Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Halvard Emil Sand-Larsen

Using adaptive spatial weight matrices for disease mapping

A validation study of different neighbourhood structures

Master's thesis in Applied physics and mathematics
Supervisor: Andrea Riebler
July 2025

NTNU
Norwegian University of
Science and Technology

Halvard Emil Sand-Larsen

Using adaptive spatial weight matrices for disease mapping

A validation study of different neighbourhood structures

Master's thesis in Applied physics and mathematics
Supervisor: Andrea Riebler
July 2025

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Preface

This Master's thesis rounds off my five years at the Norwegian University of Science and Technology (NTNU), studying Applied physics and mathematics. The work was undertaken in the spring semester of 2025 in the course TMA4900.

I would like to thank my supervisor professor Andrea Riebler for valuable guidance during the whole process, and she was always able to help me when I got stuck. Additionally, I would like to thank professor Miguel Angel Martinez-Beneito for providing the data, helping me with WinBUGS and for many fruitful discussions.

Finally, I would like to thank my family for their support and my friends for creating a motivating working environment, as well as some needed breaks.

Abstract

The intrinsic conditional autoregressive (ICAR) is a popular model component in disease mapping models for the discrete spatial setting. It assumes spatial stationarity, which results in a global level of spatial smoothing from a global precision parameter and a binary neighbourhood structure. Although the stationarity assumption has proven to be useful, it may not be suitable to all situations. For instance, if the area of interest is large and diverse, it seems overly simple to impose one global level of spatial smoothing. As such, more flexible models have been proposed in recent years with varying degrees of flexibility for the spatial smoothing. This thesis will focus on finding more flexible neighbourhood structures, but still with one precision parameter.

The different neighbourhood structure were compared in a validation study for male mortality data from 2019 for Spain at the province level. The ICAR is an intrinsic Gaussian Markov random field (GMRF) and its structure matrix is determined by the first-order neighbours of each region. Three adaptive GMRFs were used to investigate alternative neighbourhood structures. Two of the adaptive GMRFs were trained on multiple diseases to find the neighbourhood structure because they have to many parameters to train on a single disease. Posterior mean smoothing parameter estimates were used to build the structure matrices after the training, and the resulting structure matrix was then used in an intrinsic GMRF individually for each diseases in the validation. The different structures were then compared in terms of model fit, with respect to the deviance information criteria (DIC), Watanabe-Akaike information criteria (WAIC) and logarithmic scoring (LS).

The adaptive neighbourhood structures outperformed the ICAR, especially when increasing the number of diseases used in the training of the multivariate neighbourhood structures. The ICAR was still competitive, but this indicates that the assumption of the ICAR is too simplistic in the univariate disease setting for the provinces of Spain.

The effect of the diseases used for training the multivariate neighbourhood structures was also assessed in the validation study. Of the total 86 diseases, random subsets with a size $n \in \{10, 20, 50, 86\}$ were chosen along with the subset of cancers in the data, of which there are 26. When validating on all 86 diseases, it was better to include more diseases in the training, which is as expected. However, when validating on the cancers, the most flexible model trained on the cancer data performed the best, by a slight margin, and including the remaining diseases in the training did not improve the model fit in the validation in terms of DIC, WAIC and LS.

Sammendrag

ICAR (Intrinsic conditional autoregressive) er en mye brukt modellkomponent innen problemstillinger rundt sykdomskartlegging i diskret rom. En sentral forutsetning for en ICAR er romlig stasjonaritet, som gir en global styrke på den romlige utjevningen, altså hvor mye eller lite en region kan avvike fra naboene sine. Den er tett knyttet til en binær første ordens nabostruktur. Hvis sykdomskartleggingen foregår over et stort og heterogent geografisk område kan imidlertid forutsetningen om romlig stasjonaritet være upassende. Modellkomponenter som er mer fleksible enn en ICAR har blitt foreslått i litteraturen. Dette er metoder som kan finne mer fleksible nabostrukturer, gjerne gjennom å øke antall parametere. I denne masteroppgaven sammenlignes de ulike nabostrukturene på individuelle sykdommer for spanske provinser i 2019.

For å teste hvilke forutsetninger som er rimelige, har jeg benyttet tre mer fleksible modellkomponenter, kalt AGMRFs (adaptive Gaussian Markov random fields), og sammenlignet disse med en ICAR i en valideringsstudie. Dataene i studien er mannlige dødsfall i 2019 per provins i Spania fordelt på 86 dødsårsaker. De to mest fleksible modellkomponentene har mange parametere og trenger derfor å trene på flere sykdommer samtidig. Det viktige resultatet fra treningen er den trente strukturen, som skal sammenlignes med nabostrukturen for en ICAR.

Alle strukturene er deretter brukt i valideringsstudien for en og en sykdom av gangen. Formålet med denne sammenligningen er å undersøke om strukturen for en ICAR er optimal for en enkelt sykdom, eller om det er fordelaktig å bruke mer fleksible strukturer, hvor noen er trent på flere sykdommer. I valideringsstudien blir de ulike strukturene sammenlignet basert på DIC (deviance information criteria), WAIC (Watanabe-Akaike information criteria) og LS (logarithmic scoring).

Resultatene viser at de fleksible strukturene fungerer bedre enn den binære ICAR-strukturen. Dette var spesielt tydelig når antall sykdommer som ble brukt for å trene strukturene var høyere. Riktignok var ICAR-strukturen ikke langt unna de mer fleksible strukturene basert på de tre kriteriene. Sett under ett viser resultatene at strukturen fra en ICAR er for enkel for sykdomskartlegging for provinsene i Spania.

Relevansen av antall, og hvilke, sykdommer som ble brukt for å trene strukturene er også undersøkt i valideringsstudien. Av totalt 86 sykdommer ble tilfeldige grupper på størrelse $n \in \{10, 20, 50, 86\}$ trukket ut. I tillegg ble en gruppe med ulike typer kreft brukt, som det var 26 av. Resultatene viser at for alle sykdommene, når de blir validert for en og en, gjorde strukturene med flest sykdommer i treningsdataene det best. Når kun resultatet for de ulike kreftformene var relevant var det ikke grunn til å inkludere de resterende sykdommene i treningen, og strukturene basert på flere sykdommer gjorde det ikke bedre, tvert imot nesten verre.

Table of Contents

1 Introduction **1**

2 Bayesian hierarchical models **5**

2.1 The likelihood layer 5

2.2 The latent layer 6

2.3 The hyperprior layer 7

2.4 Illustration of priors for a linear regression example 8

2.5 Gaussian Markov random fields 10

2.6 Inference 16

2.6.1 Markov chain Monte Carlo 16

2.6.2 Integrated nested Laplace approximations 18

3 Adaptive Gaussian Markov random fields **21**

3.1 Motivation and literature review 21

3.2 Spatial AGMRFs for the validation study 24

3.2.1 Border Weighted ICAR 24

3.2.2 Region Weighted ICAR 25

3.2.3 Edge Weighted ICAR 27

4 Disease mapping **31**

4.1 Multivariate disease mapping 31

4.2 Male mortality data for multiple causes from Spain 2019 32

5 Validation study **35**

5.1 Training the models 35

5.1.1 Model setup 35

5.1.2 Design 37

5.1.3 Estimating the neighbourhood structure 38

5.2 Validating the models 40

5.3 Model choice criteria 41

5.4 Implementation 42

6 Results **53**

6.1 Comparison of different neighbourhood structures 53

6.2 Validation on real data 59

7 Discussion **69**

References **73**

Appendix **79**

A Additional GMRFs 79

A.1 Additional spatial examples 79

A.2 Temporal examples 81

A.3 Spatio-temporal examples 84

B Temporal adaptive GMRFs 89

C Sensitivity analysis for BW-ICAR 91

D Illustrative examples with INLA 93

E Additional plots from the validation study 96

E.1 Figures and tables from the training 96

E.2 Figures and tables from the validation 97

1 Introduction

Disease mapping is the modelling of a disease in space. This is an important application in the field of epidemiology (Lawson, 2013), and some recent applications in disease mapping includes modelling tuberculosis (Chen *and others*, 2023), cancers (Simkin *and others*, 2022) and COVID-19 (for example Kim *and others* (2023), Natalia *and others* (2025), Sahu and Böhning (2022)). The modelling is often concerned with the underlying risk structure and the aim is to increase the understanding of the given disease, and it can also serve as a starting point for further analysis or generate predictions for the future. Disease mapping models are typically in the Bayesian setting and random effects are often included to account for spatial dependencies that are not captured by the covariates. The emphasis is placed on spatial dependencies as it is well documented that many diseases have a large spatial component (Wakefield, 2006). This can be accomplished through spatial smoothing, where the assumption is that points in space have more in common with other points nearby in space than points far away.

A common random effect for this purpose, with almost 10000 citations, is the Intrinsic conditional autoregressive (ICAR) (Besag, 1974). The ICAR belongs to a group of models called Gaussian Markov random fields (GMRF). In a discrete spatial setting, this model is commonly associated with a first-order neighbourhood structure. This often means that areas or regions that share some part of their borders are considered neighbours. For other propositions of neighbourhood structures see for example Duncan *and others* (2017). The ICAR then enacts a spatial smoothing by preferring to keep the value for a region close to the mean of the surrounding neighbours, and a higher number of neighbours for a given region makes it harder to deviate from the assigned mean. An additional assumption for the ICAR, which will be put under scrutiny in this thesis, is the assumption of spatial stationarity, i.e. a global level of smoothing. In this thesis the unstructured random effects which are often coupled with the ICAR to handle spatial heterogeneity, see for example the Leroux (Leroux *and others*, 2000) and the BYM2 Riebler *and others* (2016) in Appendix A.1, were removed. This choice was made as the main interests are exploring the neighbourhood structure and the assumptions for the structured spatial models.

Specifically, the assumption of spatial stationarity will be put to the test for an application in a validation study. This application concerns the modelling of multiple causes of mortality for males in Spain for the year 2019. The spatial aspect arises as the data is aggregated to the province level, which gives an area of interest as shown in Figure 1.1. Note that the terms mortality causes and diseases will be used interchangeably in this thesis.

Spain has been chosen as it is a large and diverse country, for example in terms of climate, culture, economy and even languages, and the different autonomous regions have different legislation (Grant, 2019). Thus, the assumption of spatial stationarity might be too restrictive. One idea in the literature is to relax the stationarity assumption, which gives rise to adaptive Gaussian Markov random fields (AGMRF) (see for example Brezger *and others* (2007), Lu *and others* (2007), Abdul-Fattah *and*

others (2024)). Similarly to the ICAR, these can be employed in a Bayesian setting. It should be noted that AGMRFs are present in both temporal and spatio-temporal modelling as well, see for example Susmann and Alkema (2024) and Knorr-Held (2000), but this is outside the scope of this thesis. To investigate whether the stationarity assumption of the ICAR is too restrictive, the model performance for Bayesian models using different neighbourhood structures will be measured against each other in a validation study. The other neighbourhood structures will be found from training AGMRFs.

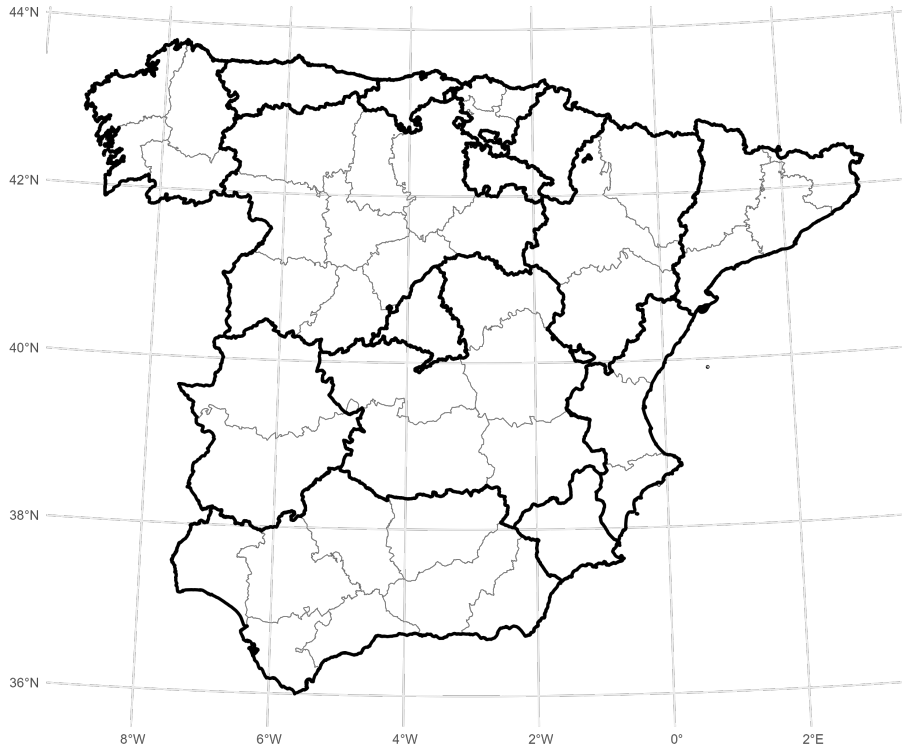


Figure 1.1: A map of mainland Spain. Grey borders divide provinces while the black borders divide autonomous regions.

The AGMRFs of interest in this thesis have been restricted to strictly spatial models with a first-order neighbourhood structure. This has been done to make them comparable to the ICAR. For example, all of the adaptive models proposed in Corpas-Burgos and Martinez-Beneito (2020) are ineligible out of the box as they include some form of an unstructured effect. The three chosen adaptive models for this thesis will vary in their degree of flexibility. The model closest to the ICAR, the Border Weighted ICAR, has two levels of spatial smoothing and will differentiate between neighbours in the same or in different autonomous regions (Aleshin-Guendel and Wakefield, 2024). Thus, in the validation it will use one precision parameter, but the structure matrix will be based on the two precision parameters fitted in the training. The next step up in flexibility is the Region Weighted ICAR, which uses

a smoothing parameter for each province (Corpas-Burgos and Martinez-Beneito, 2020). The final model is the Edge Weighted ICAR which has a unique level of smoothing for each pair of neighbours (Riddervold, 2024). Thus, four different assumptions and models for spatial smoothing will be compared. Specifically, the resulting neighbourhood structure from each model will be compared in a univariate disease setting. Note that some of these models were proposed in combination with some unstructured effect which will be removed when they are used in this thesis.

Additionally, the presence of multiple diseases in the dataset also opens the door to multivariate disease mapping [Chapter 10](Lawson, 2013). For the Region Weighted ICAR and the Edge Weighted ICAR this is especially important as they are both highly parametrized, and would have more parameters than datapoints in the univariate disease setting. Multivariate disease mapping means that multiple diseases are modelled simultaneously, with the ability to share some aspect of the Bayesian model across the diseases. Additionally, multivariate disease mapping introduces another point of interest. Namely whether the inclusion of additional diseases improves the model fit for a single disease or not. The proposition is that most diseases will have a similar spatial risk pattern. Thus, sharing of information across diseases could improve the fitted risk mapping. On the other hand, the diseases could follow differing spatial risk patterns and combining them might worsen the model fit. Investigating this effect is a secondary goal of this thesis. This comparison will be done by training the multivariate models on different subsets of diseases and using the fitted neighbourhood structures in a univariate disease validation study. The results are compared by DIC (Spiegelhalter *and others*, 2002), WAIC (Gelman *and others*, 2014) and LS (Gneiting and Raftery, 2007) both to each other and to the neighbourhood structures from the univariate models. Assessing the model fit for multivariate disease mapping will not be explored in this thesis, it will only be used to generate univariate structure matrices.

This thesis will implement and compare the ICAR with three relatively new AGMRFs for a relevant disease mapping application, spoecifically the resulting neighbourhood stuctures. These methods have not previously been compared, and it is of interest to observe how the different degrees of flexibility affects the modelling. Secondly, the fields of disease mapping and AGMRFs can be directly tied to the United Nations sustainability goals (United Nations, 2015). The most relevant being goal three, namely "Good health and well-being". Disease mapping is directly tied to health, and better disease mapping models can give further insight into diseases as well as serve as a stepping stone for further analysis. For instance, if a disease mapping model identifies that specific regions are more prevalent and have worse health outcomes compared to the remaining area of interest, this warrants further analysis into the underlying reasons. It is even possible the specific analysis for male mortality data from Spain could give some new insight. AGMRFs also play a part as they are an integral part of disease mapping models which have received much attention in recent years (MacNab, 2023). If AGMRFs are shown to outperform the traditional ICAR it could represent a paradigm shift in the whole field.

If there have been some foreign concepts so far they will all be further explained in

the rest of the thesis. Specifically, Section 2 goes in depth on the setup for Bayesian modelling, and further expands on GMRFs, which will be used in the Bayesian models. Adaptive GMRFs are introduced in Section 3 before disease mapping and the application from Spain are expanded upon in Section 4. Section 5 introduces the methodology for the validation study, which is the most important part of this thesis alongside the results presented in Section 6. Then the thesis is rounded off with a discussion in Section 7.

2 Bayesian hierarchical models

The overarching framework for this thesis is the Bayesian hierarchical model. It typically consists of three layers and the idea is that the upper levels depend on the lower levels in a hierarchical structure. Conceptually, this structure can be viewed as

1. Define the likelihood
2. Define the link function and the latent layer with necessary priors (2.1)
3. Define the necessary hyperpriors for layer 1 and layer 2.

This layered structure can also be defined mathematically as in Equation (2.2). Then $\pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta})$ is the likelihood for the observed data $\mathbf{y}$, $\pi(\boldsymbol{\eta}|\boldsymbol{\theta})$ is the density for the latent parameter $\boldsymbol{\eta}$ and $\pi(\boldsymbol{\theta})$ is the prior and hyperpriors for all the parameters of the model.

$$\begin{aligned} &1.\ \pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta}) \\ &2.\ \pi(\boldsymbol{\eta}|\boldsymbol{\theta}) \\ &3.\ \pi(\boldsymbol{\theta}). \end{aligned} \tag{2.2}$$

The posterior distribution of interest can then be defined as

$$\pi(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta})\pi(\boldsymbol{\eta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta})\pi(\boldsymbol{\eta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.3}$$

As the denominator rarely has an analytic expression multiple methods of inference have been developed, the most prominent being MCMC and INLA, which will be introduced at the end of this Section. First, a more thorough introduction to each of the three layers in a Bayesian hierarchical model, followed by Gaussian Markov random fields, an important piece of the latent layer. This theory will also be essential in regard to the validation study in Section 5.

2.1 The likelihood layer

The first layer in Equation (2.2), often called the likelihood layer, is rather straightforward for a given distribution. However, the choice of distribution is very important and should be considered carefully. Consider a dataset for count data, as in the application for this thesis. Then, the most common options for the likelihood are the Poisson and the negative binomial distributions. The negative binomial uses a mean-dispersion parametrization. One of the distributions are not inherently better than the other, but they have different strengths and weaknesses. For example, the Poisson distribution for a random variable X only has one parameter η and $E[X] = \eta = Var[X]$. Note that the symbol η is used instead of the traditional λ to follow the notation of the upcoming latent layer, where η often is used. Intuitively, variance increasing linearly with the mean makes sense, as the variance should increase when the count increases, and a linear increase appears reasonable. The negative binomial is very similar, but it adds a dispersion parameter α which allows the variance to exceed the value of the mean, specifically $Var[X] = \eta + \frac{\eta^2}{\alpha}$. This is more appropriate if the data has higher variance than the mean would indicate.

As such, some analysis of the count data should be performed prior to deciding a likelihood. Similarly, for other types of data, the user must choose between multiple likelihood candidates and weigh the advantages against the disadvantages. When a distribution has been chosen the likelihood layer is easy to define. However, it should be noted that the parameters of the likelihood typically have separate values for each datapoint in the dataset. Thus, for spatial count data $\mathbf{y} = (y_1, ..., y_N)^T$ for N regions with a Poisson likelihood, the likelihood layer becomes:

$$y_i \mid \eta_i \sim Poisson(\eta_i), \quad i = 1, ..., N.$$

The latent parameter $\boldsymbol{\eta}$ is then further defined in the latent layer.

2.2 The latent layer

The second layer in Equation (2.2), often called the latent layer, models a parameter of interest from the likelihood layer. Building on the example above, the vector $\boldsymbol{\eta} = (\eta_1, ..., \eta_N)^T$ could be modelled as a simple linear model based on measured covariates, often called fixed effects. For the Poisson likelihood a log-link is genrally used between the two layers. Thus,

$$\log(\eta_i) = \mu + \beta x_i, \quad i = 1, ..., N.$$

Alongside a linear term for the data $\mathbf{x} = (x_1, ..., x_N)^T$ a global intercept μ is included. Now, this is not sufficient for many applications as some of the important covariates often are unmeasured. Then, random effects are introduced to model the unexplained patterns in the data. Here there are two main subgroups, namely structured and unstructured random effects. In a spatial context, this differentiates whether the model considers the placement of the regions in space or not. A popular choice for the structured component is the ICAR, which will be further explained in Section 2.5. The ICAR is often combined with an unstructured effect as well, for instance an IID model. Let $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ denote the ICAR and the IID, both with dimension $(1 \times N)$, then the latent layer becomes

$$\log(\eta_i) = \mu + \beta x_i + \phi_i + \gamma_i, \quad i = 1, ..., N. \tag{2.4}$$

More generally, both the ICAR and the IID belongs to a subset of random effects commonly referred to as Gaussian Markov random fields (GMRF). This subset is especially relevant for a subgroup of Bayesian hierarchical models called latent Gaussian models (LGM) (Rue *and others*, 2016). These models require all the components of the latent layer to follow a Gaussian distribution to ensure that the combined distribution is Gaussian, i.e. that $\log(\boldsymbol{\eta}) \sim N(..., ...)$ for this example.

Formulating the priors for the latent layer has already been mentioned inadvertently. When choosing $\boldsymbol{\phi}$ to be an ICAR and $\boldsymbol{\gamma}$ to be an IID, both the random effects have effectively been assigned priors. The technical details will be expanded on in Section 2.5. Additionally, priors must be set for μ and β. Generally, $\mu, \beta \in \mathbb{R}$, so they could for example be assigned a uniform prior like $U(-\infty, \infty) \propto 1$ or a Gaussian prior $N(0, 10^{-6})$. Some other options are the Cauchy and student-t distributions. Note that of the four mentioned distributions, only Gaussian priors would be compatible

with a latent Gaussian model. Although, $N(0, \infty) \propto 1$ is equivalent to $U(-\infty, \infty)$, so one could argue that a flat uniform prior for the real line is compatible with LGMs.

The last point of interest for the latent layer is the link function. In the examples above the log-link has been used, which is the standard link for the Poisson distribution. The general idea is to model some transformation of the parameter in the likelihood which behaves better than the non-transformed parameter. The link or transformation in question is tied to the likelihood and some other typical pairs are the identity-link for a Gaussian distribution and the logit-link for a Bernoulli distribution. However, it is not a requirement to follow these conventions as most link functions can be combined with most likelihoods. Depending on the chosen likelihood layer and latent layer some hyperpriors might need specification in the following layer.

2.3 The hyperprior layer

The third and final layer in Equation (2.2) is responsible for defining the hyperpriors for the first two levels. A hyperprior is a prior for a parameter used in a higher up layer. For example if the likelihood is negative binomial, like $NegBin(\eta_i, \alpha)$ with the latent layer modelling $\boldsymbol{\eta}$ following the notation above. Then the dispersion parameter α would be given a hyperprior in the hyperprior layer. For the example with the Poisson likelihood there are no hyperpriors needed for the likelihood layer, but there are some hidden parameters for the random effects in Equation (2.4) which must be defined. As will be made clear in Section 2.5, this typically involves some variance controlling parameters like the standard deviation $\sigma \in \mathbb{R}$ or the precision $\tau > 0$, and of course $\tau = \frac{1}{\sigma^2}$. In terms of the precision τ there are a few notable options.

The first is the penalized complexity priors, denoted $\tau \sim PC(u, a)$ (Simpson *and others*, 2017). The probability $a \in (0, 1)$ and the specified value $u \in (0, \infty)$. The condition $P(\sigma > u) = a$ is satisfied when $\tau \sim PC(u, a)$, and equals an exponential prior on σ with a rate λ found from the condition. The PC prior for τ is then found by a transformation of the exponential prior for σ. PC priors have been designed to favour the base model, i.e. the simpler version of the model. Which in this case is infinite precision or a variance of zero. This means that the precision should only by decreased if the data supports it. A common choice of u and a gives $\tau \sim PC(1, 0.01)$ (Aleshin-Guendel and Wakefield, 2024). It should be noted that this puts the majority of the prior weight for τ inside $[0, 100]$ as shown in Figure 2.1a.

The second is the gamma distribution $gamma(\alpha, \beta)$, with a shape-rate parameterization. This has long been the default prior in INLA, which will be introduced in Section 2.6.2, specifically $\tau \sim gamma(1, 0.00005)$. Compared to the PC prior above, this gamma prior is more heavy-tailed for large τ. This is shown in Figure 2.1c.

The third is a uniform prior on the standard deviation σ which is then transformed to the precision τ. It should be noted that a uniform prior on σ is quite different to

a uniform prior on τ. A typical range for the standard deviation, for instance used in Corpas-Burgos and Martinez-Beneito (2020), is $\sigma \sim U(0, 5)$, but this range could vary between applications. An equivalent definition of this prior is $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$.

Three popular priors for a precision τ have been introduced. There exists other alternatives as well, for example half-normal and half-Cauchy which are discussed in Gelman (2006). However, the primary three priors which will be used in this thesis can be summed up as follows:

$$\begin{aligned} &\text{Penalized complexity:} && \tau \sim PC(1, 0.01) \\ &\text{Gamma:} && \tau \sim gamma(1, 0.00005) \\ &\text{Uniform:} && \frac{1}{\sqrt{\tau}} \sim U(0, 5). \end{aligned} \tag{2.5}$$

They have somewhat different properties, which is visualized in Figure 2.1. Notably, the PC prior in Figure 2.1a and the uniform prior in Figure 2.1b mainly give prior weight to precisions below 100 or 1 respectively before slowly converging to 0 prior weight for large precisions. The gamma prior is much wider, which is evident in Figure 2.1c. This means that the gamma prior is more likely to yield a higher precision parameter for the posterior compared to the PC and uniform priors. However, it should be noted that the gamma prior also has its max values for low precision, but as shown in Figure 2.1d, the difference between the prior density for $\tau \approx 0$ and $\tau = 2$ is negligible, while this difference is substantial for the PC and uniform priors. Additionally, these priors will be relevant for the validation study in Section 5.

Now, a short comment on the relevance of the hyperpriors for latent Gaussian models. Even though LGMs restrict the latent layer to be Gaussian, the hyperpriors for parameters in the latent layer need not be Gaussian. Thus, the hyperprior for a precision from the ICAR or IID in Equation (2.4) could be any of the three priors from Equation (2.5).

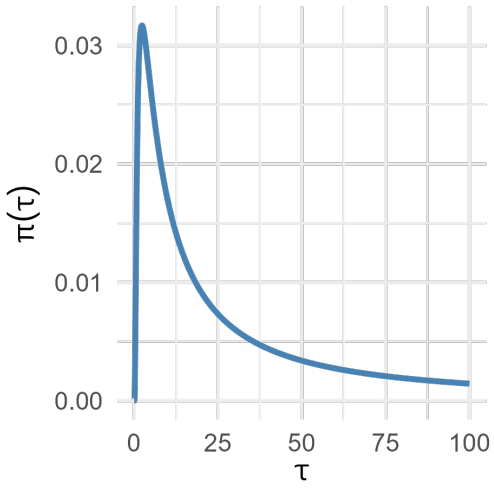
2.4 Illustration of priors for a linear regression example

As the priors and hyperpriors in a Bayesian hierarchical model clearly play a big role, lets explore this with some theory and an example of what not do to. The importance of priors can be seen directly in the definition of the posterior distribution in Equation (2.3). Note that the effect of the prior depends on the amount of data. This is because more data makes the posterior more data-driven and more dependent on the likelihood term, while the priors carry less weight. A cautionary example of the importance of priors could be in a simple regression:

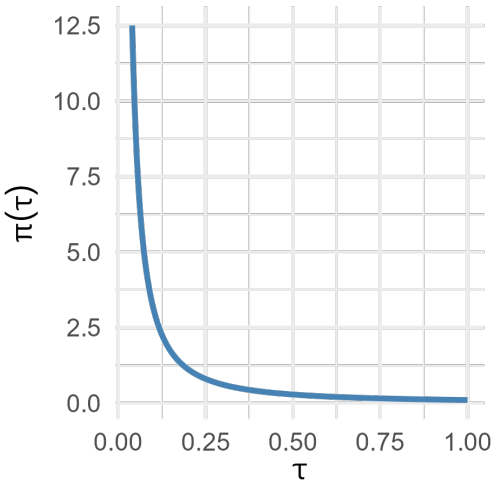
$$y_i = \mu + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \quad \text{for } i = 1, ...5,$$

where $\mathbf{x}$ is a covariate and $\mathbf{y}$ is the response variable. As a Bayesian hierarchical model it could be written as:

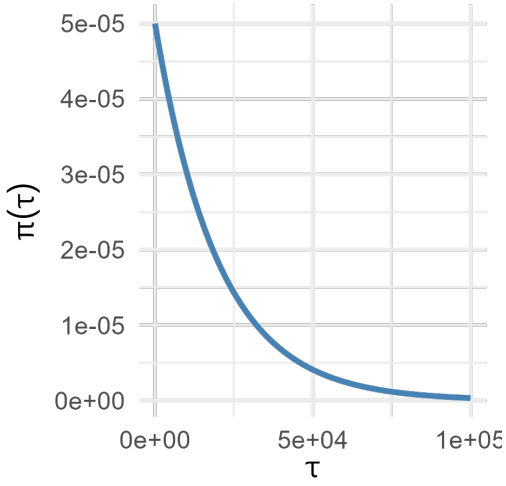
$$\begin{aligned} &1.\ y_i \sim N(\eta_i, \sigma^2), \quad \text{for } i = 1, ...5 \\ &2.\ \eta_i = \mu + \beta x_i, \quad \text{for } i = 1, ...5 \\ &\qquad \text{Prior for } \mu \text{ and } \beta. \end{aligned} \tag{2.6}$$



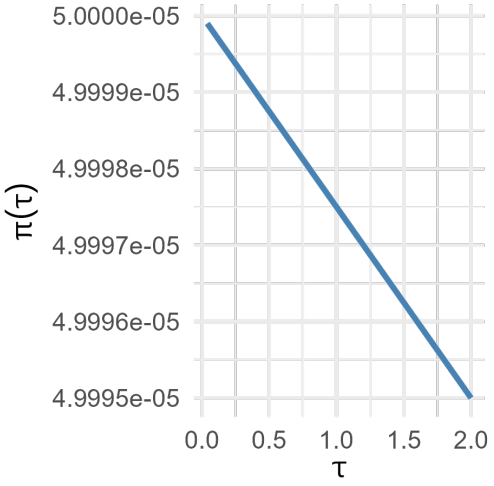
(a) $\tau \sim PC(1, 0.01)$.



(b) $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$.



(c) $\tau \sim gamma(1, 0.00005)$.



(d) $\tau \sim gamma(1, 0.00005)$.

Figure 2.1: Three different priors for a precision parameter τ. (c) and (d) shows the same prior for different intervals on the x-axis.

Equation (2.6) assumes that σ is known. Furthermore, assume one wide and one narrow prior is used for each parameter, for instance $N(0, 10^6)$ or $N(0, 10^{-2})$ for μ and $N(0, 10^6)$ or $N(2, 10^{-2})$ for β. The four combinations of these priors have been applied to the Bayesian hierarchical model in Equation (2.6) and yields the results in Figure 2.2. The data is fictitious and the wide priors in Figure 2.2a gives $\beta = 5.4$ and $\mu = 85$, which represents the best fitted values. When Figure 2.2b restricts the intercept to 0 with a strict prior the fitted line is steeper, as the intercept is lower. Note, the difference between the data points and the line increases. Similarly, when restricting the coefficient β to 2 in Figure 2.2c the intercept increases to compensate, but the error is still larger than in Figure 2.2a. Combining the two strict priors in Figure 2.2d is clearly not the way to go. Now, the strict priors do not give worse results because they are strict, but rather because they are centred at the wrong place. For instance using $\beta \sim N(5.4, 10^{-2})$ instead of $\beta \sim N(2, 10^{-2})$ in Figure 2.2c would give a similar fit as in Figure 2.2a. The overall idea is that when the value of a parameter is uncertain, a wide prior, or an uninformative prior should be used. It is usually never wrong to use uninformative priors, just note that the posterior is more data-driven. However, in some situations, when strong prior beliefs for a parameter value are warranted, a more strict, or informative prior, could be beneficial. Note that the informative priors employed in Figure 2.2 are extremely narrow, and badly centred, so the posterior parameters can hardly deviate from the given prior mean at all.

It should also be noted that no priors are truly uninformative. As mentioned, even a uniform prior has a different effect depending on the parametrization of the parameter. For further discourse on informative and uninformative priors see for example Irony and Singpurwalla (1997) and Berger (2006).

Now, as the different parts of the Bayesian hierarchical model have been introduced it is time to focus on candidate models for the latent layer. Specifically, GMRFs like the ICAR and IID need a formal definition with a theoretical background.

