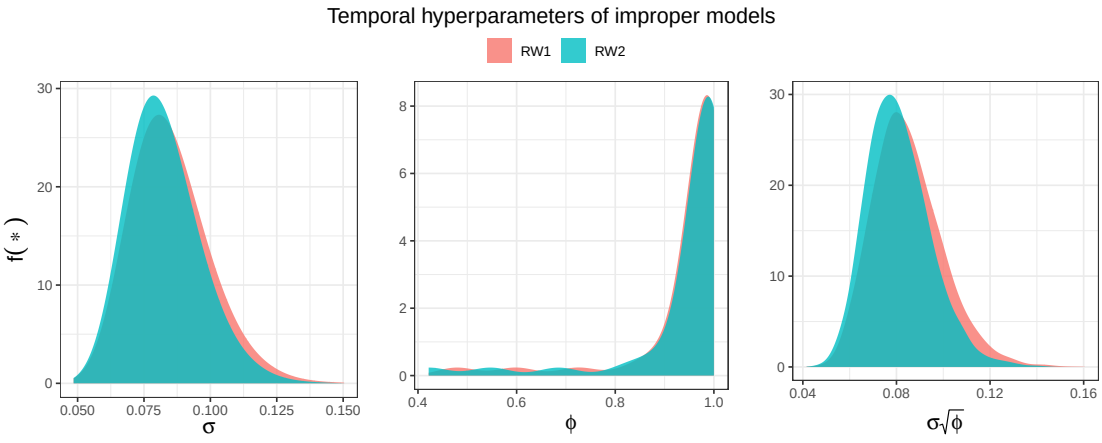
$$\tilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta},\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})}{\tilde{\pi}_{\mathrm{GG}}(\boldsymbol{x}_{-i}|x_i,\boldsymbol{\theta},\boldsymbol{y})}\Big|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^*_{-i}(x_i,\boldsymbol{\theta})}$$

where $\tilde{\pi}_{\mathrm{GG}}$ is the Gaussian approximation to $\boldsymbol{x}_{-i}|x_i,\boldsymbol{\theta}$, and $\boldsymbol{x}^*_{-i}(x_i,\boldsymbol{\theta})$ is the modal configuration. However, this is extremely expensive to compute as $\tilde{\pi}_{GG}$ has to be recomputed for each value of x_i and $\boldsymbol{\theta}$. Therefore Rue *and others* [2009] proposes modifications for optimization reasons. Firstly, approximate the modal configuration $\boldsymbol{x}^*_{-i}(x_i,\boldsymbol{\theta})$ using the Gaussian approximation of $\boldsymbol{x}$ as $\boldsymbol{x}^*_{-i}(x_i,\boldsymbol{\theta}) \approx \mathbb{E}_{\tilde{\pi}_G}[\boldsymbol{x}_{-i}|x_i]$. Secondly, by using the Gaussian approximation again, we get that the Laplace approximation can be represented as

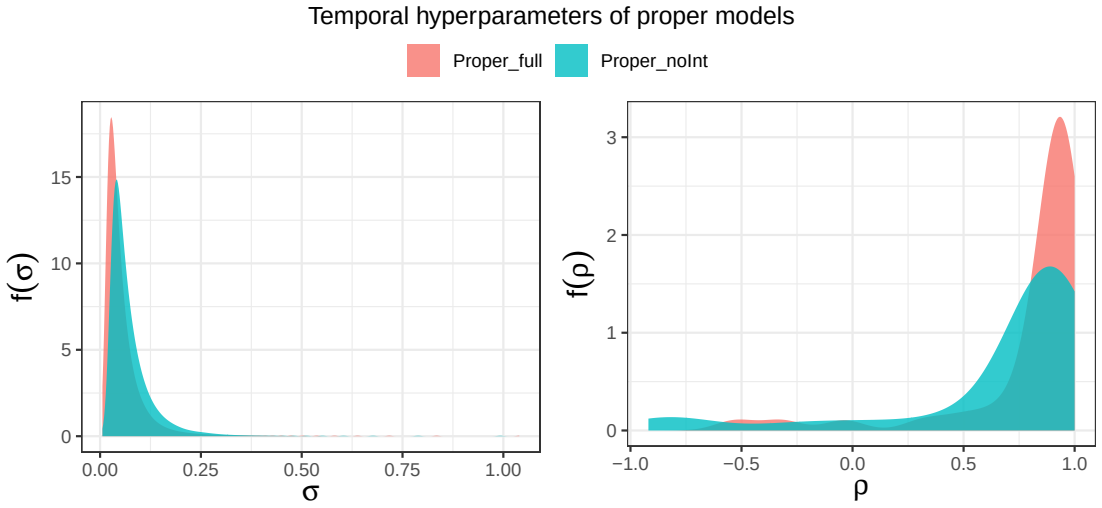
$$\tilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta},\boldsymbol{y}) \propto \mathcal{N}(x_i;\mu(\boldsymbol{\theta}),\sigma_i^2(\boldsymbol{\theta}))\exp(\text{cubic spline}(x_i))$$