2.5 Gaussian Markov random fields

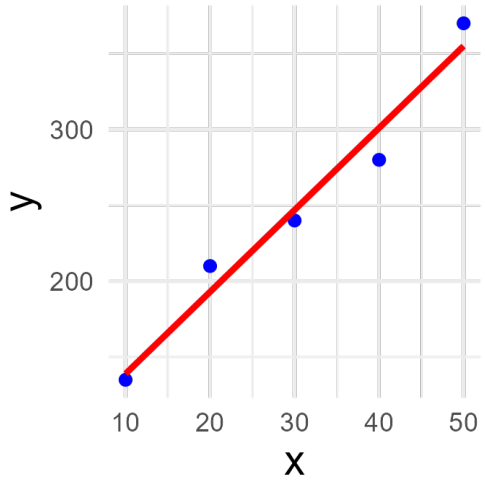
GMRFs play an important role in the latent layer of the Bayesian hierarchical models. They are especially crucial to the latent Gaussian models because of their limitation to Gaussian distributions. But what is a GMRF? Lets start with a formal definition and then explain the different parts. The following definition is from Chapter 2.2 in Rue and Held (2005), with the added assumption that $\boldsymbol{\mu} = \mathbf{0}$.

Definition: A random vector $\mathbf{x} \in \mathbb{R}^N$ is called a GMRF with respect to a labelled graph $G = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu} = \mathbf{0}$ and symmetric positive definite precision matrix $\mathbf{Q}$ iff. its density can be written as

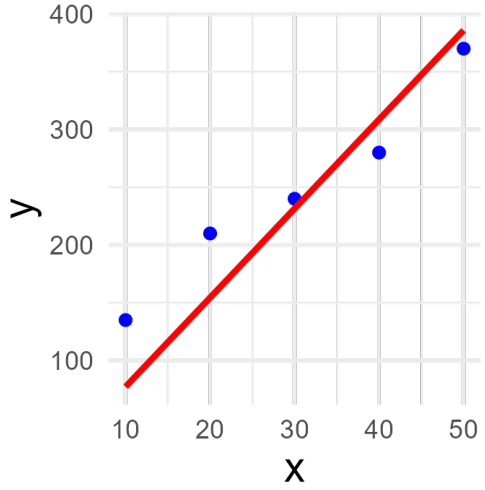
$$\pi(\mathbf{x}) = (2\pi)^{-N/2}|\mathbf{Q}|^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\right) \tag{2.7}$$

and $Q_{ij} \neq 0 \iff \{i, j\} \in \mathcal{E} \quad \forall \quad i \neq j$.

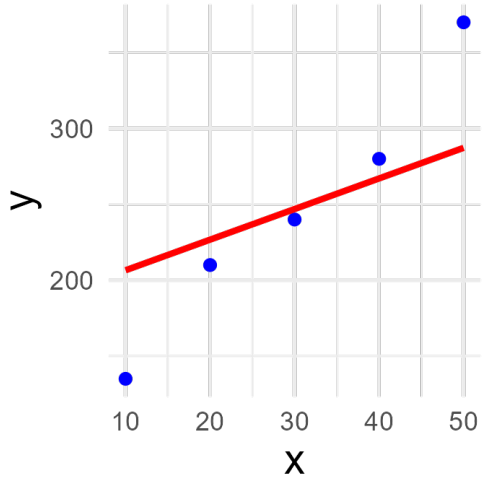
The assumption that $\boldsymbol{\mu} = \mathbf{0}$ will be used throughout this thesis for ease of notation,



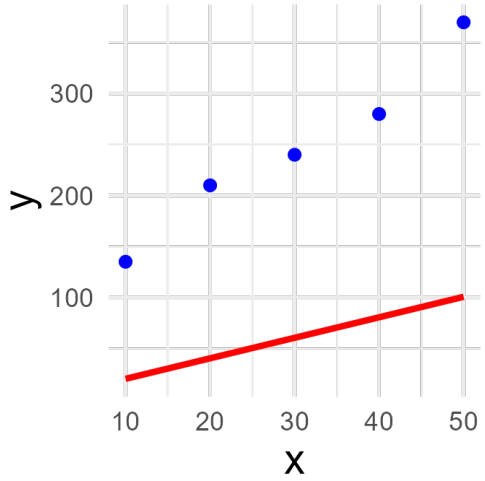
(a) $\mu \sim N(0, 10^6)$ and $\beta \sim N(0, 10^6)$



(b) $\mu \sim N(0, 10^{-2})$ and $\beta \sim N(0, 10^6)$



(c) $\mu \sim N(0, 10^6)$ and $\beta \sim N(2, 10^{-2})$



(d) $\mu \sim N(0, 10^{-2})$ and $\beta \sim N(2, 10^{-2})$

Figure 2.2: The blue points are the data points (x_i, y_i) and the red lines are the fitted linear regression for the priors specified in the subcaptions.

as well as being a common assumption for most GMRFs. The labelled graph G consists of a node set $\mathcal{V}$ and an edge set $\mathcal{E}$. As it is labelled, each node in $\mathcal{V}$ has an associated label, typically an integer, which for N nodes could be from the set $\{1, ..., N\}$. The edge set $\mathcal{E}$ tracks which nodes have an edge between them. An example could be $G^* = (\mathcal{V}^*, \mathcal{E}^*)$ with $\mathcal{V}^* = (1, 2, 3, 4)$ and $\mathcal{E}^* = ((1, 2), (1, 3), (2, 3), (2, 4), (3, 4))$, which yields the graph in Figure 2.3.

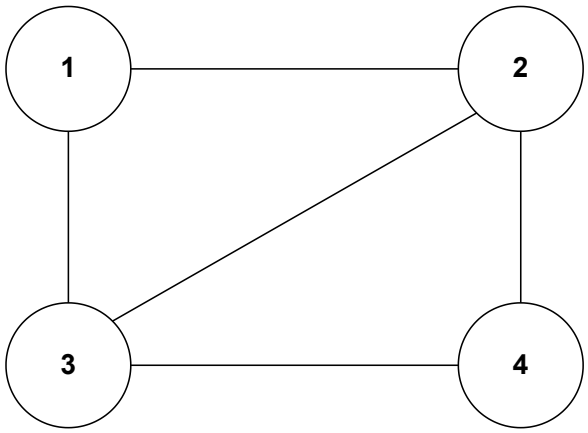


Figure 2.3: Illustration of the graph $G^* = (\mathcal{V}^*, \mathcal{E}^*)$.

Another key concept of the graph is in regard to neighbours. In this thesis the main interests lies in first-order neighbourhood structures. Thus, the neighbours of a node i are the nodes j with an edge from i. This means that i and j are neighbours if and only if $(i, j) \in \mathcal{E}$ for the associated edge set $\mathcal{E}$. For example, the neighbours of node 1 in the example graph G^* are the nodes 2 and 3. The set of neighbours for a node i is denoted $ne(i)$, so $ne(1) = \{2, 3\}$. Another useful notation for later is n_i, which represents the number of neighbours for node i. In the example graph this means that $n_1 = n_4 = 2$ and $n_2 = n_3 = 3$.

Furthermore, the definition states that all the non-zero elements on the off-diagonals of the precision matrix $\mathbf{Q}$ coincides with an edge in G. This touches on a key part of the relationship between the precision matrix $\mathbf{Q}$ and the graph structure. Namely the conditional independence between to nodes i and j which are not neighbours. Specifically, a node i is conditionally independent of all non-neighbouring nodes given all its neighbours. Formally,

$$x_i | \mathbf{x}_{ne(i)} \perp\!\!\!\perp x_j \quad \text{for } j \in \{1, ..., N\} \text{ and } j \notin \{i, ne(i)\}.$$

As off-diagonal elements Q_{ij} are only allowed to deviate from 0 when $\{i, j\} \in \mathcal{E}$, it is clear that $x_i | \mathbf{x}_{-(i,j)} \perp\!\!\!\perp x_j \iff (i, j) \notin \mathcal{E} \iff Q_{ij} = 0$. Note that a negative subscript, like $\mathbf{x}_{-i}$, denotes the vector $\mathbf{x}$ without the i-th element. For $\mathbf{x}_{-(i,j)}$ both the i-th and the j-th elements are removed from $\mathbf{x}$.

Lets recall the density for a multivariate normal which is commonly written as

$$\pi(\mathbf{x}) = (2\pi)^{-N/2} |\mathbf{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{x}\right),$$

assuming that $\boldsymbol{\mu} = 0$. If the traditional covariance matrix $\boldsymbol{\Sigma}$ is exchanged with the precision matrix $\mathbf{Q}$, the above expression equals the density for a GMRF in Equation (2.7), as $\boldsymbol{\Sigma} = \mathbf{Q}^{-1}$. This means that a GMRF is simply a multivariate normal distribution with a labelled graph G based on the structure of the precision matrix $\mathbf{Q}$. GMRFs are also traditionally defined with the precision matrix $\mathbf{Q}$ instead of the covariance matrix $\boldsymbol{\Sigma}$ because of its relationship with conditional independence as well as its sparse structure. Sparsity is advantageous for both storage as well as the necessary computations involved in model fitting and evaluation. Additionally, the precision parameter τ is often used instead of σ^2 as the variance parameter.

An important subset of GMRFs are intrinsic GMRFs, or IGMRF [Chapter 3](Rue and Held, 2005). These distributions have a precision matrix $\mathbf{Q}$ without full rank, often of rank $N - 1$, and require that $\mathbf{Q}\mathbb{1} = \mathbf{0}$ (Rue and Held, 2005). For a higher rank deficiency than 1 more restrictions are imposed, see [Chapter 3](Rue and Held, 2005). The rank deficiency means that $\mathbf{Q}^{-1}$ is no longer defined, so the generalized inverse is used instead, denoted $\mathbf{Q}^{-}$. As $\mathbf{Q}$ is not full rank, the determinant $|\mathbf{Q}| = 0$. Thus, the determinant is exchanged with the product of the non-zero eigenvalues of $\mathbf{Q}$, the pseudo-determinant, which will be written as $|\mathbf{Q}|^*$. The resulting density of an IGMRF can be written as

$$\pi(\mathbf{x}) = (2\pi)^{-(N-1)/2}(|\mathbf{Q}|^*)^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\right), \tag{2.8}$$

again with the assumption that $\boldsymbol{\mu} = \mathbf{0}$.

Another important grouping of GMRFs is the distinction between structured and unstructured GMRFs, which are subgroups of structured and unstructured random effects. The distinction arises because some GMRFs use the relative position of the datapoints, either in time, space or both, while other GMRFs discounts this aspect of the data. One group is not inherently better than the other, they just make different assumptions and are thus more suited to different purposes. For instance, unstructured effects assume that there are no relations between the different datapoints, while the structured effect assumes there is some unknown structure in the data. However, it turns out they are best at modelling the data when combined with each other. Thus, many of the most popular models in the literature combines the two, for instance like the BYM model (Besag *and others*, 1991).

Spatial GMRF examples

Now that the general definitions and some vocabulary for GMRFs have been defined, the next step is to formally define the ICAR and the spatial IID, as these traditional GMRFs are of interest for this thesis. Other spatial GMRFs are described in Appendix A.1 and temporal examples are described in Appendix A.2. For those interested, spatio-temporal models have been described in Appendix A.3. The models in Appendix A are included to give an overview of the available models, but they are not directly relevant for this thesis.

Lets start by briefly introducing the spatial modelling situation. The goal is to model the response variable in space, typically for areal data. The associated graph then represents the neighbourhood structure of the regions in space. For areal data, a pair of regions are often considered neighbours if they share some part of their border,

and all pairs of neighbours are connected by edges in the graph. Furthermore, the notation $i \sim j$ means that region i and j are neighbours. This thesis will be limited to fully connected graphs, which means that all the areas of interest must connect to each other, i.e. no islands or areas separated by areas which are not included in the data. An example of an areas of interest is shown in Figure 2.4, which shows four provinces in the north-west of Spain with the associated graph, which coincides with the earlier example graph in Figure 2.3.

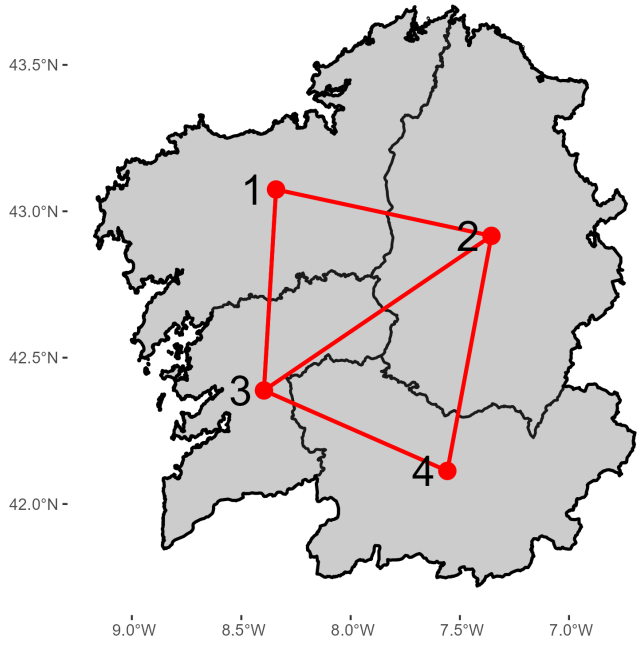


Figure 2.4: Map of the four provinces in the autonomous region Galicia in the north-west of Spain and the overlaid graph for a first order neighbourhood structure.

Intrinsic conditional autoregressive

The intrinsic conditional autoregressive model, commonly known as ICAR, is a staple in spatial modelling. It has been widely used since its introduction in Besag (1974) with almost 10000 references, and many newer models are strongly inspired by it, as will be clear in Section 3. It is an edge case of the more general conditional autoregressive (CAR) models also introduced in Besag (1974). As the name implies, the ICAR is an intrinsic GMRF, and its joint density for a vector $\mathbf{x} = (x_1, ..., x_N)^T$ for N regions, can be written as follows:

$$\pi(\mathbf{x} \mid \tau) = (2\pi)^{-(N-1)/2}(|\mathbf{Q}|^*)^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\right).$$

The precision matrix $\mathbf{Q} = \tau\mathbf{R}$ with the structure matrix $\mathbf{R}$ defined as follows:

$$R_{ij} = \begin{cases} n_i, & i = j \\ -1, & i \sim j \\ 0, & \text{else.} \end{cases}$$

Recall that n_i is the number of neighbours for region i and that $i \sim j$ means that the regions i and j are considered neighbours, which in this thesis means that i and j share some part of their borders. This is a common first-order neighbourhood structure which is generally the standard for the ICAR, but the ICAR can work with any neighbourhood structure as it only changes the structure matrix R.

In general, any neighbourhood structure is based on the assumption that regions near each other have more in common than regions further apart, which is often a reasonable assumption. Furthermore, the ICAR assumes spatial stationarity, which means that all pairs of neighbours affect each other the same, regardless of the two regions in question. This results in a stationary spatial smoothing, which means that the degree of spatial smoothing is invariant in space.

As an example, the structure matrix from Figure 2.3 and Figure 2.4 for the four regions would be

$$\mathbf{R} = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}.$$

An equivalent way of defining the structure matrix $\mathbf{R}$ for the ICAR is with a binary weight matrix $\mathbf{W}$. This matrix only accounts for the dependence between regions, and is therefore equal to 0 on the diagonal, which gives the following definition:

$$W_{ij} = \begin{cases} 1, & \text{if } i \sim j \\ 0, & \text{else.} \end{cases}$$

Note that the non-zero weights are positive. This weight matrix is then used to construct the matrix $\mathbf{D}$ as

$$\begin{aligned} \mathbf{D} &= diag(w_{1+}, ..., w_{N+}) \\ w_{i+} &= \sum_{k=1}^{N} w_{ik}, \quad i = 1, ..., N. \end{aligned} \tag{2.9}$$

This ensures that all the rows sum to 0. Then, the structure matrix is defined as the sum $\mathbf{D} - \mathbf{W}$, which coincides with the structure matrix $\mathbf{R}$ above. This definition with $\mathbf{W}$ and $\mathbf{D}$ might seem unnecessary, but it will come in handy in Section 3. With the binary weight structure, the conditional distributions are

$$x_i \mid \mathbf{x}_{-i}, \tau \sim N\left(\frac{1}{n_i}\sum_{k\sim i} x_k, \frac{1}{n_i \tau}\right), \quad i = 1, ..., N.$$

The conditional mean is the average of the neighbouring regions and the variance is inversely correlated to the number of neighbours n_i. In essence, the ICAR works as a spatial smoother, and facilitates borrowing of strength across regions. However, the assumption of spatial stationarity may in some applications be too restricting, as will be discussed in Section 4. The ICAR can also be viewed as a RW1 in space, as an ICAR for a line of regions, which would give the same neighbourhood graph as a temporal line, see Figure A.1, yields the same structure matrix as a RW1 (for details see Appendix A.2).

IID

In discrete space, the standard unstructured random effect is an IID model. The IID model captures spatial heterogeneity, which can prevent overfitting the noise aspect of the data, for an example see Figure A.2 in Appendix A.2. Formally, for a vector $\mathbf{x} = (x_1, ..., x_N)^T$ for N regions the IID model is defined as:

$$x_i \mid \sigma \overset{iid}{\sim} N(0, \sigma^2) \quad i = 1, ..., N.$$

As mentioned, the IID model will not be used in this thesis, but for most applications it should be included in some fashion.

2.6 Inference

When working with Bayesian hierarchical models, obtaining the posterior distribution is often challenging. This can be seen from Equation (2.3), where calculating the integral in the denominator rarely has an analytic form. The integral is also often of a high dimension which makes numerical approximation difficult. A counter example is when using conjugate priors, but this is often rather restrictive. Instead, the traditional methodology for computing the posterior in a Bayesian hierarchical setting is the Markov chain Monte Carlo(MCMC) sampler methods (see for example Robert and Casella (2004)). These are very flexible, but can be slow for large problems. A notable competitor to MCMC for latent Gaussian models is the integrated nested Laplace approximations, which instead of sampling methods uses approximations to obtain information about the marginal posterior distributions. Both of these methods will be used in this thesis, and a short introduction of both methodologies is in order.

2.6.1 Markov chain Monte Carlo

MCMC is a much used sampling technique, and in theory any precision can be obtained with sufficient computational power (Tierney, 1994). By viewing the denominator in Equation (2.3) as a constant, the posterior becomes

$$\pi(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta})\pi(\boldsymbol{\eta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

However, as MCMC methods generally combines the parameters in $\boldsymbol{\theta}$ and the latent layer $\boldsymbol{\eta}$ into one vector $\boldsymbol{\theta}$, the posterior simplifies to

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

The essence of a MCMC sampler is to explore the posterior distribution with the help of the likelihood and priors for the latent layer $\boldsymbol{\eta}$ and all the parameters in $\boldsymbol{\theta}$. Specifically, the three known distributions are often involved in transitions between points in the posterior distribution, and with a sufficient number of samples, an empirical approximation of the samples can be used for inference. This typically

involves sampling from the full conditional distributions. There exists multiple different methods inside the MCMC framework, with different strengths and requirements. A multitude of programs are available to use, for instance WinBUGS, JAGS, STAN and NIMBLE. This thesis will use WinBUGS, introduced in Lunn *and others* (2000), through a R-interface. As WinBUGS employs a Gibbs sampler, lets give a brief introduction to this method.

The Gibbs sampler, introduced in Geman and Geman (1984), samples from conditional distributions (Gelfand and Smith, 1990). These conditional distributions can depend conditionally on other relevant parameters, so the sampler needs $\pi(\theta_j|\boldsymbol{\theta}_{-j})$ for $j = 1, ..., J$ where J is the number of parameters involved. Note that $\boldsymbol{\theta}$ now also includes the parameters in the latent layer $\boldsymbol{\eta}$. Then, a cycle or a step typically involves sampling all the values. Assume $\boldsymbol{\theta}^1$ is known and the superscript indicates the iteration number, then

$$\begin{aligned}
\theta_1^2 &\sim \pi(\theta_1|\boldsymbol{\theta}_{-1}^1) \\
\theta_2^2 &\sim \pi(\theta_2|\theta_1^2, \boldsymbol{\theta}_{-(1,2)}^1) \\
&\vdots \\
\theta_J^2 &\sim \pi(\theta_J|\boldsymbol{\theta}_{1,...,J-1}^2, \theta_J^1).
\end{aligned} \tag{2.10}$$

This step can then be repeated L times, which is often called the length of the chain. Note that the sampling does not have to be ordered as above, it merely requires that each parameter is sampled infinitely often (Gelfand and Smith, 1990). As mentioned, when the Gibbs sampler is used for a Bayesian hierarchical model, it also samples the random effects. If the random effect $\mathbf{x}$ is included as an ICAR, then $x_i \; \forall \; i$ must also be sampled. A consequence is that the dimension of $\boldsymbol{\theta}$ often is very large. However, as the conditional distributions given in Section 2.5 only depend on a few neighbours, most parameters in $\boldsymbol{\theta}$ only depend conditionally on a few neighbours.

There are some important concepts to be aware of in regards to the Gibbs sampler, and other MCMC methods. Namely, that both the initial value $\boldsymbol{\theta}^1$ and the length of the chain L can affect the approximated posterior distribution. This is because a number of samples $\boldsymbol{\theta}^l$ will be dependent on the initial value $\boldsymbol{\theta}^1$. The idea is to let the system reach a steady state, which means that the samples are independent of the initial value. In practice, there is no way to guarantee that a steady state has been reached, but the time needed to reach a steady state generally increases with the dimension of $\boldsymbol{\theta}$ and will also vary based on the initial value $\boldsymbol{\theta}^1$ for a given model. To remedy this issue, a burn-in period is introduced. The samples taken during the burn-in are not used for approximating the empirical posterior distribution, and the length of the burn-in must be considered carefully. The length of the chain must also be chosen with care. Specifically, a too short length increases the expected error, and increases the probability that parts of the parameter space remains unexplored. Generally, a longer chain gives better inference, so the length of the chain is often a question of the time and computational resources at hand. Just to be precise, the length of the chain is the total number of simulated samples for each parameter while the burn-in removes the first samples for each parameter. Thus, the samples used for posterior inference is then the length L minus the burn-in.

WinBUGS also supplies some tools to assess whether the samples are problematic or not. For example, plots are automatically made for each of the parameters, where it might be visible that it has not reached a steady state yet and that the burn-in should be increased. Other examples are the effective sample size (ESS) and an assessment criteria $\hat{R}$ (Gelman and Rubin, 1992) which is only available when multiple independent chains have been sampled, generally in parallel. These are available for each sampled parameter in $\boldsymbol{\theta}$ and are denoted as "n. eff" and "Rhat" respectively in the R interface for WinBUGS. The ESS gives an approximation of the number of independent samples a given chain can be compared to [Chapter 18](Ross, 2022). For instance, a highly correlated chain with 1000 samples could have an ESS of 10 for a given parameter, which is of course too low. The $\hat{R}$ compares the variance between chains and inside a chain, and it should approach 1 if all chains have reached the steady state. As a rule of thump, $\hat{R}$ should be below $1,1$ [Chapter 18](Ross, 2022) and the effective sample size should be large enough, for instance above 200. If either the $\hat{R}$ or the ESS indicates the chains have not converged, the length of the chains should be increased. If this does not help, there might be a more favourable reparametrization or simply an innate issue with the model being fitted, such as high correlation (Ross, 2022). To be precise, both the ESS and the $\hat{R}$ is calculated individually for each parameter in the model.

2.6.2 Integrated nested Laplace approximations

Integrated nested Laplace approximations (INLA) is a much used methodology for posterior inference for latent Gaussian models introduced in Rue *and others* (2009). As MCMC methods often have high runtimes, one of the main advantages of INLA is the substantially shorter runtimes (Rue *and others*, 2009). This is mainly due to the sampling nature of MCMC, while INLA avoids this by numerical approximations. Furthermore, the user avoids setting parameters like burn-in and the length of the chain, which will influence the inference from a MCMC sampler. Therefore, INLA will be used for latent Gaussian models in this thesis. Note that the implementation of INLA has recently changed and now uses variational Bayes. For details on this newer implementation see Van Niekerk *and others* (2023). The following introduction to INLA is based on Martino and Riebler (2020), and focuses on the traditional implementation.

Instead of sampling from the posterior as in MCMC it approximates the marginal densities with a collection of approximations, including the Laplace approximation and numerical integration. As previously, $\mathbf{y}$ represents the data, $\boldsymbol{\eta}$ is the latent field and $\boldsymbol{\theta}$ is the parameters and hyperparameters. Specifically, the marginal densities of interest are expressed as:

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-j} \tag{2.11}$$

$$\pi(\eta_i|\mathbf{y}) = \int \pi(\eta_i|\boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{2.12}$$

The first step is to approximate $\pi(\boldsymbol{\theta}|\mathbf{y})$. The following identity can be applied:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{y}|\boldsymbol{\eta},\boldsymbol{\theta})\pi(\boldsymbol{\eta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\eta}|\boldsymbol{\theta},\mathbf{y})}.$$

As the three densities in the numerator are known through the likelihood and the priors, only the denominator needs an approximation. This is done through a Gaussian approximation by matching the mode and the curvature at the mode, denoted $\widetilde{\pi}_G(\boldsymbol{\eta}|\boldsymbol{\theta},\mathbf{y})$. Thus, the approximation of $\pi(\boldsymbol{\theta}|\mathbf{y})$ becomes

$$\widetilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{y}|\boldsymbol{\eta},\boldsymbol{\theta})\pi(\boldsymbol{\eta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\widetilde{\pi}_G(\boldsymbol{\eta}|\boldsymbol{\theta},\mathbf{y})}. \tag{2.13}$$

This approximation to a marginal density coincides with the Laplace approximation from Section 4 of Tierney and Kadane (1986).

The second step is to approximate $\pi(\eta_i|\boldsymbol{\theta},\mathbf{y})$. Here there are three different options. The most straightforward option is simply to compute the marginals of the Gaussian approximated joint density in the denominator of Equation (2.13). This is very effective, but not always very precise. Another option is to use the Laplace approximations as in Equation (2.13). Then,

$$\widetilde{\pi}(\eta_i|\boldsymbol{\theta},\mathbf{y}) \propto \frac{\pi(\mathbf{y}|\boldsymbol{\eta},\boldsymbol{\theta})\pi(\boldsymbol{\eta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\widetilde{\pi}_G(\boldsymbol{\eta}_{-i}|\eta_i,\boldsymbol{\theta},\mathbf{y})}. \tag{2.14}$$

This is a good approximation, but very computationally expensive. A third approach is in between the previous two approaches. Called simplified Laplace, it uses a Taylor's series expansion around the mode of the joint density to improve both location and skewness, before calculating the marginals.

The third, and last, step is to approximate the integrals in Equation (2.11) and Equation (2.12) by numerical integration. The first step here is to locate a subset $\{\boldsymbol{\theta}^k\}_{k=1}^K$ from the high density regions of the approximated density $\widetilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. This can be done by a grid around the mode, and some process for which points of the grid to include. Then, the numerical integration can be written as

$$\pi(\theta_j|\mathbf{y}) = \sum_{k=1}^{K} \widetilde{\pi}(\boldsymbol{\theta^k}|\mathbf{y})\Delta_k \tag{2.15}$$

$$\pi(\eta_i|\mathbf{y}) = \sum_{k=1}^{K} \widetilde{\pi}(\eta_i|\boldsymbol{\theta^k},\mathbf{y})\widetilde{\pi}(\boldsymbol{\theta^k}|\mathbf{y})\Delta_k \tag{2.16}$$

where the densities inside the sums are calculated from Equation (2.13) and one of the three methods to approximate $\pi(\eta_i|\boldsymbol{\theta},\mathbf{y})$, for each value $\boldsymbol{\theta}^k$. The weights Δ_k for $k = 1, ..., K$ will depend on the locations of the $\boldsymbol{\theta}^k$'s.

3 Adaptive Gaussian Markov random fields

As the typical assumption of stationarity in a GMRF is sometimes too restrictive, there exists a plethora of GMRFs that drop the stationarity assumption, often called adaptive GMRFs (AGMRF). Temporal situations where AGMRF are appropriate could be when some of the data was collected during events which affects the response variable, for instance war or a financial crisis, or maybe the measuring method of the response variable changed in the time period of interest. Similar applications have been examined in Aleshin-Guendel and Wakefield (2024) and Sand-Larsen (2025). In the spatial case there could similarly be spatial heterogeneity due to war and financial crisis which affects parts of the area of interest, or different ways of measuring the response variable in different areas. Additionally, geographical differences, like biomes, dividers like mountains and rivers or administrative regions could be of interest. For spatio-temporal data all of the above could occur. The definitions of the GMRFs in Section 2.5 and Appendix A all only had one precision parameter τ and a spatially invariant precision matrix $\mathbf{Q}$ from the assumption of stationarity, and thus a constant smoothing for the whole area of interest. As this is not always appropriate, this Section aims to introduce a number of proposed adaptive GMRFs as well as their strengths and use cases. First, some proposed models to give an overview of the field before the AGMRFs which will play a role in the validation study in Section 5 are defined. Again, this will mainly focus on spatial models, and two temporal models are defined in Appendix B.

3.1 Motivation and literature review

A more flexible alternative to the rigid binary weight matrix $\mathbf{W}$ is a stochastic weight matrix. There are many different approaches to making $\mathbf{W}$ stochastic, and a few of them will be introduced shortly. Furthermore, they all add another layer as the weight matrix now depends on new parameters with new priors and potential hyperparameters with additional hyperpriors.

Adaptive weight matrix with Bernoulli priors

The Wombling approach assumes that each pair of spatial regions i and j are conditionally dependent based on a Bernoulli distribution (Lu *and others*, 2007). Thus, $w_{ij}|p_{ij} \sim Bernoulli(p_{ij})$ and they let p_{ij} depend on known covariates for region i and j. The p_{ij} could also be fixed to a global value. By asserting that $w_{ij} = w_{ji}$, the weight matrix $\mathbf{W}$ is symmetric and binary. However, some adjacent pairs might have a weight $w_{ij} = 0$ and non adjacent pairs could have a weight of 1. To align Wombling with an adjacency structure it is possible to only estimate the weights for adjacent pairs, which effectively is a pruning of the corresponding ICAR weight matrix.

Adaptive weight matrix with $gamma(\alpha, \alpha)$ priors

Brezger *and others* (2007) proposed an adaptive model in the field of human brain mapping, which can be used for spatial areal data. Similarly to the Wombling method, this model makes the weights in $\mathbf{W}$ stochastic. However, they are no

longer limited to 0 or 1 and are instead given $Gamma(\alpha, \alpha)$ priors. Weights for non-adjacent regions are set to 0. Concretely,

$$\begin{aligned} w_{ij} &\sim Gamma(\alpha, \alpha), && i \sim j \\ w_{ij} &= 0, && \text{otherwise.} \end{aligned}$$

To ensure that $\mathbf{W}$ is symmetric, the restriction $w_{ij} = w_{ji}$ is imposed for all i, j. The Gamma prior uses a shape and rate parametrization. Note that the mean of the Gamma prior is always 1 and that changing α only affects the variance. The parameter α should be fixed to a prior belief, or assigned an informative or uninformative prior. Brezger *and others* (2007) propose to use $\alpha = 0.5$ based on the degrees of freedom in their application for the human brain.

Leroux with region specific mixing

Another approach, suggested in MacNab *and others* (2006), expands the Leroux model in Appendix A.1 to include region-specific mixing parameters ρ_i. Thus, each region can have its own mix of the ICAR and the iid Gaussian. With binary weights in $\mathbf{W}$ the conditional distributions are

$$x_i \mid \mathbf{x}_{-i}, \boldsymbol{\rho}, \tau \sim N\left(\frac{\rho_i}{\rho_i n_i + 1 - \rho_i} \sum_{j \sim i} x_j, \frac{\tau^{-1}}{\rho_i n_i + 1 - \rho_i}\right), \quad i = 1, ..., N.$$

Note that the resulting precision matrix is not symmetric. Thus, the covariance matrix is non-symmetric and the distribution is not a valid CAR distribution as it is not symmetric positive definite. Additionally, the authors noted that the posterior $\boldsymbol{\rho}$ barely moved from the prior $\boldsymbol{\rho}$, likely due to small amounts of data compared to the number of parameters MacNab *and others* (2006). A remedy for the symmetry issue was proposed in Congdon (2008). Now the contribution from an adjacent region j is scaled by its mixing parameter ρ_j, which gives the following conditional distributions:

$$x_i \mid \mathbf{x}_{-i}, \boldsymbol{\rho}, \tau \sim N\left(\frac{\rho_i}{\rho_i n_i + 1 - \rho_i} \sum_{j \sim i} \rho_j x_j, \frac{\tau^{-1}}{\rho_i n_i + 1 - \rho_i}\right), \quad i = 1, ..., N. \tag{3.1}$$

However, this approach introduces another problem, namely that the adaptive model no longer simplifies to the Leroux model for all constant ρ_i's, only when $\rho_i = 1 \forall i$.

Divide and conquer to reduce computation time

A potential issue with spatial models is that the runtime generally increases exponentially with the number of regions N. Thus, for datasets with large N the computational requirements can become a limiting factor. A seemingly simple remedy for this issue is to choose K disjoint subsets $\mathbf{I}_k$ of the set of regions $\mathcal{I} = \{1, ..., N\}$ such that $\mathcal{I} = \cup_{k=1}^{K} \mathbf{I}_k$, which is suggested in Orozco-Acosta *and others* (2021). Then, fit the model for each subregion $\mathbf{I}_k$ individually based on an assumption of independence between all the subregions. The global result then combines the results for all the subregions, reminiscent of a divide and conquer approach. Compared to the global model, Orozco-Acosta *and others* (2021) report that the disjoint model

greatly reduces the computation time when fit in parallel or sequentially, but yields a worse model fit in terms of DIC and WAIC. This reduction in the goodness of fit is due to the elimination of information sharing and strengthening across the disjoint subregions as the assumption of independence is not fully supported by the analysed data.

As the assumption of independence between the disjoint subregions $\mathbf{I}_k$ is quite restrictive, Orozco-Acosta *and others* (2021) further propose to include some order s of neighbours along the boundaries of the disjoint subregions. For example, for $s = 1$ the subregion $\mathbf{I}_k^s$ now includes all the first order neighbours of the subregion $\mathbf{I}_k$. This allows for sharing of information across the subregion borders while keeping the number of regions in each subregion relatively low compared to the global case. Similarly, $s = 2$ includes the second order neighbours and so on. Orozco-Acosta *and others* (2021) observed that $s > 2$ did not significantly improve the model criteria while increasing the runtime compared to $s = 1$ and $s = 2$. However, it could yield improvements in other applications. Note that the subregions $\mathbf{I}_k^s$ are no longer disjoint, and regions along the boundaries can now have multiple predictions from different subregions $\mathbf{I}_k^s$. It is also no longer straightforward to compute the model criteria DIC and WAIC. This was previously handled automatically in the model training, but this is not as straightforward when the regions are not disjoint. For further details on how to aggregate the predictions and model criteria see Orozco-Acosta *and others* (2021). This model, which can be described as a Bayesian scalable model, is mainly applicable when the number of regions is high and the runtimes are impractical. Although, it could also improve the full model if subregions actually prefers different levels of spatial smoothing. This is a key idea in the next model.

Subregion specific strength of smoothing

As the assumption of stationarity is not always applicable, Abdul-Fattah *and others* (2024) propose a flexible extension of the ICAR model. Similarly to the scalable model, they divide the set of regions $\mathcal{I}$ into K disjoint subsets $\mathbf{I}_k$ such that $\mathcal{I} = \cup_{k=1}^{K} \mathbf{I}_k$. Each of these subsets are assigned an individual precision parameter τ_k. For transitions in the same subset $\mathbf{I}_k$ the precision used is τ_k. For transitions involving regions from different subsets $\mathbf{I}_{k_1}$ and $\mathbf{I}_{k_2}$ the effective precision becomes $\tau = \frac{\tau_{k_1} + \tau_{k_2}}{2}$. These transitions are used for the off-diagonal elements of the weight matrix $\mathbf{W}$. Thus, $\mathbf{W}$ is defined as

$$w_{ij} = \begin{cases} \frac{\tau_{k_1} + \tau_{k_2}}{2}, & i \sim j \ \& \ i \in \mathbf{I}_{k_1}, j \in \mathbf{I}_{k_2} \\ \tau_k, & i \sim j \ \& \ i, j \in \mathbf{I}_k \\ 0, & \text{else.} \end{cases}$$

Note that k_1 and k_2 in the definition above indicates any of the given disjoint subsets, and there will be K individual precision parameters τ_k. Furthermore, the diagonal matrix $\mathbf{D}$ is defined as in Equation (2.9). The precision matrix $\mathbf{Q} = \mathbf{D} - \mathbf{W}$ and inherits the symmetric property of $\mathbf{D}$ and $\mathbf{W}$. Generally, this method accounts for spatial heterogeneity, specifically, it allows the precision τ to vary over space which allows for subregions with lower and higher spatial smoothing.

3.2 Spatial AGMRFs for the validation study

The proposed models above give an overview of the field and bring up some relevant ideas. However, as one of the primary interests of this thesis is to study neighbourhood structures, the idea is to gradually increase the flexibility from the baseline model, which will be the ICAR. This will be done with the three models introduced shortly, which all have varying flexibility for the spatial smoothing, and they all simplify to the ICAR when restrictions are added.

3.2.1 Border Weighted ICAR

Instead of the traditional ICAR with one global precision parameter τ and an assumption of spatial stationarity, Aleshin-Guendel and Wakefield (2024) propose a more flexible univariate model. They propose to split the edge set $\mathcal{E}$ in two and assign each group a common precision parameter. A similar temporal model for data with known shocks was proposed in the same article and is introduced in Appendix B.

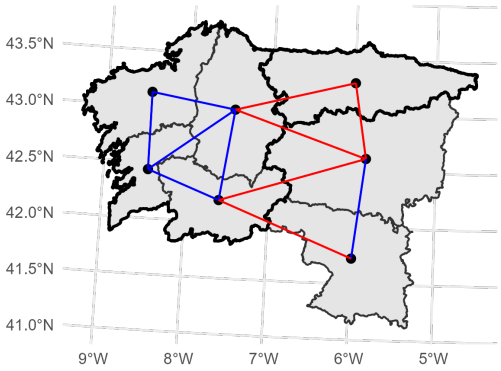
Splitting the edge set in the spatial case can be done in a few ways. For instance by splitting the regions into two groups and assigning all edges which includes the second group into its own edge group. Another option is based on different levels of spatial units, where edges in the same higher level unit are in one groups while edges across the higher level units are in the other. For the data from Spain the lower level units are provinces while the higher level units are autonomous regions. These two cases are presented in Figure 3.1. The structure of interest for this thesis is shown in the left plot. However, many other groupings of the edge set can be used, for example basing the higher and lower level spatial units on other factors like climate or geographical dividers like rivers and mountains, but these approaches will not be considered here. A different structure would only alter the definition of the weight matrix $\mathbf{W}$.

Now, lets define the Border Weighted ICAR for the structure in Figure 3.1a based on Spain. For ease of notation let A_i represent the autonomous region for a region i and the two precision parameters will be τ_1 and τ_2. Then, the chosen grouping gives the following formal definition of the weight matrix:

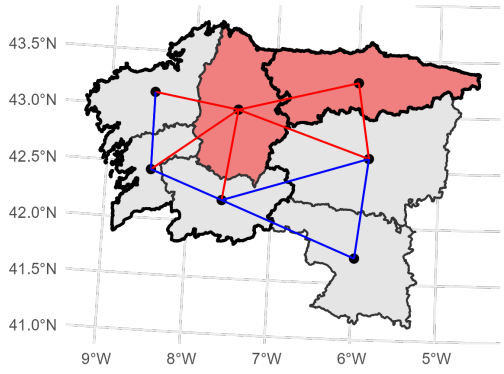
$$W_{ij} = \begin{cases} \tau_1, & A_i = A_j \ \& \ i \sim j \\ \tau_2, & A_i \neq A_j \ \& \ i \sim j \\ 0, & \text{else.} \end{cases}$$

As earlier in Equation (2.9), $\mathbf{D}$ has the row sums of $\mathbf{W}$ on the diagonal and the precision matrix $\mathbf{Q} = \mathbf{D} - \mathbf{W}$. An alternative way to construct $\mathbf{Q}$ is through individual structure matrices $\mathbf{R}^1$ and $\mathbf{R}^2$, each is responsible for one group of the edge set $\mathcal{E}$. Then,

$$R^1_{ij} = \begin{cases} 1, & A_i = A_j \ \& \ i \sim j \\ 0, & \text{else} \end{cases}$$



(a) Red edges corresponds to edges across autonomous regions while blue edges stay in an autonomous region.



(b) The two provinces in red are assumed to behave differently from the rest, and all edges including at least one of them are marked in red.

Figure 3.1: Potential structures for the Border Weighted ICAR. Province borders are grey while borders of autonomous regions are black.

$$R^2_{ij} = \begin{cases} 1, & A_i \neq A_j \ \& \ i \sim j \\ 0, & \text{else.} \end{cases}$$

Then $\mathbf{W} = \tau_1 \mathbf{R}^1 + \tau_2 \mathbf{R}^2$ and $\mathbf{D}$ and $\mathbf{Q}$ as above. Both of these methods yields the same precision matrix $\mathbf{Q}$, but the second approach makes the scaling more explicit (see Section 5.1.3) and makes the implementation in Section 5.4 easier. The full joint distribution for the model is

$$\pi(\mathbf{x} \mid \tau_1, \tau_2) = (2\pi)^{-(N-1)/2} (|\mathbf{Q}|^*)^{1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{x}\right),$$

the density for an IGMRF as in Equation (2.8). The number of regions is still N. Equivalently, $\mathbf{x} \mid \tau_1, \tau_2 \sim N(\mathbf{0}, \mathbf{Q}^{-1})$. This model will be referred to as the Border Weighted ICAR (BW-ICAR) from the focus on borders and the similarities with the ICAR from the first-order neighbourhood structure. The resulting structure for the whole of Spain can be seen in Figure 3.2. Any priors for a precision τ can be chosen for τ_1 and τ_2, but the convention is to use the same prior for both. The authors recommend using a penalized complexity prior, specifically $\tau_1, \tau_2 \sim PC(1, 0.01)$ [Chapter 4](Aleshin-Guendel and Wakefield, 2024).

3.2.2 Region Weighted ICAR

The next step up in flexibility was proposed by Corpas-Burgos and Martinez-Beneito (2020), and is quite different from the previous model. Namely, instead of dividing the edges into two sub-groups, they use the adjacency structure from the ICAR and introduce a region specific weight $c_i \in (0, \infty)$ for $i = 1, ..., N$, so $\mathbf{c} = (c_1, ..., c_N)^T$.

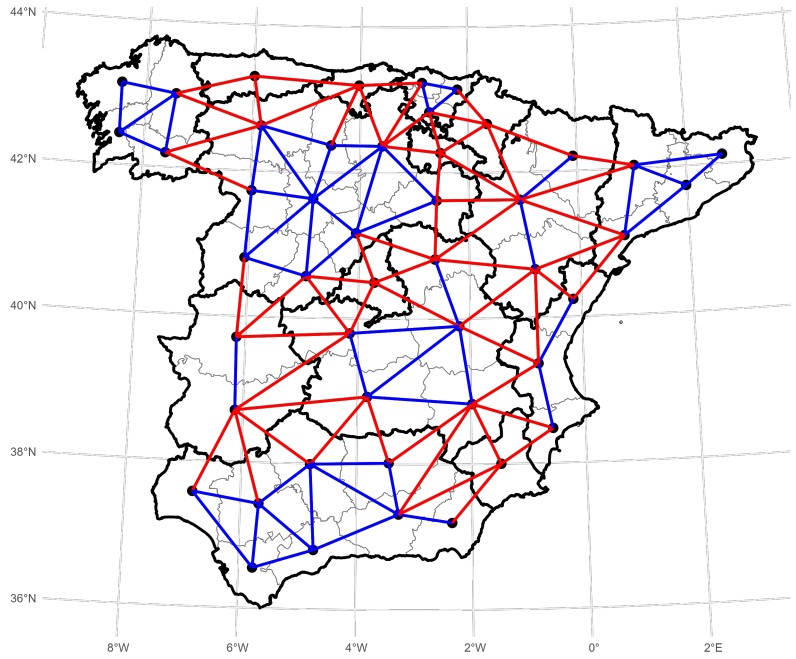


Figure 3.2: The blue edges represents neighbours in the same autonomous region while red edges represents neighbours in different autonomous regions.

Recall that the ICAR can be defined as

$$\mathbf{x} \mid \tau \sim N(\mathbf{0}, \tau^{-1}(\mathbf{D} - \mathbf{W})^{-})$$

with $\mathbf{D} = diag(n_1, ..., n_N)$ and

$$W_{ij} = \begin{cases} 1, & i \sim j \\ 0, & \text{else.} \end{cases}$$

With $\mathbf{W}$ defined as for the ICAR above, the region specific weights in $\mathbf{c}$ are incorporated as $\mathbf{W}^*(\mathbf{c}) = diag(\mathbf{c})^{1/2}\mathbf{W}diag(\mathbf{c})^{1/2}$ and $\mathbf{D}^* = diag(w^*_{1+}, ..., w^*_{N+})$ with $w^*_{i+} = \sum_{j \sim i} w^*_{ij}$. This means that the zero elements in $\mathbf{W}$ are kept at 0 in $\mathbf{W}^*$ as well. However, the non-zero weights are not the same. In $\mathbf{W}^*$ the weights can be greater and smaller than 1, and they are all governed by the region specific weights in $\mathbf{c}$. Specifically, the weights are

$$w^*_{ij}(\mathbf{c}) = \begin{cases} \sqrt{c_i c_j}, & i \sim j \\ 0, & \text{else,} \end{cases}$$

and the random effect is then

$$\mathbf{x} \mid \tau, \mathbf{c} \sim N(\mathbf{0}, \tau^{-1}(\mathbf{D}^* - \mathbf{W}^*(\mathbf{c}))^{-}). \tag{3.2}$$

The regions with a large c_i will both exert a greater influence on its neighbours and be more influenced by its neighbours in return, compared to regions with a

lower c_i. Thus, c_i can be interpreted as how spatially connected each region is to its neighbours where a $c_i = 1$ corresponds to the special dependence of the traditional ICAR. Off course, these considerations also depend on the precision τ for a given model. The interpretation of a single c_i is also inherently connected to the c_j's for the surrounding regions, as it is ultimately the product of them that is interesting.

Lastly, both τ and $\mathbf{c}$ must be assigned hyperpriors. The hyperprior for τ is defined through the standard deviation $\sigma \sim U(0, 5)$ as in Section 2.3, thus $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$. For the region specific weights in $\mathbf{c}$, two layers are used. First,

$$c_i|\alpha \sim Gamma(\alpha, \alpha), \quad i = 1, ..., N,$$

and $\alpha = \frac{1}{\sigma_\alpha^2}$ with $\sigma_\alpha \sim U(0, 5)$. The priors are based on the authors recommendation for the model. Since this model introduces many new parameters it is no longer appropriate with a univariate model, and the authors recommend using the model for multivariate modelling (Corpas-Burgos and Martinez-Beneito, 2020). This is due to the fact that the number of parameters exceeds the number of observed datapoints, which is problematic as the prior will have a large effect on the posterior inference. They also mention problems related to interpretability of the results [Chapter 3.2](Corpas-Burgos and Martinez-Beneito, 2020). For the multivariate case one precision τ is included for each disease. For a single disease the model uses one precision parameter τ and N region specific weights $\mathbf{c}$. As this model focuses on region specific weights to modify the ICAR weight matrix, it will be called the Region Weighted ICAR (RW-ICAR).

3.2.3 Edge Weighted ICAR

The third and most flexible model is inspired by Riddervold (2024), which introduced a spatio-temporal model. The spatial random effect is a more flexible extension of the standard ICAR, which works on data that breaks the stationarity assumption. Specifically, the model, called the adaptive ICAR from now on, has a unique precision τ_{ij} for each pair of neighbours i and j. In other words, the difference between all pairs of neighbours are distributed as

$$x_i - x_j|\tau_{ij} \sim N(0, \tau_{ij}^{-1}).$$

In total this gives the following structure matrix $\mathbf{Q}$:

$$\mathbf{Q}_{ij} = \begin{cases} \sum_{k|i\sim k} \tau_{ik}, & \text{if } i = j \\ -\tau_{ij}, & \text{if } i \sim j \\ 0, & \text{else.} \end{cases}$$

The structure matrix can also be used to express the joint distribution of $\mathbf{x}$ as

$$\pi(\mathbf{x}|\{\tau_{ij}\}_{i\sim j}) = (2\pi)^{-(N-1)/2}(|\mathbf{Q}|^*)^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\right) \tag{3.3}$$

where $|\mathbf{Q}|^*$ denotes the generalized determinant as defined in Section 2.5. Let $\mathbf{x} \sim IGMRF(\mathbf{0}, \mathbf{Q})$ denote that $\mathbf{x}$ follows the joint distribution in Equation (3.3).

As the adaptive ICAR introduces many new parameters, specifically $|\mathcal{E}|$, the number of edges in the edge set, one remedy is to introduce temporal data to loan strength across time. In Riddervold (2024) a first order autoregressive process is applied to the field $\mathbf{x}$. Assume that $\mathbf{x}_1 \sim IGMRF(\mathbf{0}, \mathbf{Q})$, which means that the random effect for timepoint 1 is an adaptive ICAR. Then, for T timepoints the relationships between them are as follows:

$$\mathbf{x}_t = \rho \mathbf{x}_{t-1} + \sqrt{1-\rho^2}\mathbf{v}_t, \quad t = 2, .., T,$$

with the autoregressive coefficient $-1 < \rho < 1$ and some noise $\mathbf{v}_t \sim IGMRF(\mathbf{0}, \mathbf{Q})$. A large or small parameter ρ, close to either 1 or -1, indicates little noise from year to year. However, a negative ρ corresponds to a negative correlation between years, which in most scenarios is not the expected correlation. Thus, ρ is generally limited to $0 \leq \rho < 1$. This model will be reformulated to a multivariate disease setting to loan strength across the diseases.

In this thesis the focus is not on spatio-temporal data, but instead on spatial data, both univariate and multivariate. As there is no particular structure in the data there is no reason to introduce some autoregressive component. The multivariate analysis instead uses the same weight matrix for the different response variables $\mathbf{y}$. This corresponds to fixing ρ to 0 in the temporal model above, which means that

$$\mathbf{x}_t \sim IGMRF(\mathbf{0}, \mathbf{Q}), \quad t = 1, ..., T.$$

Thus, the only sharing across time is through the precision matrix $\mathbf{Q}$ and its multitude of precision parameters. As previously mentioned, the data of interest is spatial data for multiple causes of mortality in Spain without a temporal dimension. Thus, the temporal model with $\rho = 0$ can be rewritten to the multivariate spatial scenario as:

$$\mathbf{x}_m \sim IGMRF(\mathbf{0}, \mathbf{Q}), \text{ for } m \in \mathcal{M}$$

where each $m \in \mathcal{M}$ represents a single univariate disease from the possible diseases in $\mathcal{M}$. Even though this is a valid definition of the model, an issue is that everything concerning the precision matrix $\mathbf{Q}$ is shared across all the diseases. As this includes the degree of spatial smoothing, a reparametrization is done to keep the same structure across the diseases, but which allows the scale to vary. This is done by introducing a precision τ_m for the scale for each disease. Thus,

$$\mathbf{x}_m \sim IGMRF(\mathbf{0}, \tau_m \mathbf{Q}), \text{ for } m \in \mathcal{M}.$$

Additionally, the priors are chosen to align with the priors for the RW-ICAR. Thus, the disease specific precisions τ_m are given the priors $\frac{1}{\sqrt{\tau_m}} \sim U(0,5) \forall m \in \mathcal{M}$ and the precisions for the edges $\tau_{ij} \sim gamma(\alpha, \alpha)$ with $\alpha = \sigma_\alpha^{-2}$ with $\sigma_\alpha \sim U(0,5)$. Thus, the prior mean for all the τ_{ij}'s is 1, and when they are equal to 1 the structured random effect simplifies to the ICAR. Thus, each τ_{ij} can be interpreted relative to 1, which serves as the base model. As this model has a parameter for each edge, it will be denoted the Edge Weighted ICAR (EW-ICAR).

In the univariate case this model has $1 + |\mathcal{E}|$ parameters. Overall, the increasing number of parameters of the models in a univariate setting is presented in Table 3.1. Now that all the models are introduced, the next step is to introduce the application and the data for the disease mapping from Spain.

Model	**Theoretical # parameters**
ICAR	1
BW-ICAR	2
RW-ICAR	$N + 1$
EW-ICAR	$\lvert\mathcal{E}\rvert + 1$

Table 3.1: Number of parameters for the different models in a univariate setting. N is the number of regions and $\mathcal{E}$ is the edge set.

4 Disease mapping

Now that Bayesian hierarchical models, GMRFs and AGMRFs have been introduced, lets put them all together in a disease mapping context. This is a typical application for the aforementioned models and the general setup includes N discrete spatial units, for example municipalities or counties, with the goal of modelling the risk of a disease or mortality in all the spatial units. As the data is counts of incidences per region, a Poisson likelihood and a log-link for the relative risk is a typical setup, for example

$$\begin{aligned} y_i \mid \eta_i &\sim Poisson(E_i \eta_i), \quad i = 1, ..., N \\ \log(\eta_i) &= \mu + \mathbf{z}_i^T \boldsymbol{\beta} + x_i, \quad i = 1, ..., N \end{aligned} \tag{4.1}$$

where E_i is the expected number of cases for region i and y_i is the number of observed cases. The expected cases are typically calculated from population data so that the cases for a specific age-group, say under 5 years old, on the national level is distributed to the region level proportionally to their population in that age group. This is essentially a preprocessing of the data, and will be further explained in section 4.2. Additionally, it can be considered a form of scaling as a risk $\eta_i = 1.1$ indicates that region i had 10% more cases than expected, and this holds regardless of the number of cases y_i in the region. Thus, the risk $\boldsymbol{\eta}$ can be interpreted as a relative scaling of the expected cases. Note that the risk often is denoted as R_i instead of η_i. Another common approach is to use the population at risk as an offset instead of the expected cases. In this case the risk η_i models the proportion of the population at risk that was affected.

The risk factor η_i for region i depends on an intercept μ, some covariates $\boldsymbol{z}_i$ with fixed effects $\boldsymbol{\beta}$ and a random effect x_i through a log-link. The main difference between disease models is often contained in the random effect $\mathbf{x}$. However, it is also possible to exchange the Poisson likelihood with a negative Binomial or another distribution, or use a different link function.

4.1 Multivariate disease mapping

Another major difference between disease mapping models is whether they are univariate or multivariate. The introduction above focused on the univariate case, but in some situations the multivariate approach is preferred. If a model with many parameters is used in the latent layer for the log-risk, for example the EW-ICAR, there could be insufficient data with a single disease (Corpas-Burgos and Martinez-Beneito, 2020). Thus, if several diseases have similar underlying risk factors, it is sensible to model them jointly (Knorr-Held and Best, 2001). Introducing additional diseases would make these models more data driven, and allow the posterior to differ more from the priors, if the data supports that. It could also help if there are missing data or low counts, and thus higher variability, where a multivariate approach could loan strength from the other diseases (Assunção and Castro, 2004). However, it should be noted that not all diseases are inherently correlated, and thus the diseases included in a multivariate setting should ideally have something in common. For a more in-depth introduction to disease mapping see Lawson (2013).

Multivariate disease mapping models are trained jointly on multiple diseases. The assumption is that the diseases have similar risk patterns, and thus the diseases should be chosen with care. The scale of the counts could vary substantially from disease to disease, and because of the scaling from the expected counts $\mathbf{E}$, see Equation (4.1), this is no issue. The idea is to create a Bayesian hierarchical model where parts of the model is shared and parts are disease specific. Rewriting the univariate formulation from Equation (4.1) gives:

$$\begin{aligned} y_{im} \mid \eta_{im} &\sim Poisson(E_{im}\eta_{im}), \qquad i = 1, ..., N \ \& \ m = 1, ..., M \\ \log(\eta_{im}) &= \mu_m + \mathbf{z}_i^T \boldsymbol{\beta}_m + x_{im}, \qquad i = 1, ..., N \ \& \ m = 1, ..., M \end{aligned} \tag{4.2}$$

for M different diseases. Now the intercept μ_m and the weights $\boldsymbol{\beta}_m$ are specific for each disease and the observed cases $\mathbf{y}$, expected cases $\mathbf{E}$ and the risk $\boldsymbol{\eta}$ follow the conventions of the univariate case. For the random effect $\boldsymbol{x}$ there are multiple options, but this thesis will follow the conventions from Corpas-Burgos and Martinez-Beneito (2020). Namely that the structure of the random effect is the same for all the diseases, but at a different scale. This means that the strength of spatial smoothing can vary between the diseases, but still with a similar structure. For a GMRF, this could be summarised with the following priors:

$$\mathbf{x}_m \sim N(\mathbf{0}, \tau_m^{-1}\mathbf{Q}^-), \quad m = 1, ..., M,$$

where the precision matrix $\mathbf{Q}$ is shared across all diseases and the precision τ_m is disease specific, but could share the same prior, for instance one from Section 2.3. This means that all the $\mathbf{x}_m$'s could have different degrees of spatial smoothing dependent on the fitted value of τ_m. For further reading on multivariate disease mapping see Chapter 10 in Lawson (2013).

4.2 Male mortality data for multiple causes from Spain 2019

Most of the theoretical background has been covered, so lets introduce the dataset of interest as well as some preprocessing of the data. The dataset of interest in this article concerns male mortality data from Spain at the province level from 2019, retrieved from Instituto Nacional de Estadística (2020a). The year 2019 was chosen to avoid any impact from the COVID-19 pandemic and to focus on a more or less average year. The mortality date is available from https://ine.es/dynt3/inebase/en/index.htm?padre=7924 by navigating through "Basic cause of death", then "Provincial results" and finally "Deaths by causes (reduced list), sex and province of residence" (Instituto Nacional de Estadística, 2020a). The dataset covers a multitude of mortality causes, which will also be referred to as diseases. An overview of the different diseases are presented in Table 4.1. A subset of particular interest is the largest grouping of diseases called tumours, which covers 33 different types of cancers. This dataset will supply the disease counts $\mathbf{y}$ used in Equations (4.1) and (4.2), but additional information is needed to calculate the expected number of cases. This data is also from Instituto Nacional de Estadística. One data set is from the same link as previously, https://ine.es/dynt3/inebase/en/index.htm?padre=7924, and then by navigating through "Basic cause of death" followed by "National results" and then "Deaths by causes (reduced list), sex and age" (Instituto Nacional

Overarching disease groups	# diseases	# cases
Infectious and parasitic diseases	8	3061
Tumours	33	67951
Diseases of blood and haematopoietic, and certain disorders involving the immune mechanism	2	878
Endocrine, nutritional and metabolic diseases	2	5868
Mental and behaviour disorders	4	7875
Diseases of the nervous system and sensory organs	3	10288
Diseases of the circulatory system	9	54511
Diseases of the respiratory system	6	26310
Diseases of the digestive system	5	11446
Diseases of the skin and subcutaneous tissue	1	629
Diseases of the osteomuscular system and of the conjunctive tissue	3	1730
Diseases of the genitourinary system	4	6235
Pregnancy, childbirth and puerperium	1	0
Affections originated in in the perinatal period	1	331
Congenital malformations, deformations and chromosomal abnormalities	3	412
Symptoms, signs and abnormal clinical and laboratory findings, not classified elsewhere	4	4922
External causes of mortality	13	10236
Total	102	212683

Table 4.1: Overview of the dataset for male mortality data from Spain 2019 at national level and aggregated by groups of disease causes.

de Estadística, 2020a). This dataset contains disease counts on a national level with the population divided into age groups spanning five years. For instance 0-4, 5-9, 10-14 and so on up to above 94 years old. The set of age groups will be referred to as $\mathcal{A}$. The last necessary data is simply the population counts for these age groups for each province, which are available from https://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p08/l0/&file=03002.px&L=0 (Instituto Nacional de Estadística, 2020b). Then, the expected counts can be expressed as

$$E_{im} = \sum_{a \in \mathcal{A}} \frac{P_{ai} D_{am}}{P_a}, \quad i = 1, ..., N \ \& \ m = 1, ..., M. \tag{4.3}$$

Here P_{ai} represents the number of people in age group a which lives in province i, P_a is the total number of people in age group a nationally, so $P_a = \sum_{i=1}^{N} P_{ai}$, and $D_{am} = \sum_{i=1}^{N} y_{am}$ is the national cases for age group a and disease m. The result E_{im} is then the expected disease count for disease m in province i, assuming the likelihood of the disease is spread proportionally across the whole of Spain with respect to the population in each age group. This is emphasized as $\frac{D_{am}}{P_a}$ in Equation (4.3) represents the proportion of age group a who died from disease m at a national level, before it is scaled by the population in age group a in province i, i.e. P_{ai}. Then this is summed up across all age groups $a \in \mathcal{A}$.

That sums up the strictly necessary preprocessing of the data. However, as this thesis focuses on first-order neighbourhood structures, certain provinces are not of interest. Specifically, all provinces not directly connected to mainland Spain will be removed. This includes the island provinces Islas Baleares, Las Palmas and Santa Cruz de Tenerife as well as Ceuta and Melilla since they are not connected to mainland Spain, situated at the northern coast of Africa. This leaves 47 provinces. Thus, the area of interest is the mainland of Spain as shown in Figure 4.1. The five provinces not connected to the mainland are the five that have been removed.

Figure 4.1: A map of Spain divided into provinces with the name of each province. The image is from Fernández and S. (2010).

Some preprocessing of the diseases is also advised. As can be seen in Table 4.1, some of the diseases are rather rare, one even has 0 cases. Thus, "Pregnancy, childbirth and puerperium" is not particularly relevant for males. The same goes for specific diseases in the other groups, for example uterine cancer. In order to ensure the diseases are prevalent enough to include in the analysis in Section 5, an arbitrary limit of an average of two cases per province, i.e. $2 \cdot 47 = 94$, was enforced. This reduced the total number of diseases to 86, and the subset tumours now contains 26 different cancers. The reason the subset tumours have been singled out is because it is a large subset, and it seems likely that cancers have a similar risk pattern across Spain. It is less clear that for example lung cancer and homicides have a similar risk pattern. Thus, it could be interesting to compare multivariate models trained on only cancer data versus all the disease data.

5 Validation study

As the general models of interest were defined in Section 3.2 and the dataset of interest was introduced in Section 4.2, the next step is to define the validation study. This involves training, validating and comparing the models. The main interest is in whether the standard structure matrix associated with the ICAR is the best option, or if more flexible neighbourhood structures can perform better. The comparison will use univariate diseases and they will be evaluated in terms of DIC, WAIC and LS (see Section 5.3). The overall procedure will be split in two parts, training the models and then validating on individual diseases.

5.1 Training the models

Lets start by defining the Bayesian hierarchical models for the disease mapping application. The different AGMRFs, i.e. BW-ICAR, RW-ICAR and EW-ICAR, will be used in the latent layer. The two univariate models, namely the ICAR and the BW-ICAR, will build on Equation (4.1), where they will replace the general x_i. No covariates will be used. As the ICAR does not need any training, it will be revisited in Section 5.2.

5.1.1 Model setup

The full Bayesian hierarchical model for the BW-ICAR is

$$\begin{aligned}
&\text{1. Likelihood:} && y_{im} \mid \eta_{im} \sim \text{Poisson}(E_{im}\eta_{im}), \text{ for } i \in \mathcal{I} \text{ and } m \in \mathcal{M} \\
&\text{2. Latent layer:} && \log(\eta_{im}) = \mu_m + x_{im}, \text{ for } i \in \mathcal{I} \text{ and } m \in \mathcal{M} \\
& && \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, (\tau_{m1}\mathbf{R}^1 + \tau_{m2}\mathbf{R}^2)^-), \text{ for } m \in \mathcal{M} \\
& && \mu_m \sim \mathcal{U}(-\infty, \infty) \propto 1, \quad \text{for } m \in \mathcal{M} \\
&\text{3. Hyperpriors:} && \tau_{m1}, \tau_{m2} \sim PC(1, 0.01), \quad \text{for } m \in \mathcal{M}
\end{aligned} \tag{5.1}$$

for diseases $m \in \mathcal{M}$, provinces $i \in \mathcal{I}$ and with $\mathbf{R}^1$ and $\mathbf{R}^2$ as defined in Section 3.2.1. Recall that the superscript $()^-$ represents the generalized inverse. The penalized complexity prior was chosen because the authors recommend the PC priors (see Section 2.3), and specifically with the given parameters [Chapter 4](Aleshin-Guendel and Wakefield, 2024). A small sensitivity analysis was carried out in Appendix C, which showed that the prior choice has some effect. For instance $gamma(0, 0.00005)$ performed bad for DIC and WAIC, and it was closer in terms of LS. As the random effect $\mathbf{x}_m$ is an IGMRF, a sum-to-zero constraint is applied to each $\mathbf{x}_m$ through the INLA definitions to make the intercepts μ_m interpretable, more on this in Section 5.4. Note that this model has no sharing of information across the diseases and is a univariate disease model for the individual diseases.

For the two multivariate models, the RW-ICAR and the EW-ICAR, their Bayesian hierarchical models are based on Equation (4.2). For a set of diseases $\mathcal{M}$ and a set of regions $\mathcal{I}$, which are the provinces of mainland Spain as described in Section 4.2,

the full Bayesian hierarchical model for the RW-ICAR is

$$
\begin{aligned}
&\text{1. Likelihood:} && y_{im} \mid \eta_{im} \sim \text{Poisson}(E_{im}\eta_{im}),\ \text{for } i \in \mathcal{I} \text{ and } m \in \mathcal{M} \\
&\text{2. Latent layer:} && \log(\eta_{im}) = \mu_m + x_{im},\ \text{for } i \in \mathcal{I} \text{ and } m \in \mathcal{M} \\
& && \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \tau_m^{-1}(\mathbf{D}^* - \mathbf{W}^*(\mathbf{c}))^-),\ \text{for } m \in \mathcal{M} \\
& && \mu_m \sim \mathcal{U}(-\infty, \infty) \propto 1, \quad \text{for } m \in \mathcal{M} \\
&\text{3. Hyperpriors:} && \frac{1}{\sqrt{\tau_m}} \sim U(0,5),\ \text{for } m \in \mathcal{M} \\
& && \mathbf{c}_i | \alpha \sim gamma(\alpha, \alpha),\ \text{for } i \in \mathcal{I} \\
& && \alpha = \frac{1}{\sigma_\alpha^2} \text{ with } \sigma_\alpha \sim U(0,5)
\end{aligned}
\tag{5.2}
$$

with $\mathbf{D}^*$ and $\mathbf{W}^*(\mathbf{c})$ defined as in Section 3.2.2. Note the the restriction $c_i > 0.001$ is enforced to increase numerical stability, more on this in Section 5.4. The weights $\mathbf{c}$ will be shared across all the diseases in the set $\mathcal{M}$ while the amount of smoothing determined by the precision parameter τ_m can vary between diseases. This means that all the diseases will have the same smoothing structure from the $\mathbf{c}$, but the degree of smoothing is disease specific. The prior on $\mathbf{x}_m$ coincides with Equation (3.2) and the priors on the intercepts $\boldsymbol{\mu}$ are uniform over the real line, which essentially means that all values of μ_m are equally likely, and hence proportional to 1. For $\boldsymbol{\tau} = (\tau_1, ..., \tau_M)^T$ all the precision parameters have the same identical hyperprior. Lastly, as $\mathbf{x}_m$ is an IGMRF, a soft sum-to-zero constraint is imposed of the form

$$
\sum \mathbf{x}_m \sim N(0, 0.1) \quad m \in \mathcal{M}, \tag{5.3}
$$

which wants the sum of $\mathbf{x}_m$ for each of the diseases to be close to zero. A variance of 0.1 was chosen as it was used in the implementation of the model in Corpas-Burgos and Martinez-Beneito (2020), more on this in Section 5.4. The advantage of a soft sum-to-zero constraint versus a hard sum-to-zero constraint is increased robustness, and faster convergence for larger N (Morris, 2025).