The cubic spline of x_i is fitted to the difference of the log-densities of $\tilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ and $\tilde{\pi}_G(x_i|\boldsymbol{\theta},\boldsymbol{y})$. Quadrature integration is used to normalize the density.

C Supplementary results

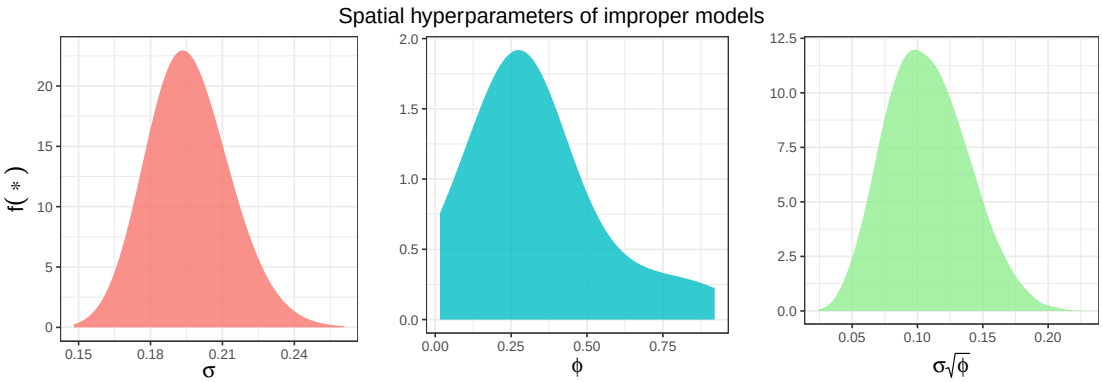


(a)

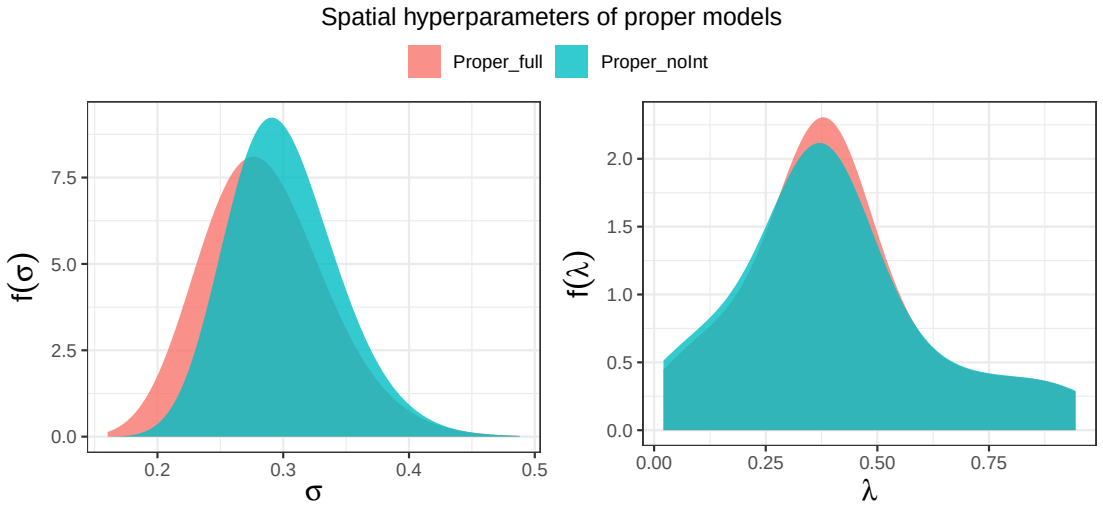


(b)