For the EW-ICAR only the priors in the latent layer and the hyperpriors in the third layer will change from Equation (5.2). Thus, the Bayesian hierarchical model is

$$
\begin{aligned}
&\text{1. Likelihood:} && y_{im} \mid \eta_{im} \sim \text{Poisson}(E_{im}\eta_{im}),\ \text{for } i \in \mathcal{I} \text{ and } m \in \mathcal{M} \\
&\text{2. Latent layer:} && \log(\eta_{im}) = \mu_m + x_{im},\ \text{for } i \in \mathcal{I} \text{ and } m \in \mathcal{M} \\
& && \mathbf{x}_m \sim N(\mathbf{0}, (\tau_m \mathbf{Q})^-),\ \text{for } m \in \mathcal{M} \\
& && \mu_m \sim \mathcal{U}(-\infty, \infty) \propto 1,\ \text{for } m \in \mathcal{M} \\
&\text{3. Hyperpriors:} && \frac{1}{\sqrt{\tau_m}} \sim U(0,5) \text{ for } m \in \mathcal{M} \\
& && \tau_{ij} \sim gamma(\alpha, \alpha) \text{ for } i \sim j \ \&\ i, j \in \mathcal{I} \\
& && \alpha = \sigma_\alpha^{-2} \text{ with } \sigma_\alpha \sim U(0,5).
\end{aligned}
\tag{5.4}
$$

The precision matrix $\mathbf{Q}$ is defined from the edge precisions τ_{ij} as in Section 3.2.3. Similarly to the c_i's for the RW-ICAR, the edge precisions τ_{ij} are limited to values greater than 0.001, again to increase numerical stability. As $\mathbf{x}_m$ is an IGMRF, a

sum-to-zero constraint is imposed as for the RW-ICAR as in Equation (5.3). The priors for the disease specific mean is the same as in the RW-ICAR in Equation (5.2). Additionally, the prior for each τ_{ij} is the same as the priors for the region weights $\mathbf{c}$ as well. This is because they both simplify to the ICAR if they are equal to 1.

5.1.2 Design

Now that the Bayesian hierarchical models have been defined, lets continue with the training procedure. For the ICAR and the BW-ICAR this will be done individually for each disease as these models are univariate. The ICAR will actually skip the training stage as its structure matrix is rigid, and was defined in Section 2.5. For the two multivariate models a few different models will be trained to assess the effect of including a given subset of the 86 diseases. Specifically, n diseases will be chosen at random with values of n chosen from $\{10, 20, 50, 86\}$. Additionally, as a large part of the diseases are different types of cancer, both multivariate methods are also trained on a subset of the diseases where only the 26 cancers are included. These different training sets are designed to investigate the effect of the number of diseases used in the training, as well as to compare the effect of including diseases with no apparent similarities. For instance, whether the validation for cancers will be improved by including the non-cancers in the training or not. The only relevant result from the training stage is the fitted neighbourhood structure, or in other words the resulting precision matrix $\mathbf{Q}$.

In regard to randomly choosing diseases based on n, for $n = 10$ the diseases are chosen randomly. Then for $n = 20$, 10 additional diseases are randomly chosen as well as the 10 diseases for $n = 10$. Similarly, $n = 50$ adds 30 new diseases. This is done to avoid situations where $n = 20$ randomly chooses diseases which do not represent the overall diseases well while $n = 10$ chooses 10 which do represents the diseases well, and hence outperforms $n = 20$. This is more likely to occur if the diseases are sampled randomly for each n without building on the previous n, but that is still a valid option to use. Note that this sampling is based on a given seed in R and that the subsets would change if another seed is used. When it is relevant, the procedure should be repeated for a different seed to investigate if the results are a general result for the models or simply a facet of the sampled subsets of diseases.

For ease of notation all the trained models will be given shortened names, for instance when named in a plot. For instance "BW" represents the BW-ICAR, "RW_10" represents the RW-ICAR with 10 diseases and "EW_20" represents the EW-ICAR with 20 diseases and so on. Additionally, the models trained only on cancer data will be denoted as "RW_C" and "EW_C". The ICAR model will simply be denoted as "ICAR". The training will be done by either INLA or WinBUGS. Even though all the models are GMRFs, and thus compatible with LGMs, the RW-ICAR and the EW-ICAR are not compatible with INLA. Mainly because the number of parameters is very high, which breaks one of the assumptions of INLA. Namely that the number of hyperparameters is low, generally under 20 Rue *and others* (2016). The number of parameters for each model in the latent layer in the univariate case can be seen in Table 5.1. Lastly, at the primary interest of this thesis is the comparison of

Model	Theoretical # parameters	# parameters for Spain
ICAR	1	1
BW-ICAR	2	2
RW-ICAR	$N+1$	$47+1$
EW-ICAR	$\lvert\mathcal{E}\rvert+1$	$111+1$

Table 5.1: Number of parameters for the different precision matrices in a univariate setting. N is the number of regions and $\mathcal{E}$ is the edge set.

different neighbourhood structures for a single disease, the model fit when training the multivariate models will not be considered. In other applications where the interest is multivariate or joint disease mapping, these properties are more relevant.

5.1.3 Estimating the neighbourhood structure

After the models have been trained, the most interesting part is the structure of spatial smoothing in the precision matrices. This neighbourhood structure is the only part of the training which is passed on to the validation stage. However, for the priors to carry the same information in the validation stage, some scaling of the precision matrices must be done. Consider two precision matrices $\tau\mathbf{Q}$ and $2\tau\mathbf{Q}$. Then a prior on τ has a different effect in the two cases. In the disease mapping application, the matrices can vary in scale, as in the simplistic example, but also in the specific entries of the matrices as a facet of the adaptive nature of the spatial smoothing. Thus, some scaling is necessary to ensure the transferability of priors between the different models and that they are assigned the same prior information (Sørbye and Rue, 2014). If this is not done, comparing the models would be unfair.

This scaling will be based on the geometric variance as introduced in Sørbye and Rue (2014). This is the geometric mean of the marginal variances associated with a precision matrix, and is defined as

$$\sigma^2_{GV}(\mathbf{x}) = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log\left([\mathbf{Q}^-]_{ii}\right)\right), \tag{5.5}$$

which assumes that $\mathbf{x} = (x_1, ..., x_N)^T \sim N(\mathbf{0}, \mathbf{Q}^-)$. As this property is tied to the variance, and hence the precision, the idea is that $\sigma^2_{GV}(\mathbf{x})$ should be equal to 1 for any $\mathbf{Q}$, and this will be accomplished by scaling the precision matrix $\mathbf{Q}$. Specifically, calculate $\sigma^2_{GV}(\mathbf{x})$ for a precision matrix $\mathbf{Q}$, and denote the result as $\sigma^2_{ref}(\mathbf{x})$. Then, $\mathbf{x}^* \sim N(\mathbf{0}, (\mathbf{Q}^*)^-)$ with $\mathbf{Q}^* = \sigma^2_{ref}(\mathbf{x})\mathbf{Q}$. Then $\sigma^2_{GV}(\mathbf{x}^*) = 1$. The last part of the idea is to combine the scaled precision matrix $\mathbf{Q}^*$ with a precision parameter τ such that

$$\mathbf{x} \sim N(\mathbf{0}, (\tau\mathbf{Q}^*)^-). \tag{5.6}$$

Then the prior placed on the precision parameter τ carries the same meaning regardless of the original matrix $\mathbf{Q}$, within reason. A similar scaling is introduced alongside the BYM2 model in Appendix A.1.

Lets expand on this idea with an example for the BW-ICAR. Specifically, the geometric variance is calculated for the matrix $\mathbf{Q} = \mathbf{R}^1 + \mathbf{R}^2$ and denoted as σ^2_{GV}, with $\mathbf{R}^1$ and $\mathbf{R}^2$ as defined in Section 3.2.1. Then $\mathbf{R}^k_* = \sigma^2_{GV}\mathbf{R}^k$ for $k = 1, 2$. The scaling is represented with a subscript in this example as previous notation for the 2 matrices $\mathbf{R}^1$ and $\mathbf{R}^2$ complicated the notation. Then $\mathbf{Q}_* = \mathbf{D}_* - \mathbf{W}_*$ with $\mathbf{W}_* = \tau_1\mathbf{R}^1_* + \tau_2\mathbf{R}^2_*$ and $\mathbf{D}_*$ defined by $\mathbf{W}_*$ as in Equation (2.9). The full joint distribution for the model is

$$\pi(\mathbf{x} \mid \tau_1, \tau_2) = (2\pi)^{-(N-1)/2}(|\mathbf{Q}_*|^*)^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}_*\mathbf{x}\right),$$

the density for an IGMRF as in Equation (2.8). Note that the superimposed $*$ references the generalized determinant while the superscript $*$ indicates that the matrix is scaled. Equivalently, $\mathbf{x} \mid \tau_1, \tau_2 \sim N(\mathbf{0}, \mathbf{Q}^-_*)$. This example was slightly more involved with 2 sub-matrices that makes up $\mathbf{Q}$, which is mainly necessary for ease of implementation, which will be covered in Section 5.4. This scaling will be used when fitting the ICAR and the BW-ICAR models, which alters Equation (5.4) to the scaled versions of $\mathbf{R}^1$ and $\mathbf{R}^2$ in the latent layer. There is not used any scaling when training the multivariate models, RW-ICAR and EW-ICAR. As the main interest lies in the neighbourhood strictures this is okay for the training stage. However, it still affects the transferability of priors.

All the resulting precision matrices will be scaled as above and reused in the validation study. Lets illustrate this with an example. Consider the EW-ICAR trained on $n = 86$ diseases. The output returned from WinBUGS through the R library R2WinBUGS contains the relevant parts from the MCMC simulations, Including the simulations for all the edge specific weights τ_{ij}. Creating the precision matrix $\mathbf{Q}_{EW\_86}$ is done by placing the posterior means for all the edges at their assigned places, i.e. $W_{ij} = W_{ji} = \tau_{ij}$, which defines a weight matrix $\mathbf{W}$. Then $\mathbf{D}$ has the row-sums of $\mathbf{W}$ on the diagonal and $\mathbf{Q}_{EW\_86} = \mathbf{D} - \mathbf{W}$. Thus, the scaled precision matrix $\mathbf{Q}^*_{EW\_86} = \sigma^2_{GV}(\mathbf{x})\mathbf{Q}_{EW\_86}$ with $\mathbf{x} \sim N(\mathbf{0}, \mathbf{Q}^-_{\mathbf{EW\_86}})$. For any of the RW-ICARs the edge precisions are calculated as $\tau_{ij} = \sqrt{c_i c_j}$ for the posterior mean estimates of the weights c_i, and the precision matrices are constructed as in the example for the EW-ICAR. The use of scaling can be summarised in Figure 5.2. Note that scaling in the training and validation is done to ensure the interpretability of the priors and that the priors for different models carry the same meaning.

	ICAR	BW	RW_X	EW_X
Is the model trained?	No	Yes	Yes	Yes
Is the model scaled in the training?	—	Yes	No	No
Is the model scaled in the validation?	Yes	Yes	Yes	Yes

Table 5.2: Summary of where scaling is applied. The _X for RW and EW indicates any of the five possible subsets of diseases as they all follow the same conventions. Namely $n \in \{10, 20, 50, 86\}$ and only cancers.

5.2 Validating the models

The scaled precision matrices from the training are passed to the validation stage, denoted as $\mathbf{Q}_k^*$. The subscript k will indicate the fitted model that yielded the given precision matrix. An IGMRF with a single precision parameter τ_{mk} and the scaled precision matrix $\mathbf{Q}_k^*$ will be fitted to each disease $m \in \mathcal{M}$ with the following prior for the GMRF:

$$\pi(\mathbf{x}_{mk}) = (2\pi)^{-(N-1)/2}(|\tau_{mk}\mathbf{Q}_k^*|^*)^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}^T\tau_{mk}\mathbf{Q}_k^*\mathbf{x}\right). \tag{5.7}$$

The superscripts indicates that a matrix is scaled or that the generalized determinant is used, depending on the context. The disease set $\mathcal{M}$ is generally all the 86 diseases. Note that the precision τ_{mk} will have the same prior for each disease m, but it can have different realizations for each disease. The full Bayesian hierarchical model for the validation process is

$$\begin{aligned}
&\text{1. Likelihood:} && y_{im} \mid \eta_{imk} \sim \text{Poisson}(E_{im}\eta_{imk}), \text{ for } i \in \mathcal{I} \text{ and } m \in \mathcal{M} \\
&\text{2. Latent layer:} && \log(\eta_{imk}) = \mu_{mk} + x_{imk}, \text{ for } i \in \mathcal{I} \text{ and } m \in \mathcal{M} \\
& && \mathbf{x}_{mk} \sim \mathcal{N}(\mathbf{0}, (\tau_{mk}\mathbf{Q}_k^*)^-), \text{ for } m \in \mathcal{M} \\
& && \mu_{mk} \sim \mathcal{U}(-\infty, \infty) \propto 1, \quad \text{for } m \in \mathcal{M} \\
&\text{3. Hyperpriors:} && \tau_{mk} \sim gamma(1, 0.00005), \quad \text{for } m \in \mathcal{M}, \text{ or} \\
& && \tau_{mk} \sim PC(1, 0.01), \quad \text{for } m \in \mathcal{M}, \text{ or} \\
& && \frac{1}{\sqrt{\tau_{mk}}} \sim U(0, 5), \quad \text{for } m \in \mathcal{M}.
\end{aligned} \tag{5.8}$$

This is repeated for each model k from the training stage with the associated scaled precision matrix $\mathbf{Q}_k^*$, as well as for the ICAR. The ICAR uses the standard structure matrix defined in Section 2.5 scaled by the geometric variance. For the precision τ_{mk} three different priors are considered to see how the prior on the precision affects the validation. For the BW-ICAR the validation is performed as in Equation (5.8) and the resulting model is denoted as BW_2 while the model fit for the training stage is still called BW. The difference between the two models is in the prior on the precision τ and also in the fact that BW has two precision parameters with priors that affects the calculation of the likelihood. Thus, the results for BW_2 should be more comparable to the results for the EW-ICARS and RW-ICARs than the BW. Still, BW has been included as it is the standard way of using the model, and to give an impression of how it compares. Note that the whole validation procedure is done with INLA, while the training used both INLA and WinBUGS, depending on the model.

To be clear, this is not a cross-validation for the diseases, as the data for the validation is used in the training as well. For the full models, i.e. with 86 diseases, it would have been preferable to use for instance leave one disease out cross-validation, but this was dropped due to high runtimes. For the subset with fewer diseases I am unsure what a cross-validation would look like as the model trained on the subset is validated on all the data, so parts of the validation is not used in the training data, while some is. Additionally, for most of the datasets the removal of a disease from

the training stage would likely have little impact on the validation results, and it is also not crucial to the comparison of the models as they are all treated the same way. For the univariate models, performing a cross-validation of diseases is not even possible.

5.3 Model choice criteria

To compare the different models, three model criteria have been chosen, namely deviance information criteria (DIC), Watanabe-Akaike information criteria (WAIC) and logarithmic scoring (LS). As there are different conventions for these criteria they need to be introduced alongside a brief introduction to their interpretation.

Lets start with DIC, which is a common criteria in Bayesian modelling. The DIC is a take on the famous trade-off between accuracy and model complexity and it was introduced in Spiegelhalter *and others* (2002). DIC is essentially a penalized form of a likelihood criteria. Formally, it takes the negative log-likelihood and adds a penalizing term based on the effective number of parameters. This means that for two models with the same log-likelihood, the model with fewer effective parameters gets a lower DIC, and it thus preferred. The quantification of the number of effective parameters is not always straightforward for a Bayesian hierarchical model. This is because the effect of the priors and the random effects are hard to quantify. This penalization term for the DIC generalizes to the number of parameters for a standard linear model (Gelman *and others*, 2014). However, there are many ways to approximate this size, and a different penalization term gives rise to the WAIC.

The modern WAIC is very similar to DIC, although, when introduced in Watanabe (2010) it was not on the deviance scale. As there are many famous information criteria on the deviance scale, such as AIC, BIC and DIC, the WAIC definition on the deviance scale in Gelman *and others* (2014) has seen more use than the original definition, and will also be used in this thesis. This is also the WAIC criteria implemented in INLA, with the penalizing term as in Equation 11 in Gelman *and others* (2014). Note that the WAIC on the deviance scale only differs from the DIC in the calculation of the effective number of parameters.

Lastly, lets introduce LS, formally called average proper logarithmic scoring. A slightly altered definition from Gneiting and Raftery (2007) with a flipped sign is simply

$$\mathrm{LS}(\boldsymbol{\pi}, \mathbf{y}) = -\frac{1}{N}\sum_{i=1}^{N} \log \pi_i(y_i),$$

where $\boldsymbol{\pi}$ is a vector containing the predictive densities for each data point in $\mathbf{y}$. A good candidate for calculating these predictive densities is the conditional predictive ordinates, or CPO. For a data point y_i, $\mathrm{CPO}_i = \pi(y_i|\mathbf{y}_{-i})$ (Gómez-Rubio, 2020). Thus, the LS can be approximated by the average of the log-likelihood of observing each data point y_i given the rest of the data. Furthermore, this involves fitting N models with essentially the same cost as the full model, alike a leave-one-out cross validation. As such, it is often approximated in some way and it is trivial to obtain the LS from CPOs. As INLA can calculate approximations of the CPOs

during model training, this will be used in this thesis, see Held *and others* (2010) for further details. Additionally, INLA includes a binary variable for each CPO to indicate if the approximation is good. For an INLA object `res`, these are stored as `res$cpo$failure` while the CPO is stored as `res$cpo$cpo`. Thus, calculating the LS from the CPOs can be done by `-mean(log(res$cpo$cpo))`. Likewise, the DIC and WAIC can be accessed as `res$dic$dic` and `res$waic$waic`.

In regards to interpretation, the three criteria prefer models with lower values. Some papers have commented on some issues with the DIC for Bayesian hierarchical models like the ones in this thesis, for example Plummer (2008). One concern raised is that the DIC is known to underestimate the penalizing term for these models. As such, it it preferred to base model inference on proper scoring rules, like LS, but the DIC is still included as it is very popular.

5.4 Implementation

As mentioned earlier both INLA and WinBUGS will be used to train and validate the models. Specifically, the EW-ICAR and the RW-ICAR will be trained with WinBUGS as they are not compatible with INLA, while the remaining models are trained with INLA. For WinBUGS, an interface in R will be used which relies on the library R2WinBUGS (Sturtz *and others*, 2005) as well as the library pbugs (FISABIO, 2023) to easily run parallel chains.

RW-ICAR

Lets start with the implementation of the RW-ICAR. The implementation is largely based on the model code in the Appendix of Corpas-Burgos and Martinez-Beneito (2020). The main changes concern removing the spatial effect $\boldsymbol{\theta}$ and the calculation of the standard mortality rate, abbreviated as SMR, as these are not needed in this study. The model is implemented as

```
RW_ICAR_NOiid <- function() {
  # Likelihood
  for (i in 1:Nareas) {
    for (m in 1:Ndiseases) {
      Y[i, m] ~ dpois(lambda[i, m])
      # Modeling of the mean for each region and disease
      log(lambda[i, m]) <- log(E[i, m]) + mu[m] + phi[i,m]
      # Prior distribution for spatial effects
      phi[i, m] ~ dnorm(mean.phi[i, m], prec.phi[i, m]) #uses
       ↪  precision
    }
  }
  # Adjacencies for c and phi
  for (i in 1:n.adj) {
    sqrt.c.adj[i] <- sqrt(c[adj[i]])
    for (m in 1:Ndiseases) {
      phi.adj[i, m] <- phi[adj[i], m]
```

```
    }
  }
  # Precision for the conditional distribution of spatial effects
  for (m in 1:Ndiseases) {
    prec.phi[1, m] <- pow(sd.phi[m], -2) * sqrt(c[1]) *
    ↪  sum(sqrt.c.adj[index[1]:index[2]])
    for (i in 2:Nareas) {
      prec.phi[i, m] <- pow(sd.phi[m], -2) * sqrt(c[i]) *
      ↪  sum(sqrt.c.adj[(index[i] + 1):index[i + 1]]) # the +1 is to
      ↪  correct the indexes, works as all regions have at least one
      ↪  neighbour
    }
  }
  # Mean for the conditional distribution of spatial effects
  for (m in 1:Ndiseases) {
    mean.phi[1, m] <- inprod2(sqrt.c.adj[index[1]:index[2]],
      phi.adj[index[1]:index[2],
      ↪  m])/sum(sqrt.c.adj[index[1]:index[2]])
    for (i in 2:Nareas) {
      mean.phi[i, m] <- inprod2(sqrt.c.adj[(index[i] + 1):index[i +
      ↪  1]], phi.adj[(index[i] + 1):index[i + 1],
      ↪  m])/sum(sqrt.c.adj[(index[i] + 1):index[i +1]])
    }
    # Soft sum-to-zero restriction for spatial effects
    ceros[m] <- 0
    ceros[m] ~ dnorm(sum.phi[m], 10) #uses precision
    sum.phi[m] <- sum(phi[, m])
  }
  # Prior distributions for c
  for (i in 1:Nareas) {
    c[i] ~ dgamma(tau, tau) %_% I(0.001, ) # ensures c[i] is above
    ↪  0.001
  }
  tau <- pow(sd.c, -2)
  sd.c ~ dunif(0, 5)
  # Other prior distributions
  for (m in 1:Ndiseases) {
    sd.phi[m] ~ dunif(0, 5)
    mu[m] ~ dflat()
  }
}
```

The code follows Equation (5.2). Note that the random effect $\mathbf{x}$ in the Equation is called `phi` in the code as x is used multiple other places in the R code, for example for plotting. Additionally, the indexes as subscripts in Equation (5.2) are now formulated as a matrix indexing. Furthermore, `Nareas` equals I and `Ndiseases` equals M. The code is then essentially defining all the dependencies between the different parts of the model. The least straightforward part is the indexing. As $\boldsymbol{\phi}_m$ and $\mathbf{c}$ are stored as vectors, additional vectors are needed to identify the neighbours.

Those are called `adj`, which contains the neighbours of all regions i in a ordered fashion, so first all the neighbours of region $i = 1$ up to $i = I$, and `index`, which contains the cumulative sum of the number of neighbours for the regions, with the number 1 added as the first element of the vector. As an example consider the graph in Figures 2.3 and 2.4. Then, `index` equals $[1, 2, 5, 8, 10]$ and `adj` equals $[2, 3, 1, 3, 4, 1, 2, 4, 2, 3]$. These lists are then used together to handle the spatial structures for instance in lines $20 - 32$. The soft sum-to-zero constraint on lines $34 - 36$ is also noteworthy. Essentially, a vector of length M with zeros is given a Gaussian "prior" with a mean as the sum of the spatial structured effect $\boldsymbol{\phi}_m$ for each disease m and a precision of 10, as `dnorm` in R2WinBUGS uses the precision parametrization (Gelman *and others*, 2005).

EW-ICAR

I extended the code above to implement the EW-ICAR. The model is now based on the Bayesian hierarchical model in Equation (5.4). The primary difference between the two implementations is the change from the I region specific weights $\mathbf{c}$ to the vector for edge precisions $\boldsymbol{\tau}$ of length 111, which is the number of edges in the correspondence graph. To handle this mapping, a vector of lists is used to keep track. This list is a result of iterating through the neighbours for each region, and for each pair of neighbours adding the index of the edge to both regions. Consider again the graph in Figures 2.3 and 2.4, then `tau.list` equals $[[1, 2], [1, 3, 4], [2, 3, 5], [4, 5]]$, where the numbers $1 - 5$ correspond to one of the five edges in the graph. They are numbered following the indexes of the spatial regions/nodes and then in increasing index for the neighbour. So, edge 1 is the edge between nodes 1 and 2 while edge 3 is the edge between 2 and 3. This list then indicates which precisions in $\boldsymbol{\tau}$ that affects each region i, which is the list `Tau2` in the code. The indexing list is used on line $26 - 28$, and the result is used on lines $19 - 24$ and $30 - 35$. It is worth noting that the dummy list created on lines $26 - 28$, `Tau2.tau.list`, could be viewed as unnecessary, as in R all the indexing of the lists could be handled on the same line. While this is true in R, WinBUGS does not support that degree of nested indexing and the program crashes if the list accessing is not split as in the code. Additionally, the priors for $\boldsymbol{\tau}$ is done on lines $42 - 46$, and the rest of the code is very similar to the RW-ICAR.

```
1  EW_ICAR_NOiid <- function() {
2    # Likelihood
3    for (i in 1:Nareas) {
4      for (m in 1:Ndiseases) {
5        Y[i, m] ~ dpois(lambda[i, m])
6        # Modeling of the mean for each region and disease
7        log(lambda[i, m]) <- log(E[i, m]) + mu[m] + phi[i,m]
8        # Prior distribution for spatial effects
9        phi[i, m] ~ dnorm(mean.phi[i, m], prec.phi[i, m]) #uses
          ↪  precision
10     }
11   }
12   # Adjacencies for phi
13   for (i in 1:n.adj) {
```

```
    for (m in 1:Ndiseases) {
      phi.adj[i, m] <- phi[adj[i], m]
    }
  }
  # Precision for the conditional distribution of spatial effects
  for (m in 1:Ndiseases) {
     prec.phi[1, m] <- pow(sd.phi[m], -2) *
      ↪  sum(Tau2.tau.list[index[1]:index[2]])
     for (i in 2:Nareas) {
       prec.phi[i, m] <- pow(sd.phi[m], -2) *
        ↪  sum(Tau2.tau.list[(index[i]+1):index[i+1]])
     }
  }
  #to avoid nesting too many lists inside each other, caused crashes
  for(i in 1:n.adj){
    Tau2.tau.list[i]<-Tau2[tau.list[i]]
  }
  # Mean for the conditional distribution of spatial effects
  for (m in 1:Ndiseases) {
      mean.phi[1, m] <- inprod2(Tau2.tau.list[index[1]:index[2]],
      phi.adj[index[1]:index[2],
       ↪  m])/sum(Tau2.tau.list[index[1]:index[2]])
    for (i in 2:Nareas) {
      mean.phi[i, m] <-
       ↪  inprod2(Tau2.tau.list[(index[i]+1):index[i+1]],
       ↪  phi.adj[(index[i] + 1):index[i + 1],
       ↪  m])/sum(Tau2.tau.list[(index[i]+1):index[i+1]])
    }
    # Sum-to-zero restriction for spatial effects
    ceros[m] <- 0
    ceros[m] ~ dnorm(sum.phi[m], 10) #uses precision
    sum.phi[m] <- sum(phi[, m])
  }
  # Prior distributions for Tau2 - all the edge precisions
  for (t in 1:(n.adj/2)){
    Tau2[t] ~ dgamma(alpha, alpha) %_% I(0.001, )
  }
  alpha <- pow(sd.alpha, -2)
  sd.alpha ~ dunif(0, 5)

  # Other prior distributions
  for (m in 1:Ndiseases) {
    sd.phi[m] ~ dunif(0, 5)
    mu[m] ~ dflat()
  }
}
```

Convergence diagnostics

Both of these models are then run in WinBUGS for all the five subsets of diseases described in Section 5.1. However, as mentioned in Section 2.6.1, some important parameters for the MCMC methods must be defined. The most crucial parameters are the burn-in and the length of the Markov chain to be simulated. There are no general rules for this. However, as increasing the number of diseases included also increases the number of parameters, a longer Markov chain with a longer burn-in is often necessary to obtain good mixing.

Assessment of convergence must be checked after the simulations, which has been done through the efficient number of draws (ESS) [Chapter 18](Ross, 2022) for each parameter as well as the $\hat{R}$ for each parameter (Gelman and Rubin, 1992). These were introduced in Section 2.6.1. Note that parameter in this context not only refers to the typical model parameters like a precision τ, but also realizations of the random effects, for instance ϕ_{im} for a region i and disease m. For $\hat{R}$ the rule of thumb stating that $\hat{R}$ should be less than 1.1 has been employed [Chapter 18](Ross, 2022), and this was strictly upheld for parameters considered important, for instance the weights $\mathbf{c}$ in the RW-ICAR and the precisions $\boldsymbol{\tau}$ is the EW-ICAR. However, if only a few "non-cruical" parameters are above 1.1, like a few ϕ_{im}, the results have been deemed satisfactory. In terms of ESS, no hard limit have been set, but the ESS was above 100 for almost all the parameters, and often above 500. All of the Markov chains were run with three parallel Markov chains, and the burn-in and length of the Markov chains can be seen in Table 5.3.

Disease subset	Burn-in	Length
10	5000	10000
20	5000	10000
Cancer	10000	25000
50	10000	25000
86	10000	30000

Table 5.3: Table of burn-in and length of the chains for the different subsets of diseases. A Markov chain is simulated of the given length, and the first samples are discarded per the burn-in.

ICAR

The ICAR, the BW-ICAR and the adaptive model used in the validation is implemented with INLA. The ICAR is already implemented as a function and the important part is to make INLA scale the model, which is set to TRUE by default, but the argument is included to highlight that it is done. The scaling keeps the effect of the prior for the precision parameter constant across all the models in the validation. For instance, the ICAR with the penalized complexity prior specified in Equation (5.8) for a single disease becomes

```
formula_ICAR <- Y ~ offset(log(E)) +
      f(ID, model = "besag", graph = adj_matrix, scale.model = TRUE,
```

```
3       constr = TRUE, hyper = list(prec = list(prior = "pc.prec",
        ↪  param = c(1, 0.01))))
```

Note that INLA refers to the ICAR as the `besag` instead, and that a soft sum-to-zero constraint is added by default when using the `besag` model by the `constr` argument. This has been set to true here to emphasize that it is used. The `adj_matrix` is the standard adjacency structure matrix for the ICAR as defined in Section 2.5. In the code it is constructed from a shapefile for the provinces of mainland Spain. The expected cases are called `E` and are included as an offset.

BW-ICAR

Implementing the other two models is more involved and I utilized the `rgeneric` framework in INLA. This involves defining a new model in INLA with the necessary parts, which can then be used similarly to an already implemented model. For example, if a model for the BW-ICAR is defined, then the `model = "besag"` can be exchanged with `model = BW_model`. The first step is to define the `rgeneric` as below.

```
1  inla.rgeneric.BWICAR = function(
2    cmd = c("graph", "Q", "mu", "initial",
     ↪  "log.norm.const","log.prior", "quit"),
3    theta = NULL)
4  {
5  #Input:
6  #df: contains the relevant information for the areal data
7  #prior_str is either Gamma, PC1, PC3 or U
8  #AMat is the adjacency matrix, so 1 for bordering regions
   ↪  representing a first order structure, else 0
9
10   envir = parent.env(environment())
11
12   # The link between the internal parameters and the precisions
13   interpret_theta <- function() { return(list(tau1 = exp(theta[1L]),
14                                       tau2 = exp(theta[2L])))}
15
16   graph <- function() {return(Q())}
17
18   # Defining the precision matrix
19   Q <- function() {
20     N <- nrow(df) #df is passed as an argument, N is the number of
       ↪  regions
21     R1 <- matrix(0, nrow = N, ncol = N) #inside autonomous regions
22     R2 <- matrix(0, nrow = N, ncol = N) #across autonomous regions
23
24     #uses the adjacency matrix called AMat
25     non_zero_indices <- which(AMat == 1, arr.ind = TRUE)
26
27     for (k in seq_len(nrow(non_zero_indices))) {
```

```
      i <- non_zero_indices[k, 1]
      j <- non_zero_indices[k, 2]

      if(df[["Cod_CCAA"]][i] == df[["Cod_CCAA"]][j]){ #the name of the autonmous regions
        R1[i, j] <- -1
        R1[i, i] <- R1[i, i] + 1
      }
      else{
        R2[i, j] <- -1
        R2[i, i] <- R2[i, i] + 1
      }
    }
    #Scaling with geometric variance
    gv <- exp(1 / N * sum(log(diag(INLA:::inla.ginv(R1 + R2)))))
    R_star_list <- list(R1 = R1*gv, R2 = R2*gv)

    p <- interpret_theta()
    Q <- R_star_list$R1 * p$tau1 + R_star_list$R2 * p$tau2
    return(inla.as.sparse(Q)) #sparse representation
  }

  mu <- function() {return(numeric(0))}

  initial <- function() {return(c(4, 4))}#The default

  log.norm.const <- function() {return(numeric(0))}

  log.prior <- function() {
    p <- interpret_theta()
    if (pr_str == "gamma")  {
      prior <- dgamma(p$tau1, shape = 1, rate = 0.00005, log = TRUE) + log(p$tau1) +
        dgamma(p$tau2, shape = 1, rate = 0.00005, log = TRUE) + log(p$tau2)}
    else if (pr_str == "PC3") {
      prior <- inla.pc.dprec(p$tau1, u = 3, alpha = 0.05, log=TRUE) + log(p$tau1) +
        inla.pc.dprec(p$tau2, u = 3, alpha = 0.05, log=TRUE) + log(p$tau2)}
    else if (pr_str == "PC1") {
      prior <- inla.pc.dprec(p$tau1, u = 1, alpha = 0.01, log=TRUE) + log(p$tau1) +
        inla.pc.dprec(p$tau2, u = 1, alpha = 0.01, log=TRUE) + log(p$tau2)}
    else if (pr_str == "U") {prior <- -0.5*log(p$tau1) - 0.5*log(p$tau2)}
    return(prior)
  }

```

```
  quit <- function() {return(invisible())}

  #to ensure theta is defined
  if (!length(theta)) theta = initial()

  vals <- do.call(match.arg(cmd), args = list())
  return(vals)
}
```

The most relevant parts are defining the precision matrix $\mathbf{Q}$ in the function `Q()` and defining the prior on the log-scale for the involved parameters. `Q()` divides the edges across or inside autonomous regions into `R1` and `R2`. Then they are scaled with a common factor before being combined with their own precision parameter as discussed in Section 5.1. In terms of the log-prior, both of the two precision parameters are given the same prior in the function `log.prior()` and the log of the parameter is added because of the transformation to the log-scale. There are four different priors to choose from as they were used in the sensitivity analysis in Appendix C. When implementing an IGMRF with `rgeneric`, these two functions are typically all that must be altered, along with potentially more or different values for the mean and initial values for the parameters in `mu()` and `initial()`. The input briefly explained in the comment at the top is passed to the `rgeneric` function and is for example used to keep track of the autonomous region for each province and the neighbourhood structure. For a more thorough introduction to the `rgeneric` framework see Section 4.3 in Sand-Larsen (2025) and Rue (2021). The specific model that can be used in the INLA formula is created as

```
BW_model <- inla.rgeneric.define(inla.rgeneric.BWICAR,
        df = ss_provinces, AMat = adj_matrix, pr_str = "PC1")
```

where the arguments `df`, `AMat` and `pr_str` are used inside the `inla.rgeneric.BWICAR` function. A formula similar to the ICAR with the new `BW_model` can look like

```
formula_BW <- Y ~ offset(log(E)) + f(ID, model = BW_model,
  extraconstr = list(A = matrix(1, nrow = 1, ncol = nAreas), e = 0))
```

Note that the soft sum-to-zero constraint now is added by enforcing a linear constraint of the form $\mathbf{Ax} = e$ and that there is no specified prior as this is handled inside the `BW_model`.

Validation model

For the validation model the precision matrix $\mathbf{Q}$ from the training stage is passed as an input before being scaled by the geometric variance and multiplied by the precision parameter. For the prior the argument `pr_str` indicates which of the three priors to use in `log.prior()`. Otherwise it is mostly the same as the BW-ICAR, and the result is as follows:

```r
inla.rgeneric.Validation = function(
  cmd = c("graph", "Q", "mu", "initial",
  ↪  "log.norm.const","log.prior", "quit"),
  theta = NULL)
{
  #Input:
  #R as the unscaled precision structure matrix
  #nAreas is the number of provinces
  #pr_str indicates which prior to use

  envir = parent.env(environment())

  interpret_theta <- function() {return(list(tau = exp(theta[1L])))}

  graph <- function() {return(Q())}

  Q <- function() {
    p <- interpret_theta()
    #Scaling with geometric variance
    gv <- exp(1 / nAreas * sum(log(diag(INLA:::inla.ginv(R)))))
    return(inla.as.sparse(p$tau * gv * R))
  }

  mu <- function() {return(numeric(0))}

  initial <- function() {return(4)}#The default

  log.norm.const <- function() {return(numeric(0))} #INLA computes it

  log.prior <- function() {
    p <- interpret_theta()
    if (pr_str == "gamma")  {prior <- dgamma(p$tau, shape = 1, rate =
    ↪  0.00005, log = TRUE) + log(p$tau)}
    else if (pr_str == "PC") {prior <- inla.pc.dprec(p$tau, u = 1,
    ↪  alpha = 0.01, log=TRUE) + log(p$tau)}
    else if (pr_str == "U") {prior <- -0.5*log(p$tau)}
    return(prior)
  }

  quit <- function() {return(invisible())}

  #to ensure theta is defined
  if (!length(theta)) theta = initial()

  vals <- do.call(match.arg(cmd), args = list())
  return(vals)
}
```

To ensure reproducibility of the results, all the code used in this thesis is available at https://github.com/Halvardgithub/Master--Comparison-of-three-adaptive-approaches.

The README also has a section dedicated to the versions for the relevant libraries used in the R-code. It should be noted that INLA sometimes crashed when running the validations for multiple priors after each-other. However, it always worked to simply restart R and try again, but I am not quite sure what causes it to crash, and I did not identify any patterns, besides when first running the validations for 86 diseases for one prior, and then beginning a new validation cycle for a new prior without restarting R in between. Some of the files I have generated are to large for GitHub and have instead been uploaded to Zenodo, available from https://zenodo.org/records/15799852. These files can be downloaded and used to generate the Figures and Tables in this thesis. For more details see the README in the GitHub repository.

Now that the validation study has been introduced with a brief explanation of the implementation, the next step is to review the results from the training and validation in the following Section.

6 Results

After implementing and running the models as described in Section 5, there are a few results to consider. The first subsection will focus on results from the training step, which includes explorations of the resulting precision matrices and the resulting weights for the edges in the neighbourhood graphs, which will be visualized in a few complementing plots. The second subsection will focus on the validation step, and specifically on comparing the performance of the different neighbourhood structures with respect to the criteria introduced in Section 5.3. Figure 4.1 will come in handy as specific regions will be mentioned at points in the upcoming analysis.

6.1 Comparison of different neighbourhood structures

In terms of the neighbourhood structures, the two multivariate methods, RW-ICAR and EW-ICAR, are the most interesting. Both because of their flexibility and as the same structure, i.e. the precision matrix $\mathbf{Q}$ in the prior, is used for the whole subset of diseases they are trained on. As Equation (5.2) and Equation (5.4) combine $\mathbf{Q}$ with a disease specific precision parameter τ_m, some care must be taken in the visualizations. The issue is that for the RW-ICAR with $n = 10$ the precision τ_m could take values in $(1, 2)$ while for $n = 20$ maybe it instead could take values in $(5, 8)$. Then just visualizing the edges based on $\mathbf{Q}$ would result in different scales, even though $\tau_m\mathbf{Q}$ might be on a similar scale for a disease m in both cases. Thus, the resulting $\mathbf{Q}$'s have been scaled by the average precision parameter τ_m over all the diseases included in the training, to make the scales comparable across the models and the different subsets. Lets make this clear with an example. Consider the EW_86 model, then a different posterior is generated for each $\tau_m,\ m = 1, ..., 86$. The average mean precision across all the diseases is calculated as $\overline{\tau} = \frac{1}{86}\sum_{m=1}^{86} E_{post}[\tau_m]$. Then $\overline{\tau}\mathbf{Q}$ will be visualized for each combination of the two models and four subsets, where the subset for cancers have been exempted.

For the four different values of n, the number of diseases, the resulting edge maps are presented in Figure 6.1, based on the definitions and implementation from Section 5. The thickness of each edge represents the precision given to the edge in the corresponding precision matrix $\mathbf{Q}$, and scaled by the average τ_m for the subset, i.e. $\overline{\tau}$. A thicker edge represents a stronger spatial smoothing than a thinner edge. Across the different plots this is hard to compare as the scales for the edges vary substantially, although they have the same interpretation.

Figure 6.1 shows that the RW-ICAR has much lower max edge weights than the EW-ICAR, but there are generally more edges close to the highest edge for the RW-ICAR than the EW-ICAR. This means that the RW-ICAR has multiple moderate connections compared to a few strong connections in the EW-ICAR. This is likely due to the flexibility difference in the two models. Furthermore, the EW-ICAR appears to make clusters of $2-4$ provinces which are strongly related, especially for $n = 50$ and $n = 86$. However, some provinces, like Madrid (the autonomous region with a single province in the centre of Spain) and Jaen (top right province of the large bottom autonomous region), are weakly connected to all their neighbours, at

least for $n = 50$ and $n = 86$. The RW-ICAR also concludes that Madrid is an outlier for higher n, in the sense that it is weakly connected to its neighbours and thus has a low degree of spatial smoothing, which is also reflected in its specific region weight being relatively small, see Table E.1. It should be noted that the interpretation of individual weights c_i is closely related to the weights $\mathbf{c}_{ne(i)}$ as $Q_{ij} = \sqrt{c_i c_j}$. Furthermore, the standard deviations shown in Table E.1 are rather large.

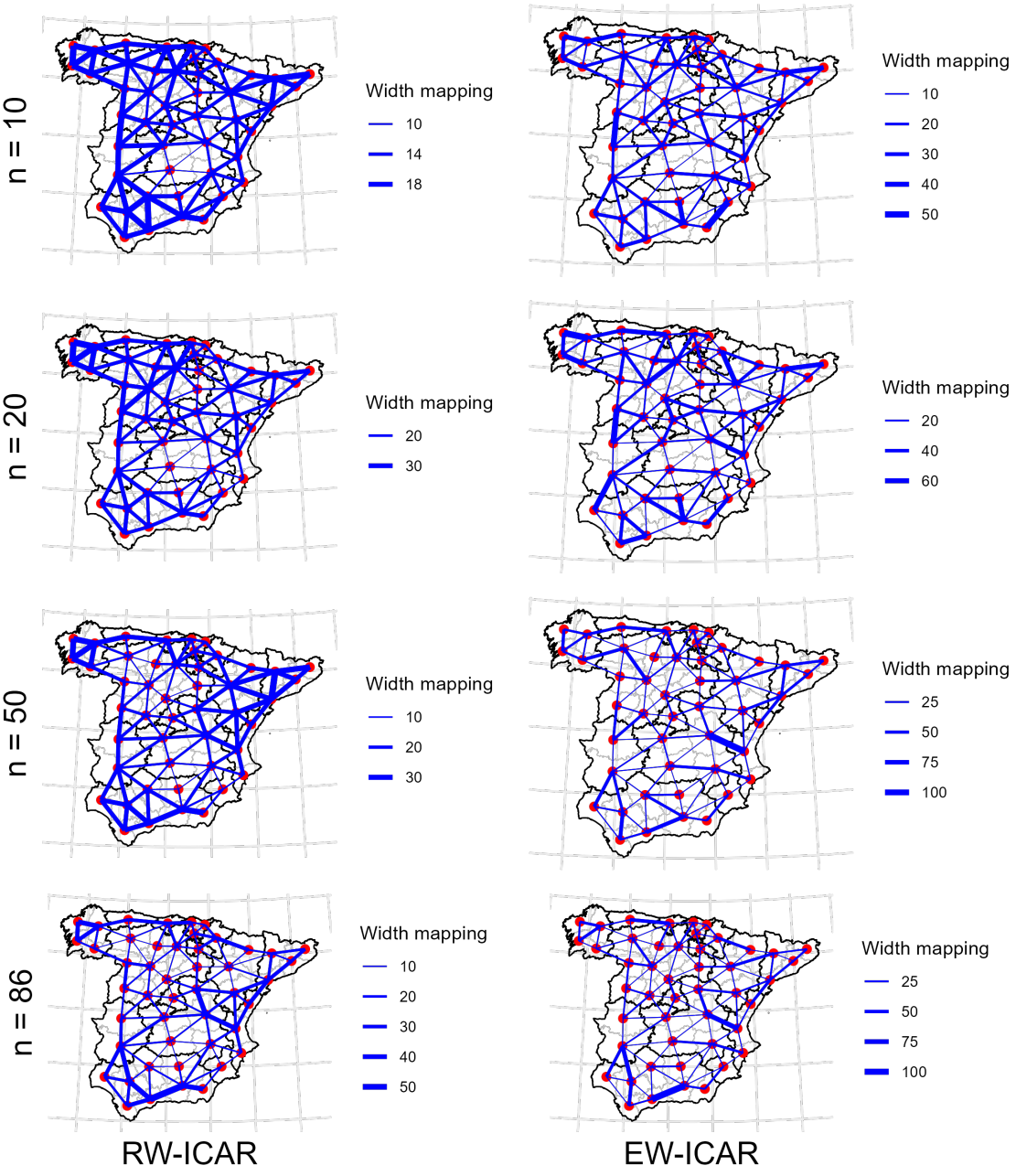


Figure 6.1: The resulting graphs show the precision $\overline{\tau} Q_{ij}$, the scaled posterior mean of Q_{ij}, for each of the eight combinations from combining RW-ICAR and EW-ICAR with $n \in \{10, 20, 50, 86\}$. Province centres are red dots and the edges are blue with varying thickness to indicate the value of $\overline{\tau} Q_{ij}$. Note that all the plots have a different mapping between the value and the thickness of the edge.

Lastly, a short note on comparing the weights $\mathbf{c}$ between different subsets. Since they are scaled by a disease specific τ_m for each subset, if the weights $\mathbf{c}$ are larger for one subset, this could simply be because the average precision across the disease $\overline{\tau} = \frac{1}{n}\sum_{m=1}^{n} E_{post}[\tau_m]$ has decreased. Clearly, the weights $\mathbf{c}$ and the precision $\boldsymbol{\tau}$ is closely intertwined.

Overall, the structures of the edge maps are rather different, and thus the associated neighbourhood structure will also be different. Including more diseases clearly has an effect, but it remains to be seen how it affects the validation study.

Another way of visualizing the precisions is to sum all the edges connected to each province in Figure 6.1. This precision represents the conditional precision for each region when the values for all the neighbouring regions are known. Thus, the provinces with a higher precision are more constrained by the values of their neighbours, while provinces with a lower precision can more easily deviate from their neighbours. The plots for the described precision can be seen in Figure 6.2. The precision for a province for a given n is generally similar for both models. However, the patterns vary somewhat for different n. Consider for example Leon in the top left of Spain. For $n = 10$ it is light blue corresponding to a precision around 120, while for $n = 50$ and $n = 86$ it is dark blue corresponding to a precision around 50, and this occurs for both models. This could indicate that the subset of diseases for $n = 10$ benefits from strong spatial smoothing, i.e. high precision, for Leon, while this apparently is not the case when all the diseases are included. Another explanation is related to the number of neighbours for each province. For the RW-ICAR most of the edges in Figure 6.1 for $n = 10$ and $n = 20$ are of a similar magnitude, thus the total precision for each province is highly correlated with the number of neighbours for each province. This is supported by the fact that most of the provinces with a lower precision are external provinces with few neighbours, although there are some internal provinces, like Jaen, that have multiple neighbours and yet a small total precision.

For $n = 86$ both models generally agree on which provinces are more or less spatially dependent. For example, the three light blue models in the south has some of the highest total mean posterior precisions, and thus are the provinces with the strongest degree of spatial smoothing. This indicates that the regions follow similar trends as their neighbours. As a counter example, consider the north-west of Spain for $n = 50$ and $n = 86$. Even though there is a large autonomous region with nine provinces, the degree of spatial smoothing is low, compared to the rest of Spain. This indicates that there are differences in the underlying risk between the provinces, and also in the observed cases. Note that this is not the case for $n = 10$ and $n = 20$. As mentioned, the total precision can be interpreted in terms of the conditional distributions, and the same goes for the individual edges. For the RW-ICAR and the EW-ICAR the conditional distributions for the structured random effect are

$$x_i|\mathbf{x}_{-i}, \boldsymbol{\tau} \sim N\left(\frac{\sum_{j\sim i}\tau_{ij}x_j}{\sum_{j\sim i}\tau_{ij}}, \left(\sum_{j\sim i}\tau_{ij}\right)^{-1}\right), \quad i = 1, ..., I. \tag{6.1}$$

The mean of $x_i|\mathbf{x}_{-i}, \boldsymbol{\tau}$ is a weighted mean of the neighbouring x_j's where the weights are the precisions for the edges connecting i and j. Additionally, the total precision,

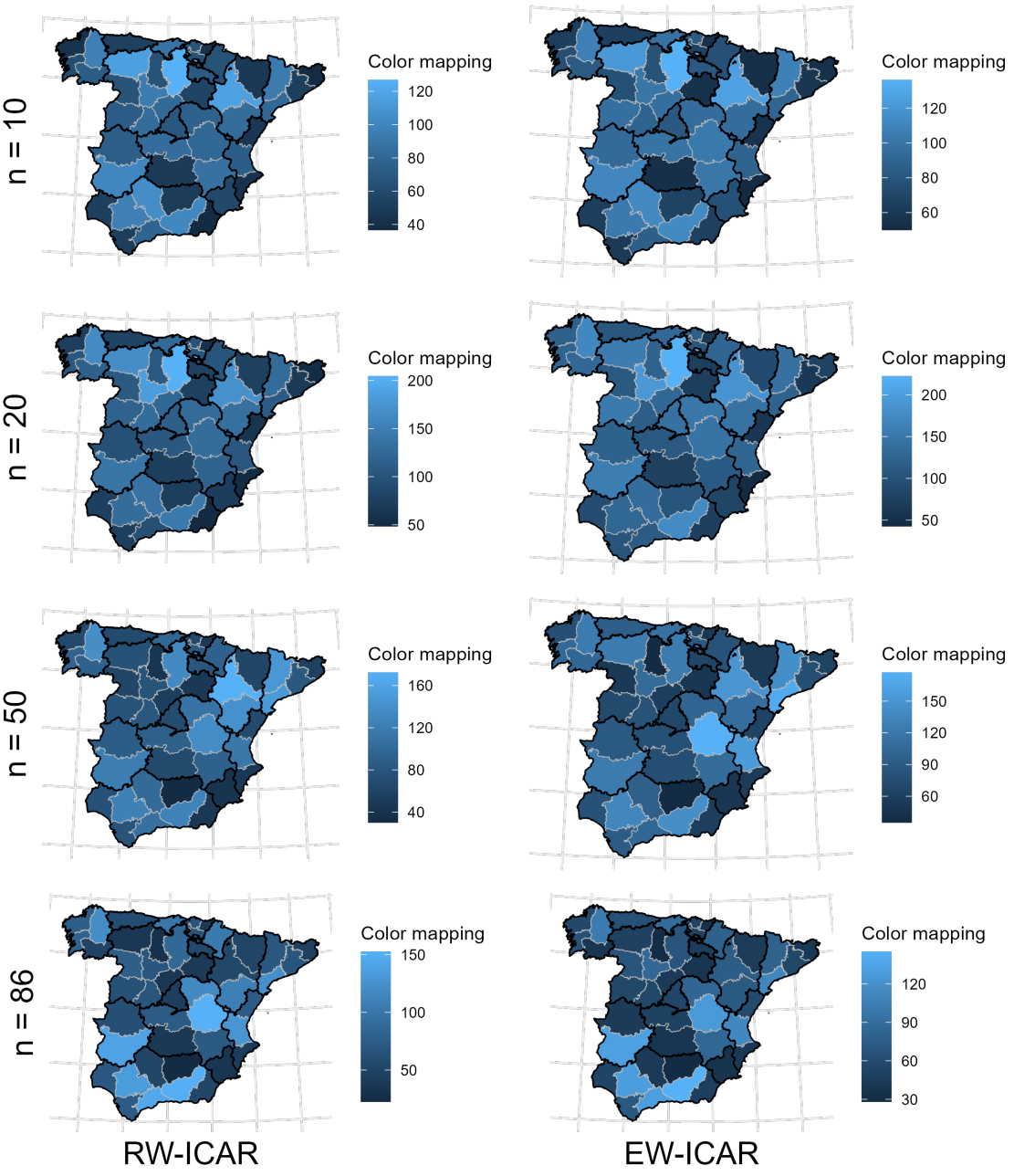


Figure 6.2: The total posterior mean precision for each province for the two models and the different subsets based on n.

or the sum of the precision parameters for all edges for a region i, is then the inverse conditional variance for $x_i|\mathbf{x}_{-i}, \boldsymbol{\tau}$. Thus, a high total precision indicates a high degree of spatial smoothing as the x_i in question is discouraged from deviating from the weighted mean of its neighbours. And vice versa for a low total precision.

The total precision shown in Figure 6.2 appears to behave similarly for both methods, while the edges in Figure 6.1 appear more different. To make this comparison easier, both Figures have been summarized with scatter-plots in Figure 6.3, as well as the computed correlation between them. This clearly shows that the total precision for each province is strongly correlated at almost above $0,9$ for the four subsets of diseases. These results indicate that the models pick up on similar trends in the data, but also that they distribute the total precision for each province in different ways, as mentioned earlier. In regard to the two other models, namely the ICAR and the BW-ICAR, both of the multivariate models appear to deviate from the assumptions of the aforementioned models. Recall that the ICAR assumes spatial stationarity, which would correspond to an edge graph where all the edges are the same width. And the BW-ICAR splits the edge set in two, as in Figure 3.2. As mentioned, both of these assumptions do not appear to be supported by the edge structures in Figure 6.1, but lets supplement this visual analysis with a mathematical analysis. For both the multivariate models the posterior mean edge precisions shown in Figure 6.1 will be split in two groups according to the assumption of the BW-ICAR. Then, the two groups are compared at a log-scale, a transformation to more closely align with the Gaussian distribution, with an ANOVA test, which assumes the data follows a Gaussian distribution. Specifically, the ANOVA test compares the variance inside groups with the variance across groups, and quantifies how likely it is that both groups where drawn from the same distribution. The p-values in Table

	RW-ICAR	EW-ICAR
n=10	0.075	0.064
n=20	0.002	0.219
n=50	0.397	0.470
n=86	0.288	0.830

Table 6.1: p-values for the comparison of the log mean posterior precision of edges inside or across autonomous regions.

6.1 indicates that there is no significant difference, with significance level $\alpha = 0.05$, between edges inside an autonomous region and edges across autonomous regions, except for $n = 20$ for the RW-ICAR. The p-value is also relatively small for both models with $n = 10$. However, the values for both $n = 50$ and $n = 86$ are larger. These results could indicate that for the first 20 chosen diseases, there is a difference between these two edge groups, but this difference disappears when more diseases are included. Thus, it is likely that this difference would disappear altogether with a different seed for choosing diseases. Alternatively, the models need a sufficient amount of data to discover an accurate structure and deviate substantially from the priors, at which point there is no difference between the two groups. When using a different seed for choosing the random disease subsets, the results are similar and can be seen in Table E.2. This adds credibility to the hypothesis that these p-values are independent of the chosen subsets, and likely instead a facet of the models.

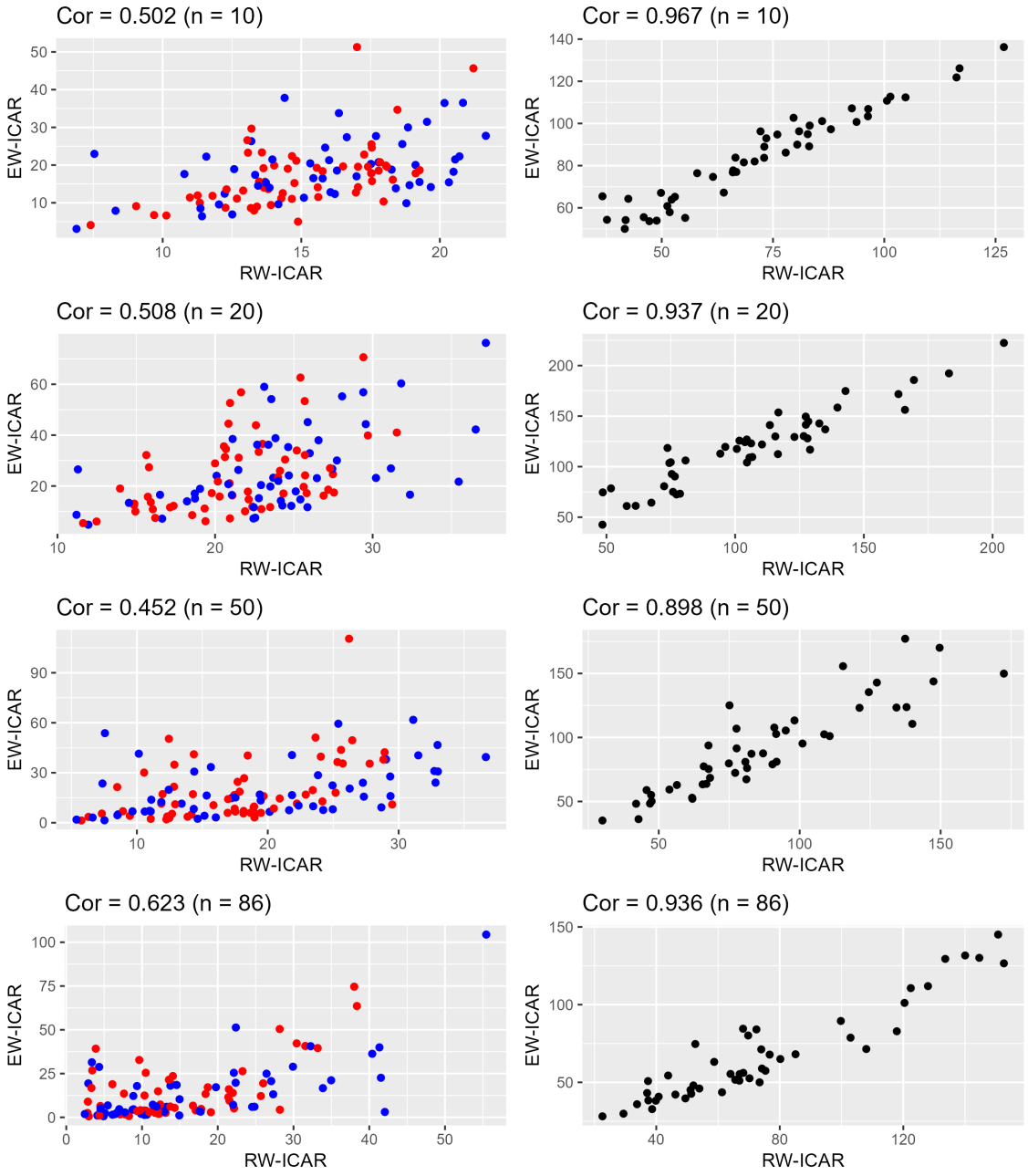


Figure 6.3: For all the plots the correlation is shown between the values for the RW-ICAR and the EW-ICAR alongside the given value of n. **Left column:** Posterior mean precisions τ_{ij} for all edges. Edges in the same autonomous region are in blue while the others are red. **Right column:** Total precision for each province.

Overall, the p-values indicates that there is no significant difference between edges inside or across autonomous regions, especially for higher n. This implies that the assumption in the BW-ICAR is unwarranted, and too restrictive for this application. However, it could still outperform the ICAR in the validation described in Section 5.2. For the application in Aleshin-Guendel and Wakefield (2024) the assumption was found to improve the model fit. Furthermore, both the multivariate models clearly deviate from the binary neighbourhood structure of the ICAR. So, it will be interesting to see the effect of this in the validation study.

Another point of interest is the correlation between the precision matrices for each of the models across the different subsets. This quantifies the degree of change in neighbourhood structures between the different values of n. For both the RW-ICAR in Table 6.2 and the EW-ICAR in Table 6.3 the highest correlation lies on the first off-diagonal. This makes sense as the "neighbouring" n-values have most of their diseases in common, while the larger subsets adds additional different diseases. The correlation for the RW-ICAR is generally higher than for the EW-ICAR, except for between $n = 50$ and $n = 86$. This is also supported by Figure 6.1, where the RW-ICAR changes less drastically when new diseases are introduced compared to the EW-ICAR. Furthermore, correlation is scale-invariant, which means that the different scales for the edges has no effect. These results have also been checked with a second seed for choosing the random subsets, and the results are similar for both the RW-ICAR and the EW-ICAR. These results can be seen in Figure E.3 and Figure E.4.

	n=10	n=20	n=50	n=86
n=10	1.00	0.77	0.61	0.39
n=20	0.77	1.00	0.54	0.37
n=50	0.61	0.54	1.00	0.71
n=86	0.39	0.37	0.71	1.00

Table 6.2: Correlation between the estimated edge weights τ_{ij} for different values of n for the RW-ICAR.

	n=10	n=20	n=50	n=86
n=10	1.00	0.71	0.41	0.23
n=20	0.71	1.00	0.47	0.30
n=50	0.41	0.47	1.00	0.78
n=86	0.23	0.30	0.78	1.00

Table 6.3: Correlation between the estimated edge weights τ_{ij} for different values of n for the EW-ICAR.

6.2 Validation on real data

After exploring the structures of the estimated precision matrices from the training step, lets present the results of the validation on real data. As mentioned in Section

5, all the models will be fit to each individual disease. This is done by using scaled structure matrix $\mathbf{Q}_k^*$ and combining it with a diseases specific precision parameter τ_m, as described in Section 5.2. Three model criteria are used to assess and compare the model fit for the different neighbourhood structures, namely DIC, WAIC and LS as described in Section 5.3. To visualize this a box plot is employed for each structure, denoted by the model it comes from, for each decision criteria. However, as the scale of the decision criteria vary between the diseases because of different scales for number of cases as well as varying suitability of the model assumptions, a reference level must be used. For this purpose the ICAR was chosen as the reference. Thus, for a given disease and a given criteria the scores for all the models are subtracted by the score for the ICAR. For example

$$DIC^*_{RW\_20}(m) = DIC_{RW\_20}(m) - DIC_{ICAR}(m), \quad \forall\, m$$

for the DIC. This value represents the difference between the given model, in this case the RW_20, and the ICAR for each disease m. Thus, a $DIC^*_{RW\_20}(5) = -1$ means that the RW_20 had a DIC one lower than the ICAR for disease 5. Similarly, a positive value indicates that the ICAR outperformed the RW_20. Thus, negative values in the box plots means that the neighbourhood structure form the model in question outperformed the structure of the ICAR for a given disease and vice versa.

The conventions used to generate the box plot is the default use of `geom_boxplot` which generates a box that outlines the 25% and 75% quantiles, along with a line for the median, or 50% quantile, inside the box. Additionally, the whisker is a line to the value that is furthest from the box up to a max distance of 1.5 times the width of the box. Points outside the whiskers are denoted as outliers and plotted as dots, these can be seen in Appendix E.2.

Note that the results often will be framed as comparing the models, even though the comparison is between the neighbourhood structures for the different models, most of which are trained on the disease data. As this is somewhat cumbersome to state repeatedly, this meaning is implied it the following analysis, where often only the models are referenced. Thus, saying that the ICAR outperformed the EW_20, means that the neighbourhood structure from the ICAR outperformed the neighbourhood structure from the EW_20 model when validated on individual diseases.

The resulting plot when the prior for the validation precision $\tau \sim gamma(0, 0.00005)$ can be seen in Figure 6.4. This shows that the ICAR outperforms is on par with all the adaptive models in terms of the median for DIC and WAIC. For WAIC, it should be noted that for higher n, and for the BW_2, the 75% quantile is close to zero along with the median, and that the 25% quantile is below 0 for higher n and for the BW. For DIC all the adaptive models outperform the ICAR for 75% of the diseases, except for the BW. Although, both the 75% quantial and the median is essentially 0 for all the models. The 25% quantiles are again lowest for BW and higher n. Thus, essentially all the adaptive models perform better than the ICAR, even though the median is at 0. Additionally, both the BW-ICAR and the EW-ICAR appear to perform better with higher n, and the best for $n = 86$. This is mainly visible through the decreasing 25% quantile. Overall the BW and the EW_86 perform the best for DIC and WAIC.

In terms of LS the BW model is confidently the worst, including the ICAR, while the remaining models outperform the ICAR for at least 75% of the diseases. The best model in terms of LS is EW_86 as it has the lowest median and the lowest 25% quantile. None of the models outperformed the ICAR for all of the diseases, although some models nearly do, like the EW_86. Overall, all of the adaptive models outperformed the ICAR on average for LS, except for BW, and the ICAR was the worst in terms of DIC and WAIC. None of the adaptive models are always better than the ICAR for a given criteria for all the diseases, even with their increased complexity and runtimes. Some are even substantially worse for specific diseases, see for example WAIC for EW_C and LS for BW in Figure E.1 in Appendix E.2.

The results are somewhat different when $\tau \sim PC(1, 0.01)$. Beware that the y-axis in Figure 6.5 is now on a different scale for the three criteria compared to Figure 6.4. Additionally, the values for the ICAR, which is used as a reference, has also been fitted with the new prior, so the values are not directly comparable across Figures. In terms of the median, some of the models outperform the ICAR on average in terms of WAIC, and most of the models for DIC. The ICAR only outperforms the BW. For LS all of the medians outperform or are on par with the ICAR, except for BW, and most of the 25% quantiles significantly beat the ICAR, notably by a larger degree than with the gamma prior. Actually, the EW_86 outperforms the ICAR in terms of LS for almost all the 86 diseases. Overall, the EW_86 seems to perform the best for the three criteria. The general trend that the multivariate models improve for higher n is clearly present for all three criteria.