Figure 17: **(a)**: Posterior densities from `Improper_1_noInt` and `Improper_2_noInt` of (left) standard deviation of temporal effects σ, (middle) temporal mixing parameter ϕ, and (right) standard deviation of structured temporal random effect. Estimation of density achieved by sampling 10000 instances of the posterior of precision τ and mixing parameter ϕ for temporal effects. Then for each sample compute $\sqrt{\frac{\phi}{\tau}}$, which for all the estimates gives us an estimate of the density of the standard deviation of the structured temporal effect. Improper models with structured temporal effect $\boldsymbol{\alpha}$ following a RW1 colored salmon red and RW2 colored turquoise. **(b)**: Posterior densities from `Proper_noInt` colored turquoise and `Proper_full` colored salmon red of (left) standard deviation of temporal effect σ and (right) temporal correlation parameter ρ, for proper models with structured temporal effect $\boldsymbol{\alpha}$ following an AR1.

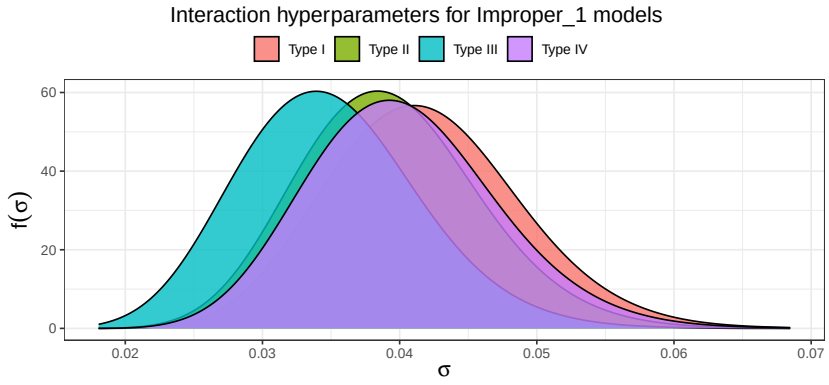


(a)

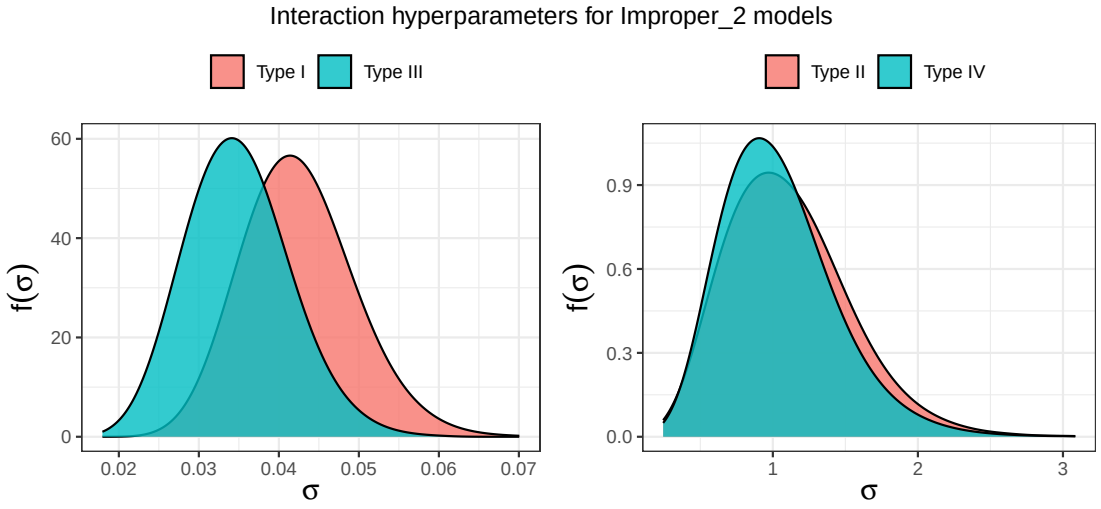


(b)