The results in Figure 6.6 are again slightly different when $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$. For WAIC, only the medians for higher values of n is below 0, alongside EW_C, and for all the models the 75% quantile is above 0, with the two lowest for BW_2 and RW_C. Thus, only a few of the models outperform the ICAR on average for WAIC, and the others are mostly on par. For DIC there are similar patterns, but most, if not all, of the models perform better here as the boxes have shifted downwards relative to the red line at 0. The two best models are EW_50 and EW_86 which outperform the ICAR for at least 75% of the diseases, and have the lowest median. For LS most of the 75% quantiles are close to 0, except for BW. This means that most of the models outperform the ICAR for 75% of the diseases, although to a varying degree. For example the EW_86 generally outperforms the ICAR with a greater margin than the RW_10 or BW_2. The outliers are included in Figure E.3 in Appendix E.

Overall, the adaptive models outperformed the ICAR on average in terms of LS regardless of the prior assigned to the precision in the validation. However, that does not mean the prior choice is effectless. The degree of outperformance is much greater for the uniform and PC priors compared to the gamma prior. In terms of DIC and WAIC the prior choice was even more influential. With the gamma prior most of the adaptive models are on par with the ICAR in terms of the median for DIC and WAIC, with BW performing notably bad. For the PC prior more of the adaptive models outperform the ICAR for DIC and WAIC. Generally, the ICAR is more competitive for WAIC, while EW_86 clearly is better for DIC. The uniform prior is relatively similar to the PC prior for DIC and WAIC in terms of the model comparison. An interesting point is that the BW was the only model to outperform the ICAR in terms of WAIC with the gamma prior for more than

50% of the diseases, as the median was below 0. However, with the uniform and PC prior the BW performed the worst in terms of WAIC. This is likely due to the prior for the validation changing between the plots, but the BW model is the same across all the the validation plots as it is the results from the training stage where two precisions are used, specifically $\tau_1, \tau_2 \sim PC(1, 0.01 =$. To conclude, most of the adaptive models outperformed the ICAR in terms of LS, and to a lesser extent for DIC and WAIC. This was also dependent on the prior used for the validation.

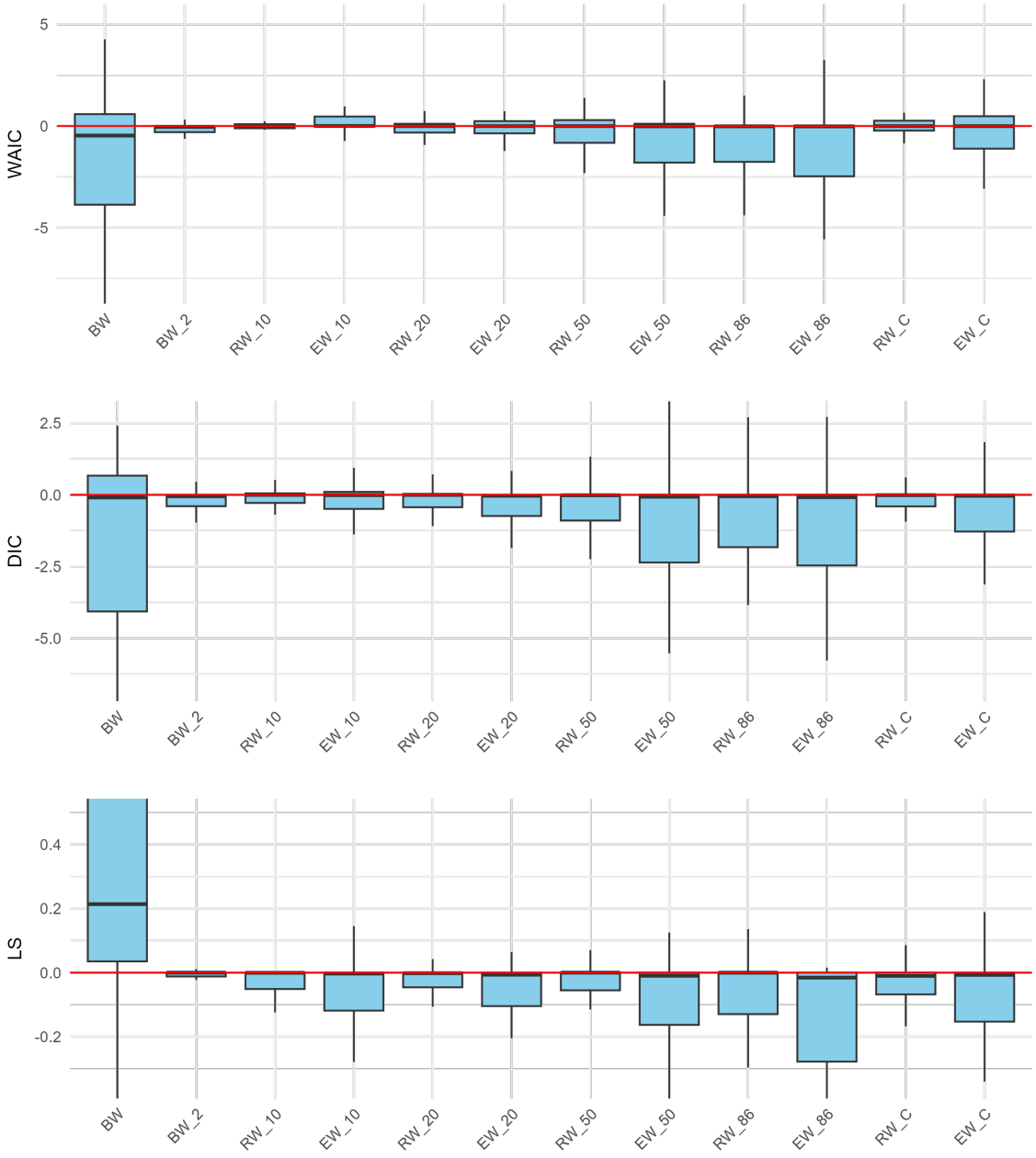


Figure 6.4: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease with $\tau \sim gamma(0, 0.00005)$. Values lower than zero indicate that the method outperformed the ICAR, which is used as a reference.

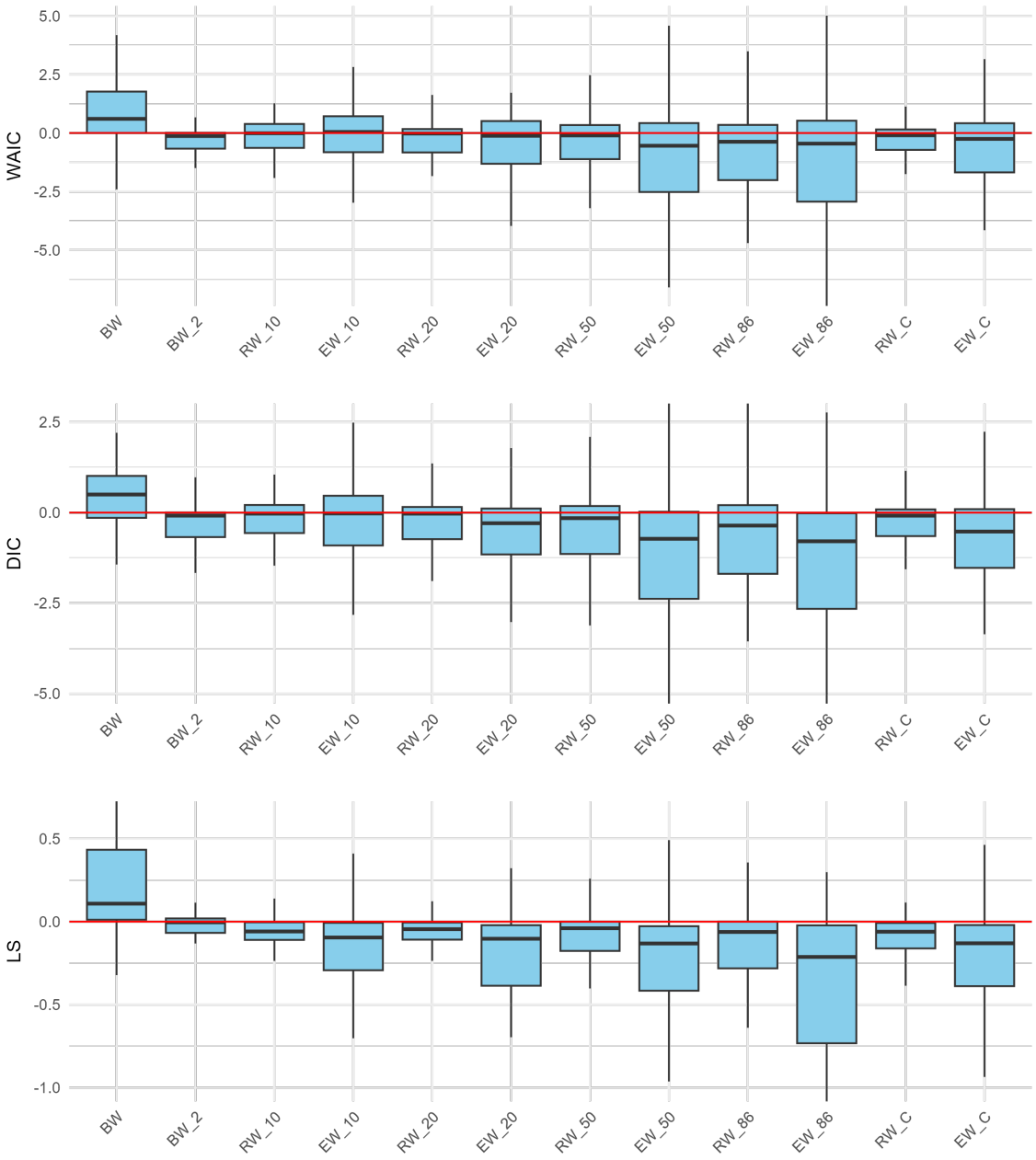


Figure 6.5: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease with $\tau \sim PC(1, 0.01)$. Values lower than zero indicate that the method outperformed the ICAR, which is used as a reference.

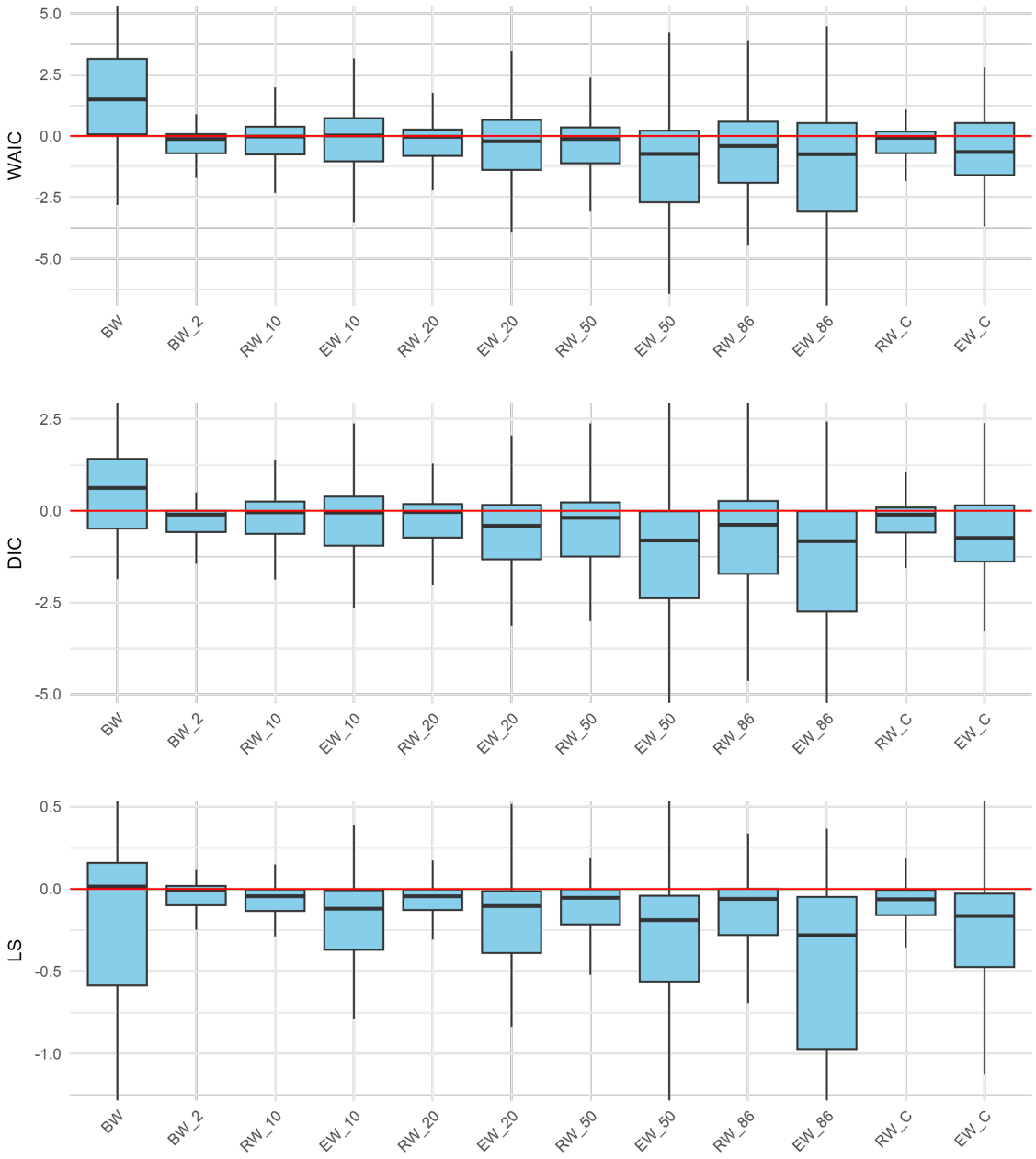


Figure 6.6: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease with $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$. Values lower than zero indicate that the method outperformed the ICAR, which is used as a reference.

It generally looks like adding more diseases improves the multivariate models, even when the diseases are chosen randomly and might not have anything in common. Furthermore, lets explore the results for a more hand-picked disease subset, namely validating purely on cancers. The main interest is then whether the methods trained purely on cancer data, RW_C and EW_C, outperforms the models with all the diseases.

Overall, there are similar trends here in terms of the three different priors for the

validation precision parameter. For the gamma prior in Figure 6.7, most of the medians for DIC and WAIC are close to 0. However, so is most of the 75% quantiles, which indicates that most of the models outperformed the ICAR, especially as the 25% quantiles are mostly below 0. In terms of the 25% quantile the BW does the best, but EW_20, EW_50 and EW_C does the best in terms of the medians for DIC. The two models trained only on cancers, RW_C and EW_C, outperforms the ICAR, and EW_C does the best of the two. Notably, they both perform better relative to the other models compared to the validation on all the diseases in Figure 6.4. All the models outperform the ICAR for LS, except for BW. There is no clear difference between RW_C and EW_C and the other adaptive models in either direction. Thus, with the gamma prior it looks like adding additional diseases had little impact. Although, for example the EW_50 outperformed the EW_86, so more diseases is not always better as it was when validating on all the diseases.

Different results can be observed for the PC prior in Figure 6.8. Most of the adaptive models are better than the ICAR for WAIC, and almost all are better for DIC. Note that EW_C performs significantly better here compared to Figure 6.5, which is as expected. All the adaptive models again outperform the ICAR for LS, except for BW, and one could argue that the EW_50 and EW_C did the best, as their 25% quantiles are the lowest, along with their medians. It is not clear that the two cancer models perform better than the other adaptive models. Although, the EW_C is among the best models here. Notably, the EW-ICARs outperformed their counterparts in the RW-ICARs.

With the uniform prior in Figure 6.9 the patterns align with the results for the PC prior. The EW-ICARs are best in terms of DIC and WAIC. For LS, the EW_86 and EW_C did the best based on the median and the 25% quantile.

On average, the adaptive models outperformed the ICAR for the cancer subset in terms of LS. Overall, the model comparison is rather similar to comparing for all the 86 diseases. However, the two cancer models, especially the EW_C, perform better on the cancer subset, which was expected. Another take away is that RW_C and EW_C performed on par with the models training on all the diseases, and it varied from prior to prior and criteria to criteria which of the training subsets of diseases yielded the best model.

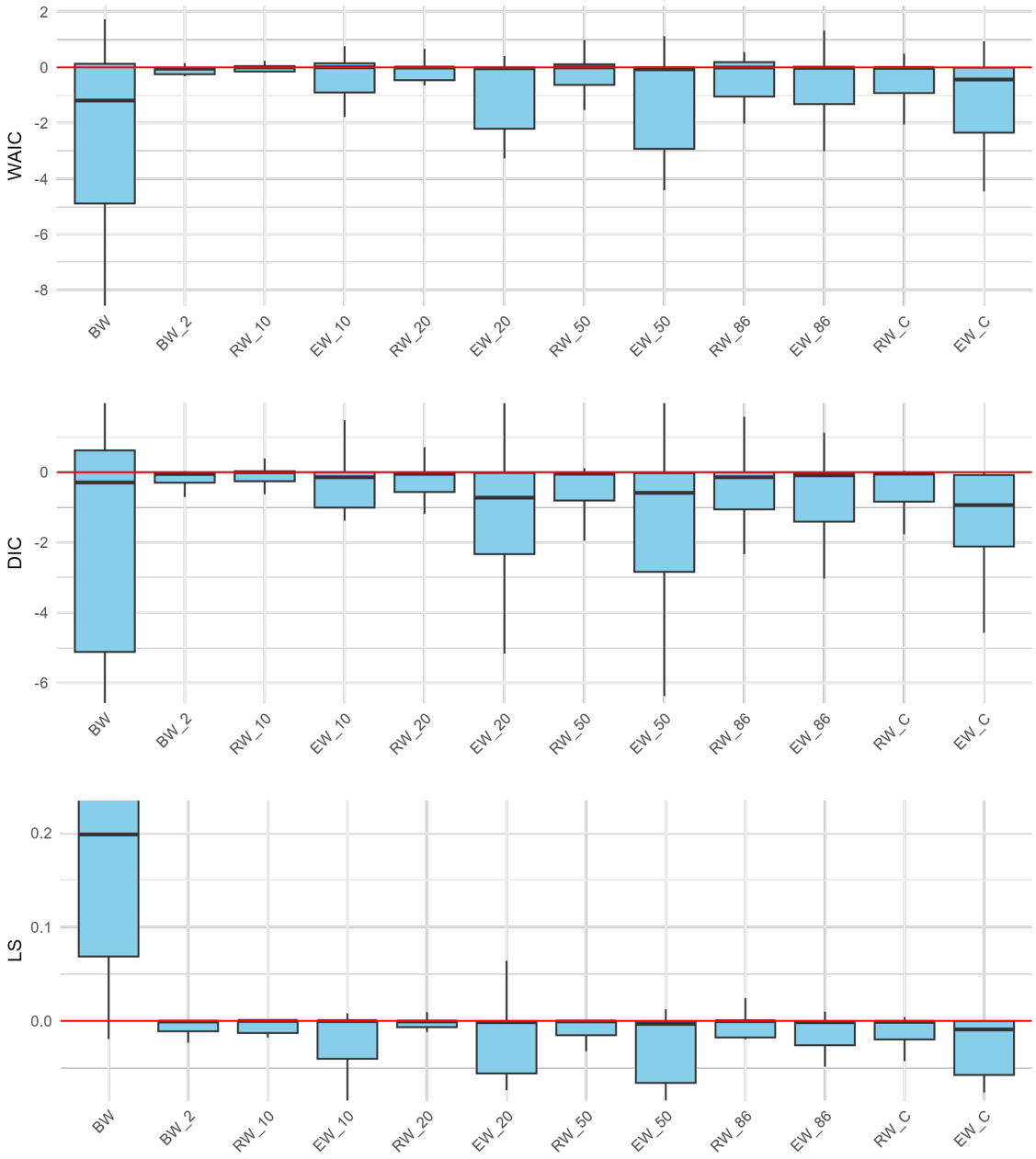


Figure 6.7: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease. The prior for the precision was $\tau \sim gamma(1, 0.00005)$. Values lower than zero indicate that the method outperformed the ICAR, which is used as a reference.

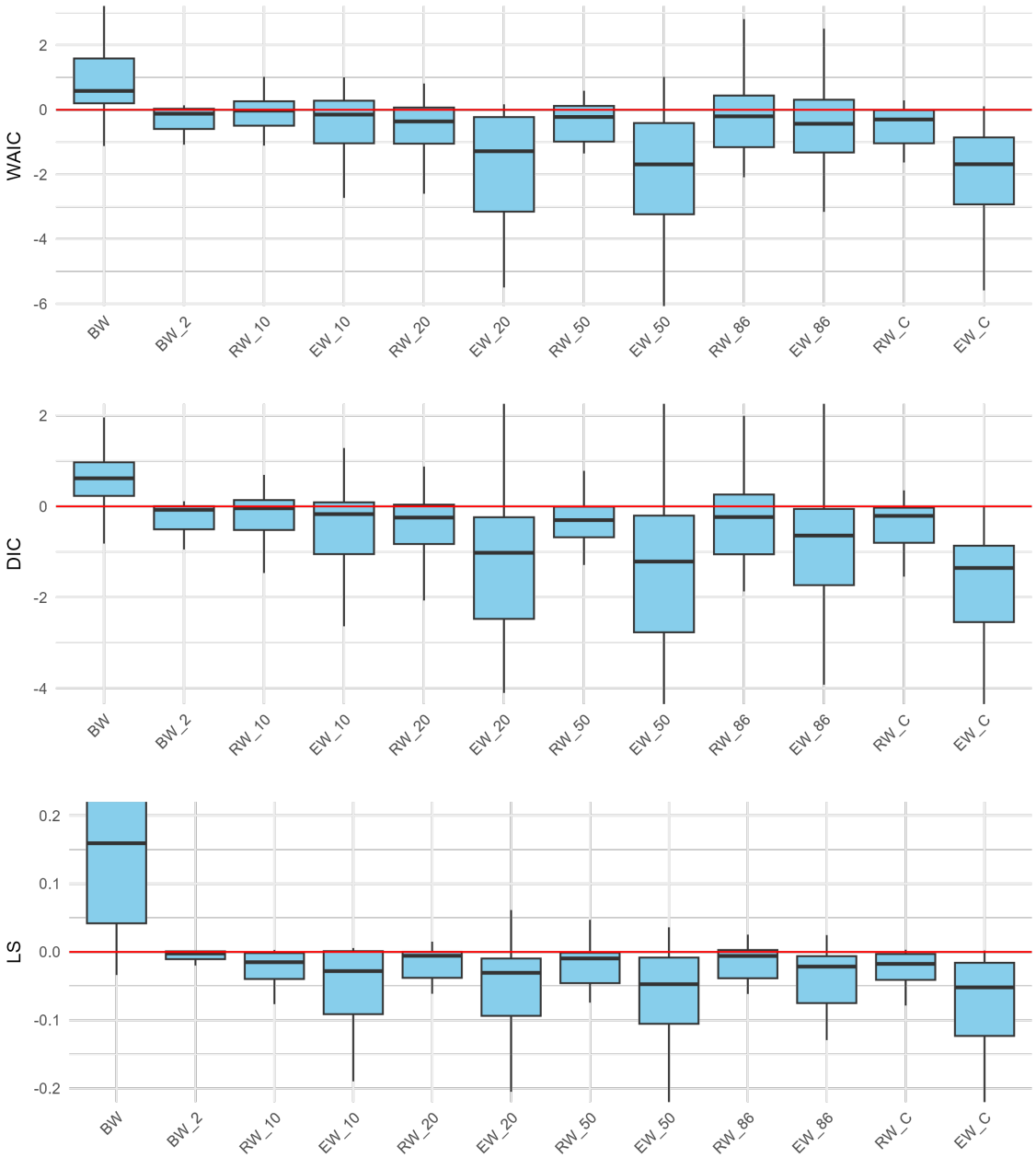


Figure 6.8: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease. The prior for the precision was $\tau \sim PC(1, 0.01)$. Values lower than zero indicate that the method outperformed the ICAR, which is used as a reference.

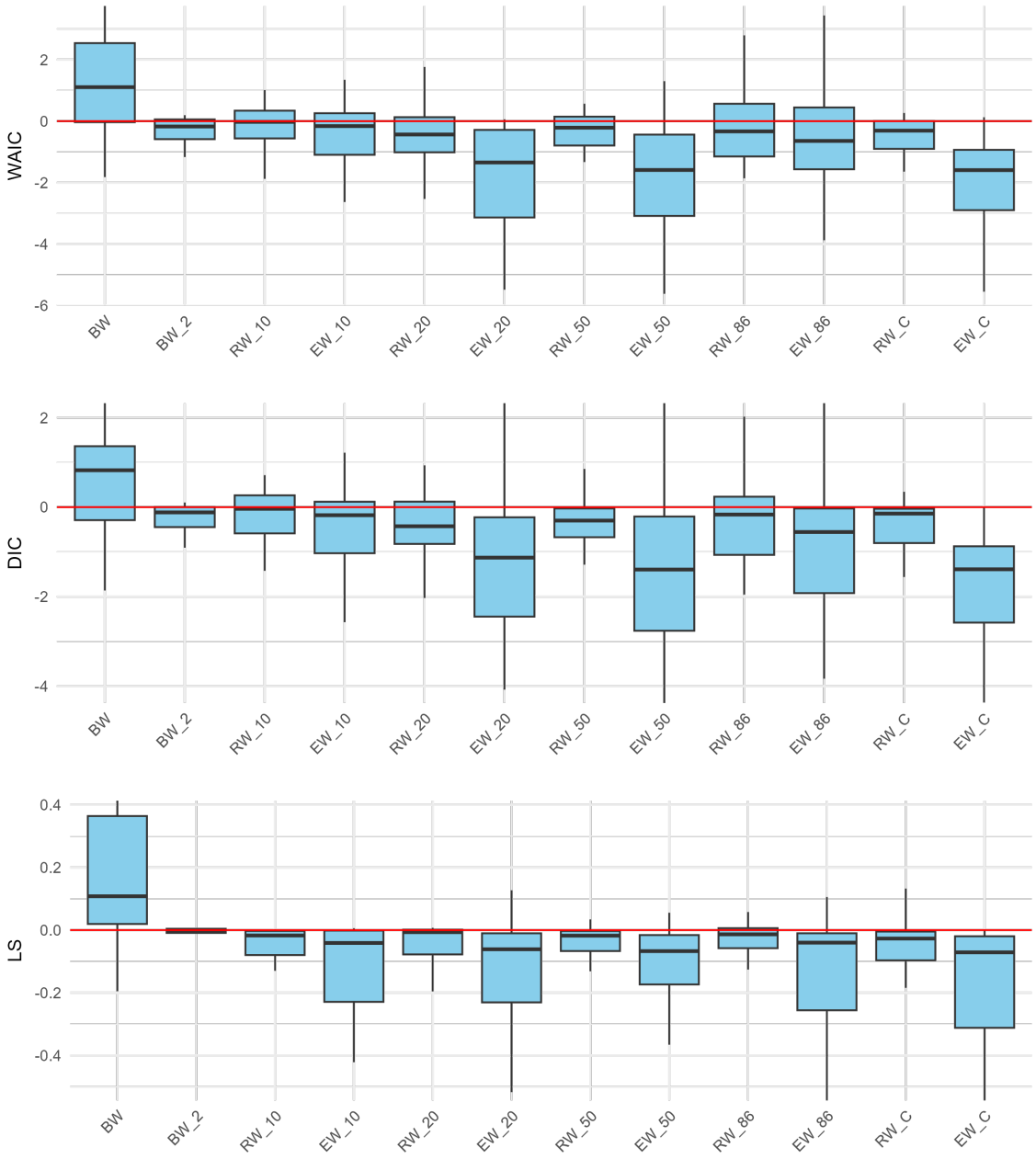


Figure 6.9: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease. The prior for the precision was $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$. Values lower than zero indicate that the method outperformed the ICAR, which is used as a reference.

7 Discussion

In terms of LS, the ICAR was generally beaten by the more flexible neighbourhood structures from the adaptive models. It was more competitive for WAIC and DIC, but the neighbourhood structures based on a higher number of diseases were still better. The ICAR was often on par with the neighbourhood structures from multivariate models with low n, i.e. 10 or 20, and also with both of the BW-ICARs, although this varied dependign on the validation prior and the specific criteria. For higher n the flexible structures performed better, and the neighbourhood structures from EW-ICARs generally did better than the ones from RW-ICARs. Thus, it seems the performance of the neighbourhood structures were aligned with the number of parameters used in the training of each one, recall Table 5.1.

The BW-ICARs performed at a similar level to the ICAR structure and were sometimes better and sometimes worse depending on the validation subset of diseases and the validation prior. The out of the box structure with two precision, denoted BW was generally the worst model, except for when $\tau \sim gamma(1, 0.00005)$. When using the fitted structure from this model and scaling it by a single precision parameter in the validation, denoted BW_2, it always outperformed the ICAR structure, although by a small margin. Both of these structures allowed for one strength of spatial smoothing for neighbours in the same autonomous region and another strength for neighbours in different autonomous regions. This assumption was not supported by the multivariate models, see Table 6.1, and the more flexible multivariate models outperformed the BW-ICARs.

The next step up in flexibility was the RW-ICAR which assumes that each region can be more or less spatially connected. This appeared to model the data well and overall outperformed the neighbourhood structure with fewer parameters, although for low n the improvement was small. However, the RW-ICARs were not always better on average than the less flexible models, see for example Figure 6.4.

The most flexible models, the EW-ICARs, overall gave rise to the best neighbourhood structures. Now the assumption is that all pairs of neighbours have their own strength of spatial smoothing, and there is no sharing of information across different pairs of neighbours. The EW-ICARs generally had the lowest DIC and WAIC and often outperformed the structure for the ICAR for at leats 75% of the diseases in question, see for example Figure 6.5. In terms of LS the EW-ICARs performed even better relative to the other models. Overall, the EW_86 model was the best across the three criteria and three priors, and the EW-ICARs generally performed better when the number of diseases n increased.

Overall, the results in Section 6 indicates that the spatial stationarity assumption for the ICAR is too restrictive for the application of this thesis. Namely, univariate disease mapping for the provinces of Spain. However, that does not mean it is not a useful assumption. The multivariate models that gave rise to the best performing neighbourhood structures, needs replicates, in this case multiple diseases, to obtain good results in the validation study. Replicates across time is for example mentioned in Riddervold (2024). It is also important that these replicates share some underlying risk structure. In applications where the access to good data with replicates of some

sort are not available, the ICAR is a suitable model. Generally, larger areas with more variability in factors relevant to the response variable indicates that the spatial stationarity assumption is unwarranted. This appears to be the case for Spain. For data rich situations the results support that more flexible neighbourhood structures are the way to go.

To clarify, when the spatial stationarity assumption for the ICAR is mentioned it is often in reference to the single global precision parameter, but it also the incudes the binary first-order neighbourhood structure. The adaptive models breaks this assumption by introducing multiple precision parameters in the precision matrix. When this trained matrix is used in the validation, it only has one precision parameter. However, I would argue the specific fixed values inside the matrix now breaks the spatial stationarity assumption, as the smoothing will vary over space because of the values no longer being binary.

Generally, increasing the amount of data in the training yielded better results in the validation. This was clear in the results from the validation study in Section 6, where the multivariate models with $n = 10$ and $n = 20$ performed worse than the ones with $n = 50$ and $n = 86$.

Another major point of interest for this thesis was the difference between model performance for neighbourhood structures trained on different subsets of the data. As mentioned, the multivariate models performed better when trained on a higher number of diseases, but they were still not better than the ICAR for all the diseases in the validation. Additionally, two neighbourhood structures were trained strictly on cancer data. When validated on all the diseases they performed worse than the structures with a higher n, but on par with the models for $n = 10$ and $n = 20$, which was as expected. However, the most relevant results came from validating only on the different cancers, the same ones the two structures were trained on. Then the two cancer models where competitive with the other multivariate models, and in some cases straight up better. For instance, EW_C often outperformed EW_86, although it was closer with EW_50. For RW_C it was on par with the other RW-ICARs, but not necessarily any better than them. In total, across the different criteria and priors for the validation precision, the EW_C performed the best,closely followed by EW_50. This indicates that it is not always beneficial to add additional diseases, and could also indicate that at least some of the other diseases in the full dataset have underlying risk structures that differ from the cancers. It should be mentioned that there could be significant differences in the underlying risk structures between the cancers as well.

These results indicate that it is not always better with more data. This stems from the fact that the multivariate models in this thesis, defined in Section 3.2, share the precision matrix $\mathbf{Q}$ across all the diseases. Thus, the structure of the spatial smoothing is shared across all the diseases, and this is only beneficial if the underlying risk structure follows similar trends. As such, analysis of these structures could be useful to construct training sets. This has not been explored in this thesis, but it could be a relevant future work.

Additionally, it is clear that training on multiple diseases has the potential to improve on the univariate disease case. This is due to the sharing of information across

the diseases, which allows for more complex models to be used, i.e. with a higher number of parameters. Thus, even if the interest is in a single disease, the inference could be improved by including other diseases with a similar underlying risk structure.

Some other avenues in which this thesis can be expanded on is including more priors in the sensitivity analysis, as well as other adaptive models. For example the divide and conquer approaches introduced in Orozco-Acosta *and others* (2021) and Abdul-Fattah *and others* (2024). This would make the validation study more robust. Another idea is to try a similar validation study on different data, for example different diseases or from a different country. It would also be interesting to try a finer spatial granularity, for instance on the municipalities of Spain instead of the provinces. This thesis also avoided the use of covariates, but as covariates could explain parts of the smoothing now performed by the random effects, this could improve the inference. This would also make the validation study more alike the state of the art in the field, where including covariates is the norm [Chapter 5](MacNab, 2022). Additionally, when using these models for an actual study, it is recommended to include some unstructured random effect, for instance an IID, along with the structured models that have been the focus of this thesis.