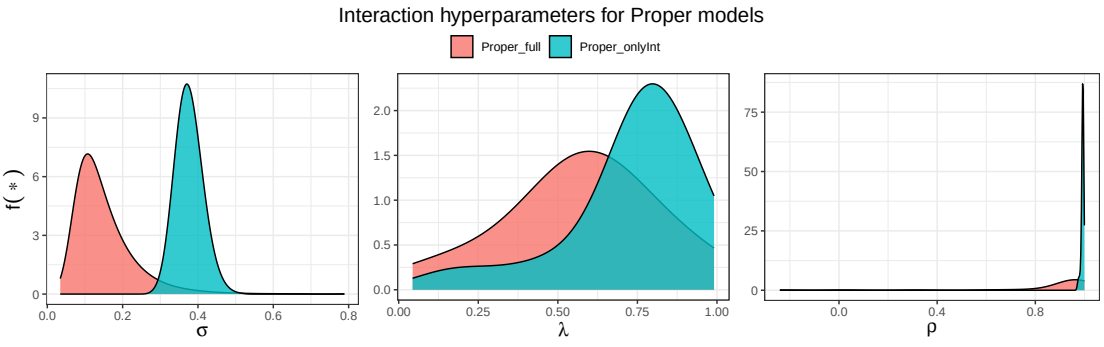
Figure 18: **(a)**: Posterior densities from `Improper_1_noInt` of (left) standard deviation of spatial effects, (middle) spatial mixing parameter ϕ, and (right) standard deviation of structured spatial random effect. **(b)**: Posterior densities from `Proper_noInt` colored salmon red and `Proper_full` colored turquoise of (left) standard deviation of spatial effect σ and (right) spatial mixing parameter λ, for a proper model with structured spatial effect $\boldsymbol{\theta}$ following a Leroux model.



(a)

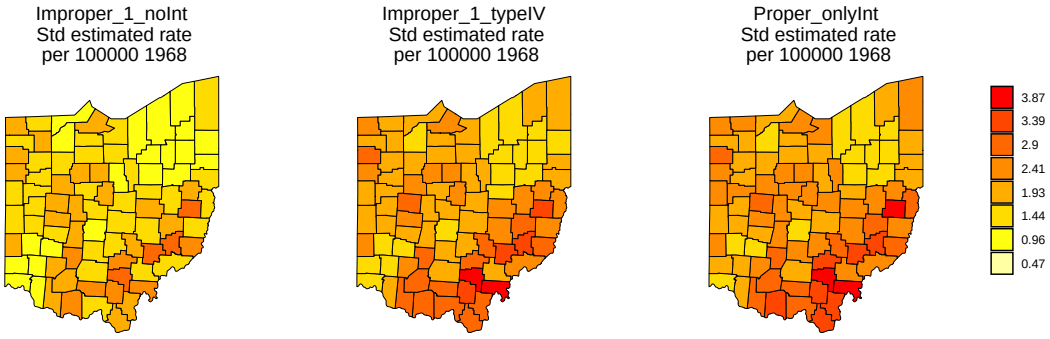


(b)

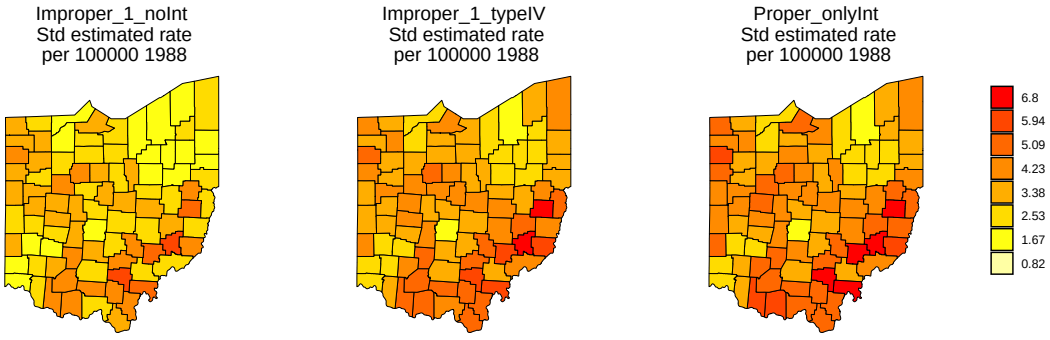


(c)

Figure 19: **(a)**: Posterior distribution of standard deviation of interaction terms of type I-IV. Here structured temporal random effect $\boldsymbol{\alpha}$ follows a RW1. **(b)**: Posterior distribution of standard deviation of interaction terms of (left) type I and III, and (right) type II and IV. Here structured temporal random effect $\boldsymbol{\alpha}$ follows a RW2. **(c)**: Posterior densities of proper models' interaction hyperparameters from `Proper_onlyInt` colored turquoise and `Proper_full` colored salmon red (left) interaction standard deviation σ, (middle) interaction spatial correlation parameter λ, and (right) interaction temporal correlation parameter ρ.

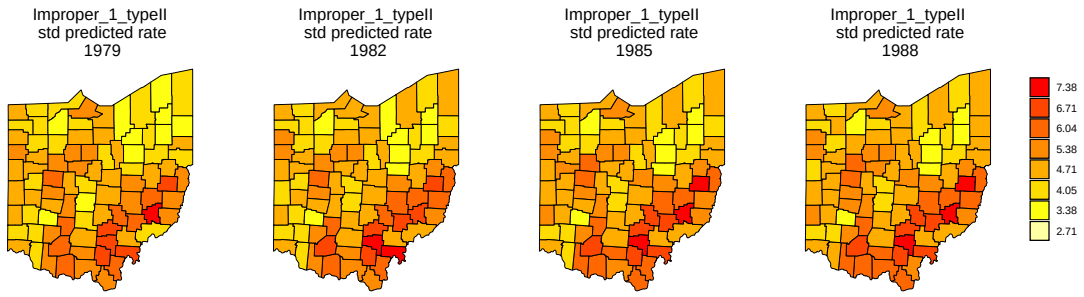


(a)

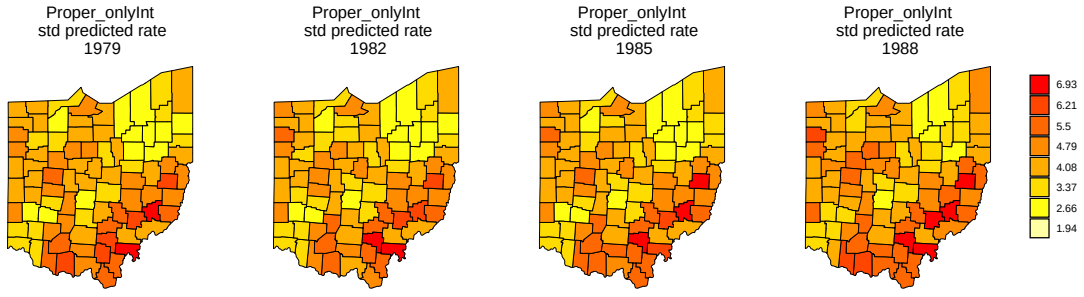


(b)

Figure 20: **(a)** Posterior standard deviation of death rate from lung cancer per 100,000 in the counties of Ohio in 1968, using the models `Improper_1_noInt`, `Improper_1_typeIV`, and `Proper_onlyInt`. **(b)** Posterior standard deviation of death rate from lung cancer per 100,000 in the counties of Ohio in 1988, using the models `Improper_1_noInt`, `Improper_1_typeIV`, and `Proper_onlyInt`.



(a)



(b)

Figure 21: **(a)** Posterior standard deviation of predicted death rate from lung cancer per 100,000 in Ohio for the years 1979, 1982, 1985, and 1988, using the model `Improper_1_typeII`. **(b)** Posterior standard deviation of predicted death rate from lung cancer per 100,000 in Ohio for the years 1979, 1982, 1985, and 1988, using the model `Proper_onlyInt`