References

Abdul-Fattah, E., Krainski, E., Niekerk, J. V. and Rue, H. (2024). Non-stationary Bayesian spatial model for disease mapping based on sub-regions, *Statistical Methods in Medical Research* **33**(6): 1093–1111.

Aleshin-Guendel, S. and Wakefield, J. (2024). Adaptive Gaussian Markov random fields for child mortality estimation, *Biostatistics* **26**(1): kxae030.

Allévius, B. (2018). On the precision matrix of an irregularly sampled AR(1) process. Accessed on 25.06.2025. Available from: https://arxiv.org/abs/1801.03791.

Assunção, R. M. and Castro, M. S. (2004). Multiple cancer sites incidence rates estimation using a multivariate Bayesian model, *International Journal of Epidemiology* **33**(3): 508–516.

Berger, J. (2006). The case for objective Bayesian analysis, *Bayesian Analysis* **1**: 385–402.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2): 192–236.

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**: 1–21.

Brezger, A., Fahrmeir, L. and Hennerfeind, A. (2007). Adaptive Gaussian Markov random fields with applications in human brain mapping, *Journal of the Royal Statistical Society Series C: Applied Statistics* **56**(3): 327–345.

Chen, Z.-Y., Deng, X.-Y., Zou, Y., He, Y., Chen, S.-J., Wang, Q.-T., Xing, D.-G. and Zhang, Y. (2023). A Spatio-temporal Bayesian model to estimate risk and influencing factors related to tuberculosis in Chongqing, China, 2014–2020, *Archives of Public Health* **81**: 42.

Congdon, P. (2008). A spatially adaptive conditional autoregressive prior for area health data, *Statistical Methodology* **5**(6): 552–563.

Corpas-Burgos, F. and Martinez-Beneito, M. (2020). On the use of adaptive spatial weight matrices from disease mapping multivariate analyses, *Stochastic Environmental Research and Risk Assessment* **34**: 531–544.

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

Duncan, E., White, N. and Mengersen, K. (2017). Spatial smoothing in Bayesian models: A comparison of weights matrix specifications and their impact on inference, *International Journal of Health Geographics* **16**: 47.

Fernández, E. G. and S., J. C. (2010). Accessed on 26.06.2025. Available from: https://commons.wikimedia.org/wiki/File:Provinces_of_Spain.svg.

FISABIO (2023). pbugs. Version 1.0.6. Available from: https://github.com/fisabio/pbugs.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**(410): 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper), *Bayesian Analysis* **1**(3): 515 – 534.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science* **7**(4): 457 – 472.

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

Gelman, A., Ligges, U. and Sturtz, S. (2005). R2winbugs: a package for running WinBUGS from R, *Journal of Statistical Software* **12**: 1–16.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.* **6**: 721–741.

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

Grant, J. E. (2019). Researching the law of the Spanish autonomous communities, *GlobaLex*.

Gómez-Rubio, V. (2020). *Bayesian Inference with INLA*, Chapman & Hall/CRC Press. Boca Raton, FL.

Held, L., Schrödle, B. and Rue, H. (2010). *Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA*, Physica-Verlag HD, Heidelberg, pp. 91–110.

Instituto Nacional de Estadística (2020a). Accessed on 25.06.2025. Available from: https://ine.es/dynt3/inebase/en/index.htm?padre=7924.

Instituto Nacional de Estadística (2020b). Accessed on 02.07.2025. Available from: https://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p08/l0/&file=03002.px&L=0.

Irony, T. Z. and Singpurwalla, N. D. (1997). Non-informative priors do not exist a dialogue with José M. Bernardo, *Journal of Statistical Planning and Inference* **65**(1): 159–177.

Kim, J., Lawson, A., Neelon, B., Korte, J., Eberth, J. and Chowell, G. (2023). Evaluation of Bayesian spatiotemporal infectious disease models for prospective surveillance analysis, *BMC Medical Research Methodology* **23**: 171.

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

Knorr-Held, L. and Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **164**(1): 73–85.

Lawson, A. B. (2013). *Bayesian disease mapping: Hierarchical modelling in spatial epidemiology*, Chapman and Hall/CRC. Boca Raton, FL.

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

Lu, H., Reilly, C. S., Banerjee, S. and BP, C. (2007). Bayesian areal wombling via adjacency modeling, *Environmental and Ecological Statistics* **14**: 433–452.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modeling framework: Concepts, structure and extensibility, *Statistics and Computing* **10**: 325–337.

MacNab, Y. C. (2022). Bayesian disease mapping: Past, present, and future, *Spatial Statistics* **50**: 100593.

MacNab, Y. C. (2023). Adaptive Gaussian Markov random field spatiotemporal models for infectious disease mapping and forecasting, *Spatial Statistics* **53**: 100726.

MacNab, Y. C., Kmetic, A., Gustafson, P. and Sheps, S. (2006). An innovative application of Bayesian disease mapping methods to patient safety research: a Canadian adverse medical event study, *Statistics in Medicine* **25(23)**: 3960–3980.

Martino, S. and Riebler, A. (2020). *Integrated Nested Laplace Approximations (INLA)*, John Wiley & Sons, Ltd, pp. 1–19. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08212.

Morris, M. (2025). Accessed on 04.07.2025. Available from: https://mc-stan.org/learn-stan/case-studies/sum_to_zero_vector.html.

Natalia, Y. A., Molenberghs, G., Neyens, T., Hens, N. and Faes, C. (2025). Empirical analysis of COVID-19 confirmed cases, hospitalizations, vaccination, and international travel across Belgian provinces in 2021, *PlOS ONE* **20**(5): 1–15.

Orozco-Acosta, E., Adin, A. and Ugarte, M. D. (2021). Scalable Bayesian modelling for smoothing disease risks in large spatial data sets using INLA, *Spatial Statistics* **41**: 100496.

Piironen, J. and Vehtari, A. (2017). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Vol. 54 of *Proceedings of Machine Learning Research*, PMLR, pp. 905–913.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison, *Biostatistics* **9**(3): 523–539.

Riddervold, O. (2024). Non-stationary spatial random walk models on graphs, *NTNU*. Available from: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3154388.

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**: 1145–1165.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*, Springer.

Ross, K. (2022). An introduction to Bayesian reasoning and methods. Available from: https://bookdown.org/kevin_davisross/bayesian-reasoning-and-methods/.

Rue, H. (2021). *Defining a latent model in R or C*. Available from: https://inla.r-inla-download.org/r-inla.org/doc/vignettes/rgeneric.pdf.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC. 6000 Broken sound parkway NW.

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.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. K. (2016). Bayesian computing with INLA: A review, *Annual Reviews* **4**: 395–421.

Sahu, S. K. and Böhning, D. (2022). Bayesian spatio-temporal joint disease mapping of Covid-19 cases and deaths in local authorities of England, *Spatial Statistics* **49**: 100519.

Sand-Larsen, H. E. (2025). Adaptive Gaussian Markov random fields for data with known shocks. Available from: https://github.com/Halvardgithub/Master-project/blob/main/Master_project_Halvard%20-%20Finished.pdf.

Simkin, J., Dummer, T., Erickson, A., Otterstatter, M., Woods, R. and Ogilvie, G. (2022). Small area disease mapping of cancer incidence in British Columbia using Bayesian spatial models and the smallareamapp R Package, *Frontiers in Oncology* **12**: :833265.

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.

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

Sturtz, S., Ligges, U. and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R, *Journal of Statistical Software* **12**(3): 1–16.

Susmann, H. and Alkema, L. (2024). Flexibly modeling shocks to demographic and health indicators with Bayesian shrinkage priors. Available from: https://arxiv.org/abs/2410.09217.

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

Tierney, L. (1994). Markov chains for exploring posterior distributions, *The Annals of Statistics* **22**(4): 1701 – 1728.

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

United Nations (2015). The 17 goals. Retrieved 18.06.2025. Available from: https://sdgs.un.org/goals.

Van Niekerk, J., Krainski, E., Rustand, D. and Rue, H. (2023). A new avenue for Bayesian inference with INLA, *Computational Statistics & Data Analysis* **181**: 107692.

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

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11**: 3571–3594.

Appendix

A Additional GMRFs

This section will give an overview of different proposed GMRF models for different situations. Specifically, spatial, temporal and spatio-temporal models. These models will not be used for any analysis in this thesis, but they give context to the models that will be used.

A.1 Additional spatial examples

Other than the ICAR and the IID introduced in Section 2.5 there exists a plethora of spatial GMRF models. The examples given below all combine a structured and an unstructured random effect in some way, but this is just a subset of the available models. As in Section 2.5, the situation is still a discrete spatial setting with N regions.

BYM

The BYM model, also known as Besag-York-Mollié, combines the structured ICAR with the unstructured IID model (Besag *and others*, 1991). This means that the random effect $\mathbf{x} = (x_1, ..., x_N)^T$ is a combination of two random effects, specifically

$$\mathbf{x} = \boldsymbol{\phi} + \boldsymbol{\theta}$$

where $\boldsymbol{\phi}$ is an ICAR and $\boldsymbol{\theta}$ is the IID, both with dimension $(N \times 1)$. These both have an associated precision denoted τ_ϕ and τ_θ. Then the conditional distributions for x_i becomes

$$x_i \mid \mathbf{x}_{-i}, \tau_\phi, \tau_\theta \sim N\left(\frac{\sum_{j\sim i} x_j}{n_i}, (n_i\tau_\phi)^{-1} + \tau_\theta^{-1}\right) \quad i = 1, ..., N.$$

The conditional mean is the mean of the neighbours from the ICAR since the mean of the Gaussian IID is 0. The conditional variance sums the variance from the ICAR and the IID, so the IID handles overdispersion in the data. The BYM can also be defined by a joint density like

$$\mathbf{x} \mid \tau_\phi, \tau_\theta \sim N(\mathbf{0}, \tau_\phi^{-1}\mathbf{R}_\phi^- + \tau_\theta^{-1}\mathbf{I}) \tag{A.1}$$

where $\mathbf{R}_\phi$ is the structure matrix for the ICAR defined in Section 2.5. Overall, the idea behind the model is that the ICAR facilitates spatial smoothing and sharing of information across regions while the IID accounts for spatial heterogeneity. It will also vary between applications if the $\boldsymbol{\phi}$ or $\boldsymbol{\theta}$ is the dominating effect, or even if there is a dominating effect. Note that an increasing τ_θ indicates a lower presence of spatial heterogeneity while an increasing τ_ϕ indicates more spatial smoothing as the regions can differ less from the mean of their neighbours. However, the BYM model has a problem of identifiability, and the comparison of the effect of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ is often challenging. This is because the resulting effect is a combination of both the ICAR

and the IID, and it is unknown what which of the sub-models contributed. There exists parametrizations of the BYM model which tries to improve certain aspects of the model, and some of them will be introduced below, alongside some models inspired by the BYM.

Leroux

A new model from Leroux *and others* (2000) introduces a mixing parameter $\rho \in [0, 1]$, which represents the mixing of the precision between the ICAR and the IID. Note that this model can not be defined as a weighted sum of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, like the BYM could. Specifically, the distribution for $\mathbf{x}$ becomes

$$\mathbf{x} \mid \tau, \rho \sim N(\mathbf{0}, \tau^{-1}((1-\rho)\mathbf{I} + \rho\mathbf{R}_\phi)^{-1}) \tag{A.2}$$

where $\mathbf{R}_\phi$ is the structure matrix for the ICAR model. Note that $\rho = 0$ corresponds to the IID model while $\rho = 1$ gives the ICAR. Additionally, for $\rho = 1$ the inverse of the matrix sum must be exchanged with the generalized inverse as mentioned in Section 2.5. The conditional distributions for the x_i's are (Leroux *and others*, 2000)

$$x_i \mid \mathbf{x}_{-i}, \tau, \rho \sim N\left(\frac{\rho}{\rho n_i + 1 - \rho}\sum_{j\sim i} x_j, \frac{\tau^{-1}}{\rho n_i + 1 - \rho}\right), \quad i = 1, ..., N.$$

Note that the while the BYM model adds the variances of the two models, the Leroux instead adds the precisions before inverting to the overall variance. This creates some issues and for example makes it harder to scale the model (Riebler *and others*, 2016).

Dean

Like the Leroux model, this is also a mixing model and it was introduced in Dean *and others* (2001). The vector $\mathbf{x}$ is now defined as

$$\mathbf{x} = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\rho}\boldsymbol{\theta} + \sqrt{\rho}\boldsymbol{\phi}\right).$$

This gives the joint distribution for $\mathbf{x}$ as

$$\mathbf{x} \mid \tau, \rho \sim N(\mathbf{0}, \tau^{-1}((1-\rho)\mathbf{I} + \rho\mathbf{R}_\phi^-)).$$

This is very similar to the Leroux distribution in Equation (A.2) , but now the inverse is taken on each matrix individually, and not on the sum of them. Note that $\mathbf{I}^{-1} = \mathbf{I}$.

BYM2

The BYM2 model, introduced in Riebler *and others* (2016), builds on the Dean model to solve another problem. Namely the transferability of priors between different data and different neighbourhood structures. This requires some form of scaling of the spatial structure matrix $\mathbf{R}_\phi$. This is because the interpretation of the precision for $\boldsymbol{\phi}$ will vary based on the graph structure, and thus on $\mathbf{R}_\phi$. For example, the conditional variance for region 2 for an ICAR based on Figure 2.4 would be $\frac{1}{3\tau}$,

while for region 1 in the same graph would be $\frac{1}{2\tau}$, because the conditional variance is inversely proportional to the number of neighbours. Thus, graphs with a higher average number of neighbours have a lower conditional variance than the precision τ would imply. As such, the prior placed on τ will also have a different interpretation in the two cases. One way to remedy this is through the geometric variance, which is the geometric mean of the marginal variances associated with a structure matrix (Sørbye and Rue, 2014). The geometric variance is defined as follows

$$\sigma^2_{GV}(\mathbf{x}) = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{\tau}[\mathbf{R}^-_\phi]_{ii}\right)\right).$$

Ideally, this value would be the same for different matrices $\mathbf{R}_\phi$ corresponding to different graphs, but that is not the case (see Riebler *and others* (2016) and Sørbye and Rue (2014)). Again because of the different number of neighbours and how they are distributed. Furthermore, the geometric variance should satisfy $\sigma^2_{GV}(\mathbf{x}) = 1/\tau$, where τ is the precision parameter of the mixed model. This is to connect the interpretation of τ to the geometric mean of the marginal variance. So, to scale $\mathbf{R}_\phi$ appropriately, first the geometric variance when $\tau = 1$ is calculated and denoted as $\sigma^2_{ref}(\mathbf{x})$. Then, the scaled structure matrix $\mathbf{R}^*_\phi = \sigma^2_{ref}(\mathbf{x})\mathbf{R}_\phi$. The model can then be specified as

$$\mathbf{x} = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\rho}\boldsymbol{\theta} + \sqrt{\rho}\boldsymbol{\phi}^*\right)$$

where $\boldsymbol{\phi}^*$ is associated with the scaled structure matrix $\mathbf{R}^*_\phi$. This ensures that the geometric variance for the vector $\mathbf{x}$ equals $1/\tau$, and the interpretability of τ is directly tied to the geometric mean of the marginal variances. Additionally, the prior placed on τ is now transferable between different situations and graph-structures, as a result of the scaling.

A.2 Temporal examples

Modelling temporal data is another important use case for GRMFs, but not the primary interest of this thesis. Thus, some temporal models with differing properties will be introduced as they might be referenced in the thesis, and they can also provide some context to the spatial models. For example the connection between the ICAR in space and the RW1 in time. The examples introduced below will be GMRFs and IGMRFs as well as structured and unstructured random effects. The graph associated with the models introduced in this Section is shown in Figure A.1.

Figure A.1: Illustration of the graph corresponding to a temporal case with N timepoints.

First order random walk

The simplest temporal random effect is a first order random walk, or RW1. The idea is that timepoint t only depends on the previous timepoint $t-1$ and so on. Thus, it is a structured random effect. For a vector $\mathbf{x} = (x_1, ..., x_N)^T$ with known x_1, or with some prior for x_1, the following illustrates the idea:

$$x_t \mid x_{t-1}, \sigma \sim N(x_{t-1}, \sigma^2), \quad \text{for } t = 2, ..., N$$

for some standard deviation σ. This principle makes it straightforward to simulate first order random walks, see Figure A.3, but it is a bit simplified in the context of random effects. Specifically, it requires some restriction applied to x_1 and it does not represent the backwards dependence, namely that the value at time $t+1$ also affects the fitted valued at time t. Thus, the more relevant conditional distributions are

$$\begin{aligned} x_t \mid x_{t+1}, \sigma &\sim N(x_{t+1}, \sigma^2), && \text{for } t = 1 \\ x_t \mid x_{t-1}, x_{t+1}, \sigma &\sim N\left(\frac{x_{t-1} + x_{t+1}}{2}, \frac{\sigma^2}{2}\right), && \text{for } t = 2, ..., N-1 \\ x_t \mid x_{t-1}, \sigma &\sim N(x_{t-1}, \sigma^2), && \text{for } t = N. \end{aligned}$$

Assuming the mean is $\mathbf{0}$ the joint density can be expressed as follows:

$$\pi(\mathbf{x} \mid \sigma) = (2\pi)^{-(N-1)/2} (|\mathbf{Q}|^*)^{1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{x}\right).$$

The precision matrix $\mathbf{Q}$ above depends on the structure matrix $\mathbf{R}$, specifically $\mathbf{Q} = \frac{1}{\sigma^2}\mathbf{R} = \tau\mathbf{R}$. Note that the pseudo-determinant is used as $\mathbf{Q}$ and $\mathbf{R}$ have rank $N-1$ and are not invertible. Furthermore, the elements of $\mathbf{R}$ are defined as follows:

$$R_{ij} = \begin{cases} 1, & \text{if } i = j \in \{1, N\} \\ 2, & \text{if } i = j \notin \{1, N\} \\ -1, & \text{if } |i - j| = 1 \\ 0, & \text{else.} \end{cases}$$

It is clear that $\mathbf{Q}\mathbb{1} = \mathbf{0}$ and that a RW1 is an IGMRF.

Second order random walk

A natural extension of the first order random walk is the second order random walk, or RW2. Naturally, this is also a structured random effect. Now the previous two timepoints influence the current timepoint. Specifically, for a vector $\mathbf{x}$ with known x_1 and x_2, the conditional distributions are:

$$x_t \mid x_{t-1}, x_{t-2}, \sigma \sim N(2x_{t-1} - x_{t-2}, \sigma^2).$$

As for the RW1, a joint density can be defined where the only difference is the precision matrix $\mathbf{Q} = \tau\mathbf{R}$ with $\mathbf{R}$ defined as follows [chapter 3.4](Rue and Held,

2005):

$$R_{ij} = \begin{cases} 1, & \text{if } i = j \in \{1, N\} \\ 5, & \text{if } i = j \in \{2, N-1\} \\ 6, & \text{if } i = j \notin \{1, 2, N-1, N\} \\ 1, & \text{if } |i-j| = 2 \\ -2, & \text{if } |i-j| = 1 \text{ and } i \vee j \in \{1, N\} \\ -4, & \text{if } |i-j| = 1 \text{ and } i \,\&\, j \notin \{1, N\} \\ 0, & \text{else.} \end{cases}$$

Again, $\mathbf{Q}\mathbb{1} = \mathbf{0}$ and the RW2 is an IGMRF. However, note that the rank of $\mathbf{Q}$ is $N-2$, so the RW2 is a second order IGMRF. Realizations for the RW2 can be seen in Figure A.3.

First order autoregressive

Another extension of the first order RW1 is a first order autoregressive model, or AR1. The dependence structure mirrors that of the RW1, so again a structured random effect. However, the strength of dependence is now more flexible. Specifically, the conditional mean of x_t is no longer just x_{t-1}, but rather a scaled mean equal to ρx_{t-1}. Thus, the conditional density for a vector $\mathbf{x}$ with known x_1 becomes:

$$x_t \mid x_{t-1}, \sigma, \rho \sim N(\rho x_{t-1}, \sigma^2), \quad \text{for } t = 2, ..., N \tag{A.3}$$

for $\rho \in \mathbb{R}$. When $\rho = 1$ the AR1 equals the RW1, so the RW1 is a special case of the AR1. Usually, ρ is constrained to the interval $(-1, 1)$, and this is assumed to hold for the results below. Furthermore, the joint density compared to the RW1 and RW2 again only differs in the precision matrix $\mathbf{Q} = \tau\mathbf{R}$, which is a scaled version of the structure matrix $\mathbf{R}$ defined as follows (Allévius, 2018):

$$R_{ij} = \begin{cases} 1, & \text{if } i = j \in \{1, N\} \\ 1 + \rho^2, & \text{if } i = j \notin \{1, N\} \\ -\rho, & \text{if } |i-j| = 1 \\ 0, & \text{else.} \end{cases}$$

Note that the precision matrix $\mathbf{Q}$ for the AR1 with $\rho = 1$ does not equal the precision matrix for the RW1, even though the RW1 is a special case of the AR1. This is because $\rho = 1$ is outside the assumption for the ρ. It is also clear that $\mathbf{Q}$ has full rank, an the AR1 is a GMRF. Realizations for the AR1 can be seen in Figure A.3.

IID

A somewhat different model essentially captures white noise and is often denoted as an IID model. While the three previous examples were structured random effects, the iid model is instead unstructured, as all the elements are independent of each other. Additionally, the elements follow an identical distribution, limited to a Gaussian one in the context of GMRFs. Thus, the model can be defined as:

$$x_t \mid \sigma \sim N(0, \sigma^2), \quad \text{for } t = 1, ..., N.$$

This can again be summarised with a joint density with the same expression as the previous models, where now the precision matrix $\mathbf{Q} = \tau\mathbf{I}$. As stated, the purpose

of the IID is to capture temporal heterogeneity. An example can be seen in Figure A.2, which shows the fitted model for a first and seventh degree polynomial. The idea is that if the seventh degree polynomial was combined with an IID effect, it could reduce to the first degree polynomial and let the IID capture the remaining spatial heterogeneity. This is a bit simplified, but the idea is to let the IID capture noise, like sampling noise or adverse events, so the rest of the model can focus on modelling the underlying trend. This can also be applied to more complex situations than the given example.

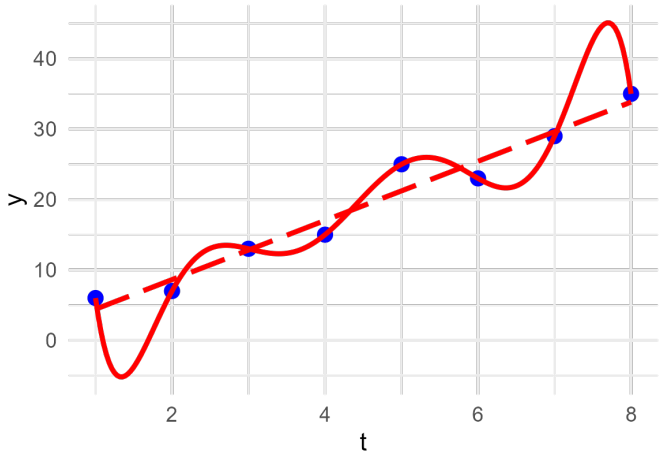


Figure A.2: The blue points are the data. The red dotted line is a linear regression while the red line is a polynomial of seventh degree.

Realizations of the IID can be seen in Figure A.3. Observe that the IID is the only model which is unstructured, which is apparent in Figure A.3. Furthermore, the RW2 is much smoother than both the RW1 and AR1, and it also reaches values much further from the start point of $y = 0$. Note that an AR1 with a ρ approaching 0 behaves similarly to the IID.

A.3 Spatio-temporal examples

Some modelling situations combines the spatial and temporal dimensions. In the field of epidemiology this could be modelling cases per region over time, for instance during the COVID-19 outbreak. In biology it could be used to model an animal population and their spatial distribution over time. The following examples of models will focus on discrete timepoints and discrete spatial units with a first order neighbourhood structure. Specifically, assume there are regions $i = 1, ..., I$ and timepoints $t = 1, ..., T$. When used in a Bayesian hierarchical model the latent layer could be

$$\eta_{it} = \mu + \alpha_t + \gamma_t + \phi_i + \theta_i \quad i = 1, ..., I \ \& \ t = 1, ..., T.$$

Here $\boldsymbol{\eta}$ is related to the response variable in some fashion, μ is an intercept and the remaining four variables are GMRFs, either in space or time. The unstructured

Figure A.3: Realizations of four temporal GMRFs. The x-axis indicate the timepoints while the y-axis represents the value of the GMRFs. The standard deviation $\sigma = 1$, number of timepoints $N = 50$ and the coefficient for the AR1 is $\rho = 0,8$, see (A.3). For the RW1, RW2 and AR1 x_1 is fixed to 0, for the RW2 x_2 is also set to 0.

effects are $\boldsymbol{\gamma}$ in time and $\boldsymbol{\theta}$ in space, typically some zero-centred iid Gaussian distribution while $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ are structured effects, for example a first order random walk in time and an ICAR in space. All of these random effects can be defined as GMRFs with an associated structure matrix $\mathbf{R}$. Then the random effects are assigned a GMRF prior as a multivariate normal with mean $\mathbf{0}$ and precision matrix $\tau\mathbf{R}$ for some precision τ. The structure matrices are $\mathbf{R}_{\boldsymbol{\gamma}} = \mathbf{I}_T$ and $\mathbf{R}_{\boldsymbol{\theta}} = \mathbf{I}_I$. As in Appendix A.2 the structure matrix for the first order random walk $\boldsymbol{\alpha}$ is:

$$\mathbf{R}_{\boldsymbol{\alpha}} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}.$$

For the ICAR $\boldsymbol{\phi}$ the structure matrix follows the definition in Section 2.5. The temporal matrices have dimensions $(T \times T)$ and the spatial matrices have dimension $(I \times I)$. So far all of the random effects only focus on either spatial or temporal effects, and the assumption is that these do not interact. This is not always an appropriate assumption, and thus interaction effects are of interest. This gives the following latent layer

$$\eta_{it} = \mu + \alpha_t + \gamma_t + \phi_i + \theta_i + \delta_{it} \quad i = 1, ..., I \ \& \ t = 1, ..., T,$$

for some space-time interaction effect $\boldsymbol{\delta} = (\delta_{1,1}, \delta_{1,2}, ..., \delta_{I,T-1}, \delta_{I,T})$ (Knorr-Held, 2000). Furthermore, $\boldsymbol{\delta}$ is also represented as a GMRF with a structure matrix $\mathbf{R}_{\boldsymbol{\delta}}$. Specifically, Knorr-Held (2000) proposes to use the Kronecker product of the structure matrices for a spatial and temporal effect, and gives four options for the model specified above.

Recall that the Kronecker product of two matrices creates a block structure where the second matrix is repeated and scaled for each element in the first matrix, specifically:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}.$$

It can be applied to matrices of any size, and if $\mathbf{A}$ is $(M \times N)$ and $\mathbf{B}$ is $(J \times K)$ the product has dimension $(MJ \times NK)$.

The first interaction, called **Type I** interaction, uses the product from $\boldsymbol{\gamma}$ in time and $\boldsymbol{\theta}$ in space, the two unstructured effects. Thus,

$$\mathbf{R}_{\boldsymbol{\delta}} = \mathbf{R}_{\boldsymbol{\gamma}} \otimes \mathbf{R}_{\boldsymbol{\theta}} = \mathbf{I}_T \otimes \mathbf{I}_I = \mathbf{I}_{T*I}$$

This means that the Type I interaction is an iid Gaussian with an independent element for each spatial region $i = 1, ..., I$ at every timepoint $t = 1, ..., T$. Clearly, this $\boldsymbol{\delta}$ is an unstructured random effect.

The **Type II** interaction instead combines a temporal structured effect with a spatial unstructured effect, namely $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. As an example consider a first order random

walk for timepoints $t = 1, 2, 3$ and an iid Gaussian for regions $i = 1, 2$. Then,

$$\mathbf{R}_{\boldsymbol{\alpha}} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{R}_{\boldsymbol{\theta}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The structure matrix for the $\boldsymbol{\delta}$ is then

$$\mathbf{R}_{\boldsymbol{\delta}} = \mathbf{R}_{\boldsymbol{\alpha}} \otimes \mathbf{R}_{\boldsymbol{\theta}} = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ -1 & 0 & 2 & 0 & -1 & 0 \\ 0 & -1 & 0 & 2 & 0 & -1 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

for $\boldsymbol{\delta} = (\delta_{11}, \delta_{21}, \delta_{12}, \delta_{22}, \delta_{13}, \delta_{23})$. Note that the rows $1, 3$ and 5 only have elements in columns $1, 3$ and 5, so clearly $\mathbf{R}_{\boldsymbol{\delta}}$ is not of full rank. Actually, the matrix can be split into two smaller matrices, one for each region i. For region $i = 1$ the structure matrix is

$$\mathbf{R}_{\delta_1} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

which is a temporal first order random walk. Region $i = 2$ will have the same reduced structure matrix. Therefore, the Type II interaction between a temporal RW1 and a spatial IID creates an independent temporal RW1 for each region with the same precision parameter τ. The rank of $\mathbf{R}_{\boldsymbol{\delta}}$ is $I(T-1)$ and a Type II interaction is appropriate if the temporal trends are region specific with no spatial structure (Knorr-Held, 2000).

The **Type III** interaction combines a temporal unstructured effect $\boldsymbol{\gamma}$ with a spatial structured effect $\boldsymbol{\phi}$. Thus,

$$\mathbf{R}_{\boldsymbol{\delta}} = \mathbf{R}_{\boldsymbol{\gamma}} \otimes \mathbf{R}_{\boldsymbol{\phi}} = \mathbf{I}_T \otimes \mathbf{R}_{\boldsymbol{\phi}}.$$

This results in an identity block matrix where each block on the diagonal is the ICAR structure matrix $\mathbf{R}_{\boldsymbol{\phi}}$. Thus $\mathbf{R}_{\boldsymbol{\delta}}$ is not of full rank, and the rank in this case is $(I-1)T$. The interpretation of the $\boldsymbol{\delta}$ is that the random effect is an independent ICAR at each time-point $t = 1, ..., T$. There is no connection between following years and this interaction type is suitable when spatial trends vary over time with no temporal structure.

The final interaction type, denoted as **Type IV**, combines the structured temporal effect $\boldsymbol{\alpha}$ and the structured spatial effect $\boldsymbol{\phi}$. The combined structure matrix becomes

$$\mathbf{R}_{\boldsymbol{\delta}} = \mathbf{R}_{\boldsymbol{\phi}} = \begin{bmatrix} \mathbf{R}_{\boldsymbol{\phi}} & -\mathbf{R}_{\boldsymbol{\phi}} & 0 & \cdots & 0 \\ -\mathbf{R}_{\boldsymbol{\phi}} & 2\mathbf{R}_{\boldsymbol{\phi}} & -\mathbf{R}_{\boldsymbol{\phi}} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\mathbf{R}_{\boldsymbol{\phi}} & 2\mathbf{R}_{\boldsymbol{\phi}} & -\mathbf{R}_{\boldsymbol{\phi}} \\ 0 & \dots & 0 & -\mathbf{R}_{\boldsymbol{\phi}} & \mathbf{R}_{\boldsymbol{\phi}} \end{bmatrix}.$$

The rank of $\mathbf{R}_{\boldsymbol{\delta}}$ is now $(I-1)(T-1)$ and the structure is more involved than the previous examples. This can be seen both in the joint distribution

$$p(\boldsymbol{\delta} \mid \tau_{\boldsymbol{\delta}}) \propto \exp\left(-\frac{\tau_{\boldsymbol{\delta}}}{2}\sum_{t=2}^{T}\sum_{i\sim j}(\delta_{it}-\delta_{jt}-\delta_{i,t-1}+\delta_{j,t-1})^2\right)$$

and the conditional distributions with mean

$$E[\delta_{it} \mid \boldsymbol{\delta}_{-it}, \tau_{\boldsymbol{\delta}}] = \begin{cases} \delta_{i,t+1} + \frac{1}{n_i}\sum_{j\sim i}\delta_{jt} - \delta_{j,t+1}, & t=1 \\ \delta_{i,t-1} + \frac{1}{n_i}\sum_{j\sim i}\delta_{jt} - \delta_{j,t-1}, & t=T \\ \frac{1}{2}(\delta_{i,t-1}+\delta_{i,t+1}) + \frac{1}{n_i}\sum_{j\sim i}\delta_{jt} - \frac{1}{2}(\delta_{j,t-1}+\delta_{j,t+1}), & t=2,...,T-1 \end{cases}$$

and precision

$$\tau_{it} = \begin{cases} n_i\tau_{\boldsymbol{\delta}}, & t=1 \text{ or } t=T \\ 2n_i\tau_{\boldsymbol{\delta}}, & t=2,...,T-1 \end{cases}$$

from Knorr-Held (2000). From the conditional mean it is clear that some second order neighbours directly influence a given δ_{it}. Specifically, temporal neighbours of the spatial neighbours of δ_{it} are present in the conditional mean. However, there are no second order neighbours purely temporally or spatially. Effectively, now temporal trends loan strength from adjacent regions and spatial trends loan strength across time. Thus, this is a suitable choice when temporal trends for adjacent regions are expected to share similarities and when spatial trends follow a temporal pattern (Knorr-Held, 2000). However, as the rank is $(I-1)(T-1)$, this rank-deficiency is typically rather high. This means that a similar number of restrictions must be imposed on the model[Chapter 3](Rue and Held, 2005). Another consequence is the high number of parameters which makes the models computationally expensive, especially when I and T are large.

Even though the RW1 and the ICAR was used as structured random effects in the examples above, they can be interchanged with any desired structured random effects. Similarly, the Gaussian IID models can be swapped with other unstructured models, for example independent models with varying precision for each element in the model. Could also use IID models with different distributions than Gaussian, but then it is no longer a GMRF.

B Temporal adaptive GMRFs

This subsection introduces two temporal adaptive GMRFs. One model for when it is assumed the shocked timepoints are known, for example an analysis of previous years when there was a war during some of the period. This is an unpractical assumption when predicting ahead in time. A second model assumes that there are shocks, but that they are at unknown timepoints.

Model for data with known shocks

For data where the timepoints of shocks are known, Aleshin-Guendel and Wakefield (2024) proposed a temporal model that incorporates this prior knowledge. This model is analogous to the BW-ICAR introduced in Section 3.2, just in time instead of space. This model splits all the timepoints in two groups, namely shocked and non-shocked. Let the set of shocked points be denoted as $\mathcal{S}$. The structure of the model closely mimics a RW1. However, there are now two precision parameters τ_S and τ_{US}. Specifically, the shocked τ_S is used for all transitions including at least one shocked timepoint. Formally, the conditional distributions for each x_i is

$$x_t|\mathbf{x}_{-t}, \tau_S, \tau_{US} \sim N\left(\frac{\sum_{j\sim t}\tau_{tj}x_j}{\sum_{j\sim t}\tau_{tj}}, \left(\sum_{j\sim t}\tau_{tj}\right)^{-1}\right) \tag{B.4}$$

with

$$\tau_{ij} = \begin{cases} \tau_S, & i \vee j \in \mathcal{S} \;\&\; |i-j| = 1 \\ \tau_{US}, & i \;\&\; j \notin \mathcal{S} \;\&\; |i-j| = 1. \end{cases} \tag{B.5}$$

Note that $t \sim j$ coincides with $|t-j| = 1$ in the temporal case. Effectively, (B.4) and (B.5) means that there are now two levels of spatial smoothing, and the expectation is that $\tau_S < \tau_{US}$, so that there is less spatial smoothing for adjacent timepoints where at least one of them is shocked. Defining the weight matrix is quite similar to the distributions for the transitions, and the result is

$$W_{ij} = \begin{cases} \tau_S, & i \vee j \in \mathcal{S} \;\&\; |i-j| = 1 \\ \tau_{US}, & i \;\&\; j \notin \mathcal{S} \;\&\; |i-j| = 1 \\ 0, & \text{else.} \end{cases}$$

Then $\mathbf{D}$ is defined as in (2.9) and the precision matrix $\mathbf{Q} = \mathbf{D} - \mathbf{W}$.

Model for data with unknown shocks

An adaptive GMRF for temporal data with unknown shocks was proposed in Susmann and Alkema (2024). In the article they focus on a univariate setting where the goal is to produce estimates for future male life expectancy at birth. The idea is that current techniques, which do not account for shocks, overestimates the uncertainty and thus get unnecessarily wide credibility intervals. Instead, the authors propose to incorporate a shocked latent effect with a Bayesian shrinkage prior, known as the regularized horseshoe prior (Piironen and Vehtari, 2017). This ensures that the shock term only activates when there is a shock in the data.

The standard model is as follows. For the true life expectancy $\boldsymbol{\eta}$, a smoothing function f which is fit as a B-spline (for details see Susmann and Alkema (2024)) and some Gaussian iid noise $\boldsymbol{\epsilon}$ the model is:

$$\eta_t = \eta_{t-1} + f(\eta_{t-1}, \boldsymbol{\beta}) + \epsilon_t.$$

Here t is a timepoint. The function $f(\eta_{t-1}, \boldsymbol{\beta})$ models the expected change from timepoint $t-1$ to t through a B-spline. The Gaussian noise $\epsilon_t \mid \sigma_\epsilon \sim N(0, \sigma_\epsilon^2)$ captures the deviations from the expected value at time t.

The shocked model introduces the effect $\boldsymbol{\delta}$ to account for unknown shocks in the data. The model can now be written as

$$\eta_t = \eta_{t-1} + f(\eta_{t-1}, \boldsymbol{\beta}) + \epsilon_t - \delta_t.$$

The delta is constrained to be positive so that it only captures negative shocks, as short term positive shocks are unnatural for life expectancy. Furthermore, it is included at every time-point with a shrinkage prior to assert that it is zero for most time-points. Specifically, the regularized horseshoe prior for δ_t is given as:

$$\begin{aligned}
\delta_t \mid \gamma_t, \sigma, v &\sim N^+(0, \sigma^2 \tilde{\gamma}_t^2) \\
\tilde{\gamma}_t^2 &= \frac{v^2 \gamma_t^2}{v^2 + \sigma^2 \gamma_t^2} \\
\gamma_t &\sim C^+(0, 1).
\end{aligned}$$

The second part is responsible for the regularization and the parameter v is crucial. The idea is that when δ_t is large the prior above converges to a $N(0, v^2)$. Thus, v can be assigned an informative prior if the magnitude of expected shocks are known beforehand. Additionally, in contrast to the prior proposed by Piironen and Vehtari (2017) the Gaussian distribution is limited to positive values since $\delta_t > 0$ for all t. Even though the given model is designed for modelling life expectancy, it is simple to modify it to suite other temporal data of interest. That could mean switching to only positive shocks, or maybe including both negative and positive shocks.

C Sensitivity analysis for BW-ICAR

This illustrates the sensitivity analysis for the BW-ICAR. Specifically, the four prior choices are

1. $\tau \sim PC(1, 0.01)$
2. $\tau \sim PC(3, 0.05)$
3. $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$
4. $\tau \sim gamma(1, 0.00005)$.

Then the BW-ICAR defined in Equation (5.1) is trained with each of the four priors and for all the 86 diseases. Three criteria are used, WAIC, DIC and LS. For all the results the model with $\tau \sim PC(3, 0.05)$ is used as a reference for each disease. Thus, positive values in the three Figures indicates that the reference prior performed better, while negative values indicates otherwise. It seems like $\tau \sim PC(3, 0.05)$ and $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$ performed the best for WAIC in Figure C.1 and DIC in Figure C.2, although $\tau \sim PC(1, 0.01)$ performs competitively. In terms of LS in Figure C.3 the $\tau \sim PC(1, 0.01)$ performed the best. As LS will be prioritized over DIC and WAIC, $\tau \sim PC(1, 0.01)$ was chosen, and will be used in the full validation study.

The model specified in Equation (5.1) will be used with the chosen prior, as well as three other priors.

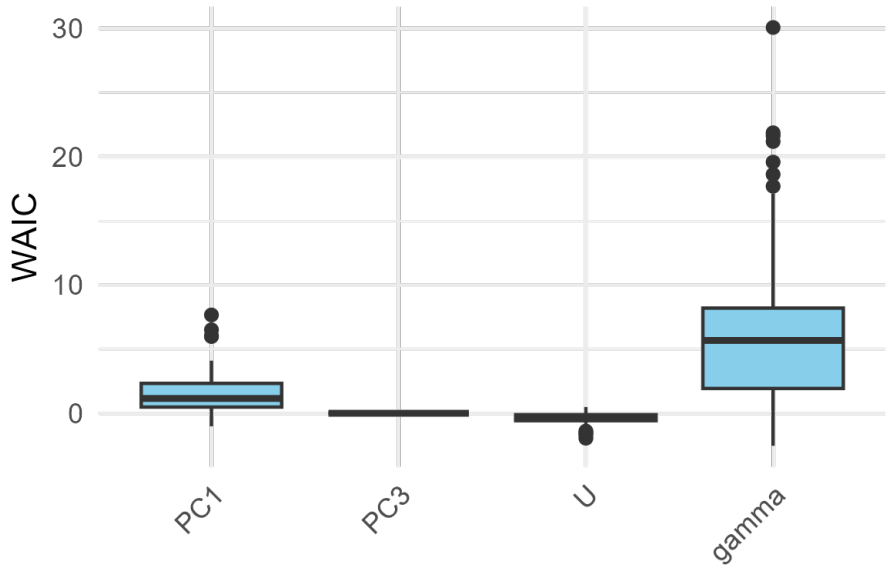


Figure C.1: WAIC for all 86 diseases for the BW-ICAR with $PC(3, 0.05)$ as the reference level.

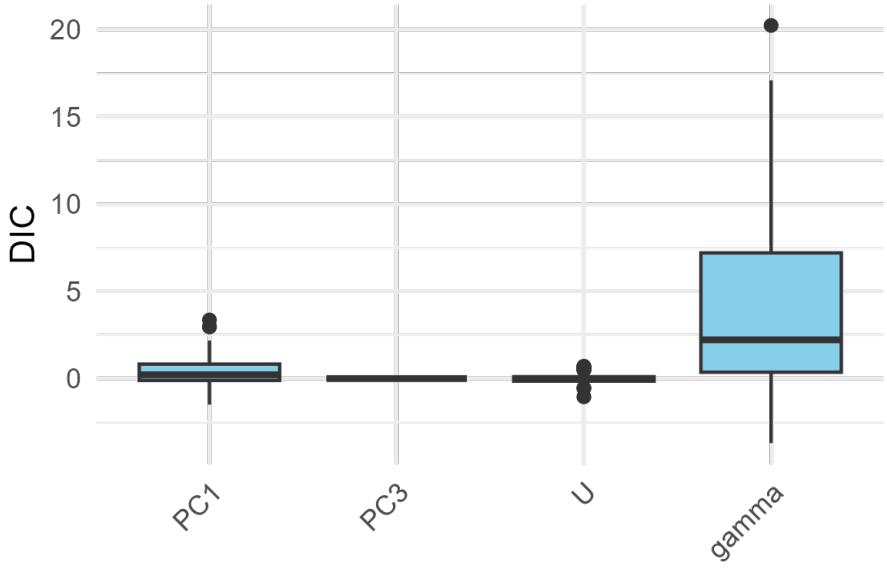


Figure C.2: DIC for all 86 diseases for the BW-ICAR with $PC(3, 0.05)$ as the reference level.

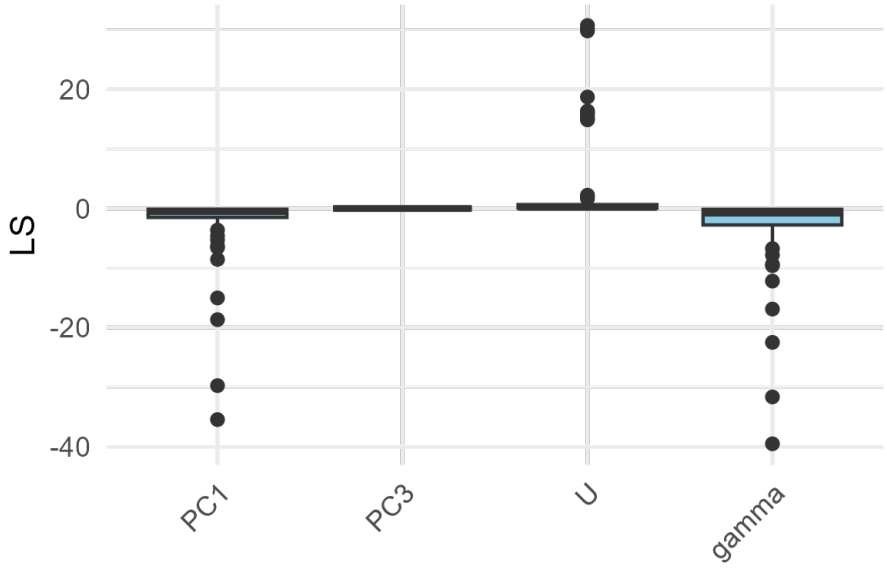


Figure C.3: LS for all 86 diseases for the BW-ICAR with $PC(3, 0.05)$ as the reference level.

D Illustrative examples with INLA

This section uses a random disease from the validation procedure, in this case disease 60. It would be the same with simulated data or any other data. The plan is to compare slightly different model definitions and arguments in INLA and test in practice if they give the same results. The initial step of collecting the data is then

```
Y <- as.vector(ObservedCases[, 60])
E <- as.vector(ExpectedCases[, 60])

data_df <- data.frame(Y = Y, E = E, ID = 1:nAreas)
```

The offset for the expected cases

There are two primary ways of including an offset, one method includes the log of the offset in the INLA formula, so

```
formula_ICAR <- Y ~ offset(log(E)) +
      f(ID, model = "besag", graph = adj_matrix, scale.model = TRUE,
      hyper = list(prec = list(prior = "pc.prec", param = c(1,
      ↪  0.01))))

res_ICAR <- inla(formula_ICAR, family = "poisson", data = data_df)
```

The other method is to include the offset in the INLA call, and specifically include the argument `E = E` where the second E is whatever you have called your offset in the code. In total,

```
formula_ICAR_2 <- Y ~ f(ID, model = "besag", graph = adj_matrix,
  ↪  scale.model = TRUE,
      hyper = list(prec = list(prior = "pc.prec", param = c(1,
      ↪  0.01))))

res_ICAR_2 <- inla(formula_ICAR_2, family = "poisson", data =
  ↪  data_df, E = E)
```

When looking at their summaries, these two models perfectly align and the two methods of including the offset E are equivalent in terms of the resulting inference.

Sum-to-zero constraints for the ICAR

Enforcing the sum-to-zero constraint can be done in a few different ways. For the ICAR, called `besag` in INLA, the argument `constr = TRUE` is used by default, which includes a sum-to-zero constraint. For other models, this is not always available and it is common to add it as
`extraconstr = list(A = matrix(1, nrow = 1, ncol = N), e = 0)` for an effect of length N. The two different methods are showed below

```
formula_ICAR <- Y ~ offset(log(E)) +
      f(ID, model = "besag", graph = adj_matrix, scale.model = TRUE,
       ↪  constr = F,
        extraconstr = list(A = matrix(1, nrow = 1, ncol = nAreas), e
         ↪  = 0),
      hyper = list(prec = list(prior = "pc.prec", param = c(1,
       ↪  0.01))))

res_ICAR <- inla(formula_ICAR, family = "poisson", data = data_df,
                  control.compute = list(cpo = T, waic = T, dic =
                   ↪  T))

#versus

formula_ICAR_2 <- Y ~ offset(log(E)) +
  f(ID, model = "besag", graph = adj_matrix, scale.model = TRUE,
   ↪  constr = T,
      hyper = list(prec = list(prior = "pc.prec", param = c(1,
       ↪  0.01))))

res_ICAR_2 <- inla(formula_ICAR_2, family = "poisson", data =
 ↪  data_df,
                  control.compute = list(cpo = T, waic = T, dic =
                   ↪  T))
```

However, they do not yield the same inference, either for the model criteria like DIC and WAIC, or for the fitted fixed and random effects. The already implemented constraint has been chosen as it is likely more sophisticated than the manual one, even though the rest of the models will use the manual one, as that is the only option available. The specific results are

```
summary(res_ICAR)
Time used:
    Pre = 0.569, Running = 0.526, Post = 0.192, Total = 1.29
Fixed effects:
             mean    sd 0.025quant 0.5quant 0.975quant  mode kld
(Intercept) -0.07 0.034     -0.138    -0.07     -0.003 -0.07   0

Random effects:
  Name          Model
    ID Besags ICAR model

Model hyperparameters:
                  mean    sd 0.025quant 0.5quant 0.975quant  mode
Precision for ID 27.49 23.16       8.06    21.10      83.97 14.99

Deviance Information Criterion (DIC) ...............: 287.93
Deviance Information Criterion (DIC, saturated) ....: 70.09
Effective number of parameters .....................: 18.93
```

```
19
20 Watanabe-Akaike information criterion (WAIC) ...: 290.26
21 Effective number of parameters .................: 16.61
22
23 Marginal log-Likelihood:  -160.90
24 CPO, PIT is computed
25 Posterior summaries for the linear predictor and the fitted values
   ↪ are computed
26 (Posterior marginals needs also
   ↪ 'control.compute=list(return.marginals.predictor=TRUE)')
27
28 > summary(res_ICAR_2)
29 Time used:
30     Pre = 0.561, Running = 0.54, Post = 0.167, Total = 1.27
31 Fixed effects:
32               mean    sd 0.025quant 0.5quant 0.975quant   mode kld
33 (Intercept) -0.069 0.034     -0.136   -0.068     -0.002 -0.068   0
34
35 Random effects:
36   Name          Model
37     ID Besags ICAR model
38
39 Model hyperparameters:
40                   mean    sd 0.025quant 0.5quant 0.975quant  mode
41 Precision for ID 36.65 39.62       9.08    25.32     128.42 16.88
42
43 Deviance Information Criterion (DIC) ...............: 288.79
44 Deviance Information Criterion (DIC, saturated) ....: 70.95
45 Effective number of parameters .....................: 17.84
46
47 Watanabe-Akaike information criterion (WAIC) ...: 292.05
48 Effective number of parameters .................: 16.64
49
50 Marginal log-Likelihood:  -160.23
51 CPO, PIT is computed
52 Posterior summaries for the linear predictor and the fitted values
   ↪ are computed
53 (Posterior marginals needs also
   ↪ 'control.compute=list(return.marginals.predictor=TRUE)')
```

E Additional plots from the validation study

This section includes some Figures and Tables from the validation study, both from the training stage and the validation stage. The section is primarily a supplement to the results in Section 6.

E.1 Figures and tables from the training

Additional Figures from the training stage. First is a Table with all of the region specific weights **c** for the different RW-ICARs with the associated standard deviation from the MCMC simulations.

Province	n=10	n=20	n=50	n=86	Cancer
Álava	0.79 (0.83)	1.15 (1.02)	1.40 (1.18)	5.82 (5.14)	1.08 (0.96)
Albacete	0.95 (0.85)	0.88 (0.69)	0.99 (0.72)	3.59 (2.33)	0.97 (0.72)
Alicante	1.06 (0.98)	0.78 (0.69)	1.65 (1.21)	7.72 (5.76)	0.95 (0.80)
Almería	1.71 (1.38)	1.54 (1.18)	3.05 (2.20)	19.52 (13.06)	1.67 (1.09)
Ávila	0.91 (0.80)	1.06 (0.94)	1.12 (0.79)	9.21 (6.67)	0.96 (0.72)
Badajoz	1.37 (1.17)	1.53 (1.20)	1.93 (1.53)	28.94 (21.45)	1.28 (1.10)
Barcelona	1.21 (1.04)	1.08 (0.96)	1.93 (1.57)	7.78 (5.96)	1.12 (0.80)
Burgos	1.38 (0.99)	1.76 (1.24)	2.15 (1.32)	5.65 (3.58)	0.92 (0.71)
Cáceres	1.55 (1.25)	1.35 (1.04)	2.41 (1.59)	6.37 (4.07)	1.54 (1.06)
Cádiz	1.15 (1.00)	1.46 (1.13)	2.15 (1.55)	9.85 (8.24)	1.52 (1.10)
Castellón	1.17 (1.08)	0.99 (0.80)	1.50 (1.08)	9.57 (7.79)	1.18 (0.88)
Ciudad Real	0.35 (0.37)	0.45 (0.45)	0.83 (0.67)	1.57 (1.10)	0.57 (0.46)
Córdoba	1.57 (1.34)	1.34 (1.06)	1.35 (0.95)	1.50 (1.11)	1.15 (0.93)
La Coruña	1.69 (1.33)	2.13 (1.39)	2.96 (1.75)	23.91 (15.39)	2.12 (1.39)
Cuenca	0.80 (0.87)	0.98 (0.91)	2.06 (1.69)	18.53 (15.71)	1.15 (0.99)
Gerona	1.48 (1.19)	1.62 (1.22)	2.56 (1.60)	26.81 (16.74)	1.81 (1.32)
Granada	1.31 (1.14)	1.55 (1.15)	2.80 (1.81)	26.38 (15.92)	1.44 (0.93)
Guadalajara	0.98 (0.91)	1.28 (1.20)	2.19 (1.80)	27.52 (20.36)	1.23 (1.07)
Guipúzcoa	1.10 (0.89)	0.89 (0.80)	1.39 (1.04)	4.02 (2.76)	1.10 (0.83)
Huelva	1.30 (1.12)	1.31 (1.11)	2.30 (1.66)	9.54 (8.36)	1.04 (0.80)
Huesca	1.19 (0.97)	1.55 (1.21)	1.51 (1.02)	11.82 (7.42)	1.70 (1.13)
Jaén	0.92 (0.92)	1.07 (0.98)	0.32 (0.34)	1.68 (1.42)	0.79 (0.76)
León	1.26 (1.11)	1.12 (0.93)	0.52 (0.53)	0.74 (0.52)	1.39 (1.05)
Lleida	1.60 (1.37)	1.51 (1.19)	3.07 (2.10)	4.34 (3.92)	1.62 (1.29)
La Rioja	1.22 (1.14)	0.66 (0.72)	1.43 (1.08)	5.47 (3.79)	0.74 (0.64)
Lugo	1.49 (1.22)	2.01 (1.44)	2.91 (1.85)	20.36 (14.11)	2.06 (1.41)
Madrid	1.10 (1.09)	1.54 (1.26)	1.10 (0.93)	2.22 (2.19)	1.48 (1.16)
Málaga	1.58 (1.32)	1.18 (0.98)	2.74 (1.89)	34.77 (24.25)	1.31 (1.10)
Murcia	0.98 (0.90)	0.88 (0.69)	0.43 (0.28)	1.68 (1.06)	0.76 (0.59)
Navarra	0.93 (0.82)	1.20 (0.92)	1.74 (1.23)	25.13 (13.53)	1.47 (1.01)
Orense	1.38 (1.16)	1.62 (1.18)	2.34 (1.51)	3.28 (2.31)	1.84 (1.27)
Asturias	1.33 (1.10)	1.53 (1.17)	2.27 (1.63)	9.67 (7.24)	1.50 (1.12)
Palencia	1.33 (1.07)	1.54 (1.25)	0.66 (0.57)	2.41 (1.92)	1.36 (1.11)

Province	n=10	n=20	n=50	n=86	Cancer
Pontevedra	1.61 (1.32)	1.94 (1.50)	2.91 (2.08)	15.26 (13.45)	2.05 (1.52)
Salamanca	1.69 (1.41)	1.98 (1.31)	2.10 (1.43)	6.88 (4.97)	2.02 (1.44)
Cantabria	1.34 (1.18)	1.33 (1.12)	2.44 (1.82)	24.48 (17.49)	1.50 (1.11)
Segovia	1.11 (1.05)	1.03 (0.90)	0.52 (0.45)	1.58 (1.36)	1.10 (0.92)
Sevilla	1.55 (1.21)	1.50 (1.21)	2.38 (1.99)	15.18 (14.52)	1.48 (1.24)
Soria	0.61 (0.61)	0.64 (0.53)	0.44 (0.28)	3.17 (2.10)	0.92 (0.73)
Tarragona	1.23 (1.07)	1.25 (1.01)	3.46 (2.56)	34.35 (23.33)	1.46 (1.18)
Teruel	1.02 (0.90)	1.29 (1.04)	1.90 (1.38)	5.87 (6.06)	1.60 (1.16)
Toledo	1.15 (1.03)	0.86 (0.83)	1.18 (1.07)	4.98 (5.29)	0.88 (0.80)
Valencia	1.24 (1.19)	1.39 (1.14)	2.64 (1.87)	23.27 (16.94)	1.52 (1.15)
Valladolid	0.82 (0.77)	1.52 (1.16)	0.87 (0.54)	4.60 (2.86)	1.13 (0.85)
Vizcaya	1.41 (1.18)	1.55 (1.14)	1.17 (0.90)	9.69 (7.90)	1.65 (1.18)
Zamora	1.28 (1.14)	1.43 (1.17)	1.41 (1.15)	31.74 (22.46)	1.55 (1.19)
Zaragoza	1.15 (1.01)	1.22 (0.93)	1.99 (1.36)	1.04 (0.75)	1.37 (1.07)

Table E.1: The weights **c** for each of the subsets of diseases for the RW-ICAR. The standard deviation is shown in brackets.

Next are some Tables for running the multivariate models when choosing the random diseases subsets for the different n with a different seed. This is done to increase the certainty regarding which results are tied to the specific models, or to the chosen random disease subset.

	RW-ICAR	EW-ICAR
n=10	0.108	0.065
n=20	0.002	0.206
n=50	0.400	0.527
n=86	0.256	0.808

Table E.2: p-values for the comparison of the log mean posterior precision of edges inside or across autonomous regions when using the second seed.

	n=10	n=20	n=50	n=86
n=10	1.00	0.81	0.57	0.38
n=20	0.81	1.00	0.58	0.40
n=50	0.57	0.58	1.00	0.70
n=86	0.38	0.40	0.70	1.00

Table E.3: Correlation between the estimated edge weights τ_{ij} for different values of n for the RW-ICAR when using the second seed.

E.2 Figures and tables from the validation

This section shows the validation box plots for all the diseases with the outliers shown, i.e. points outside the whiskers of the boxplots. These Figures are referenced

	n=10	n=20	n=50	n=86
n=10	1.00	0.70	0.40	0.23
n=20	0.70	1.00	0.45	0.31
n=50	0.40	0.45	1.00	0.79
n=86	0.23	0.31	0.79	1.00

Table E.4: Correlation between the estimated edge weights τ_{ij} for different values of n for the EW-ICAR when using the second seed.

in Section 6.

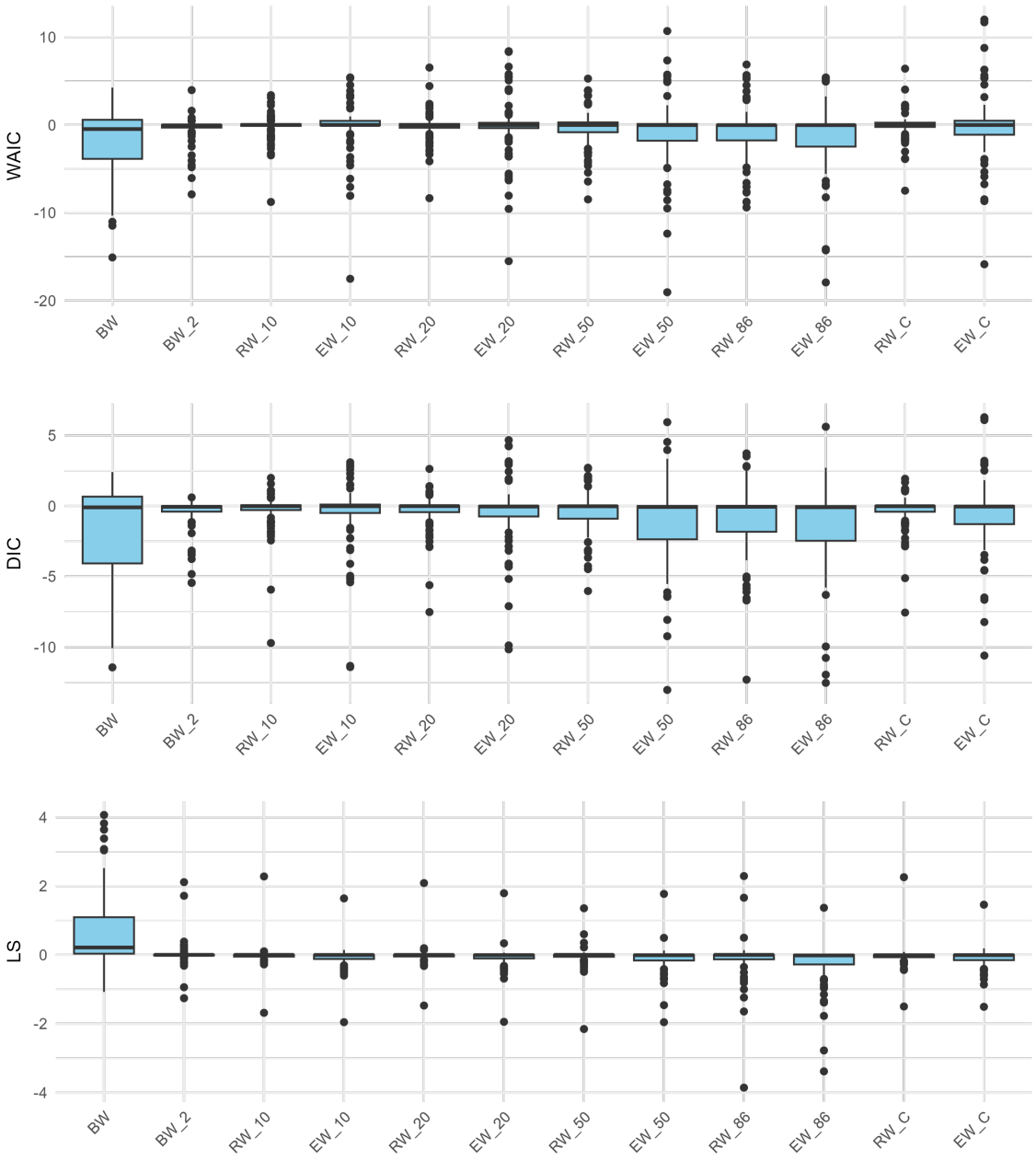


Figure E.1: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease with $\tau \sim gamma(0, 0.00005)$. Values lower than zero indicate that the model outperformed the ICAR, which is used as a reference. The black dots are individual outliers corresponding to a given disease.

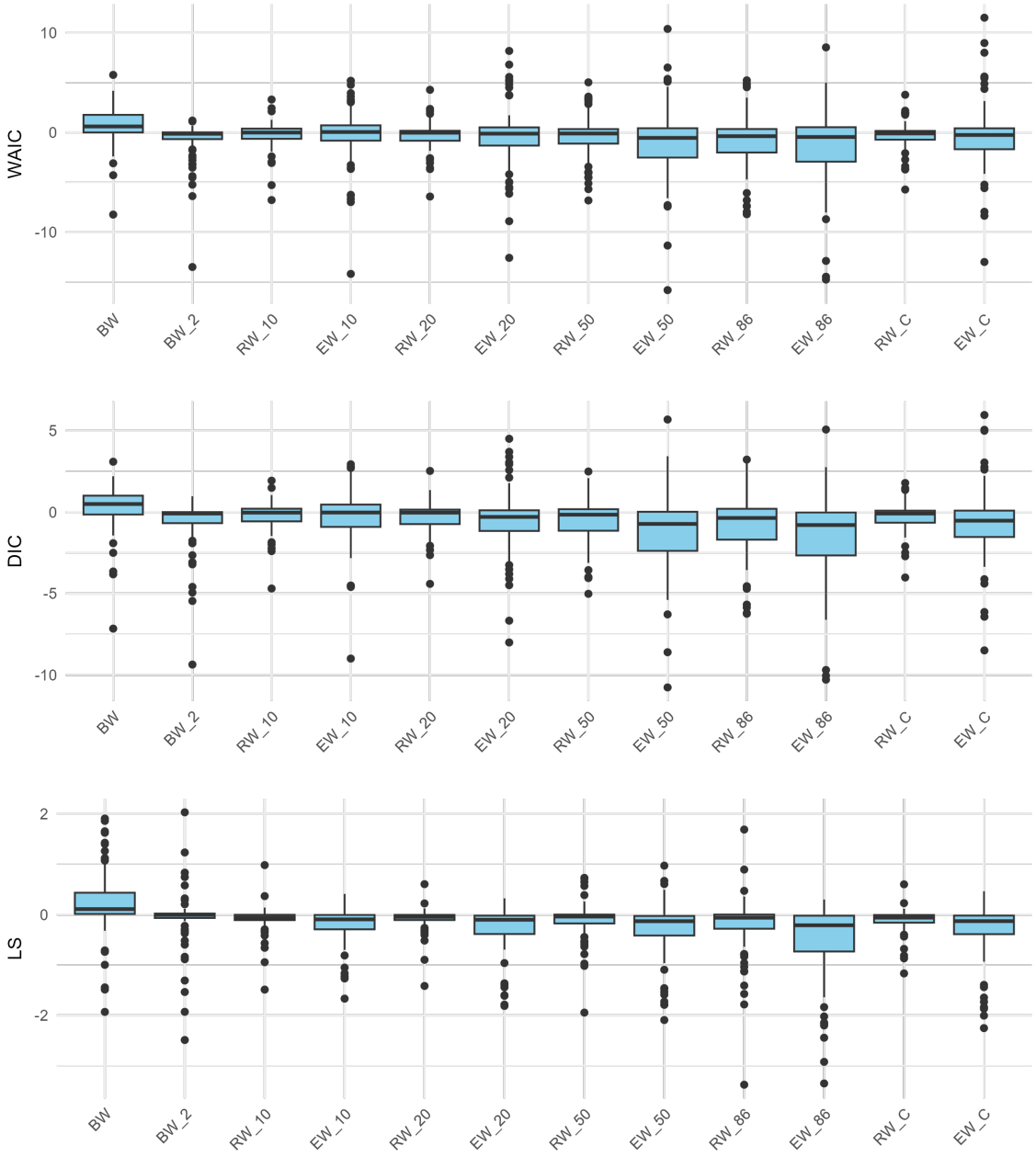


Figure E.2: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease with a $\tau \sim PC(1, 0.01)$. Values lower than zero indicate that the model outperformed the ICAR, which is used as a reference. The black dots are individual outliers corresponding to a given disease.

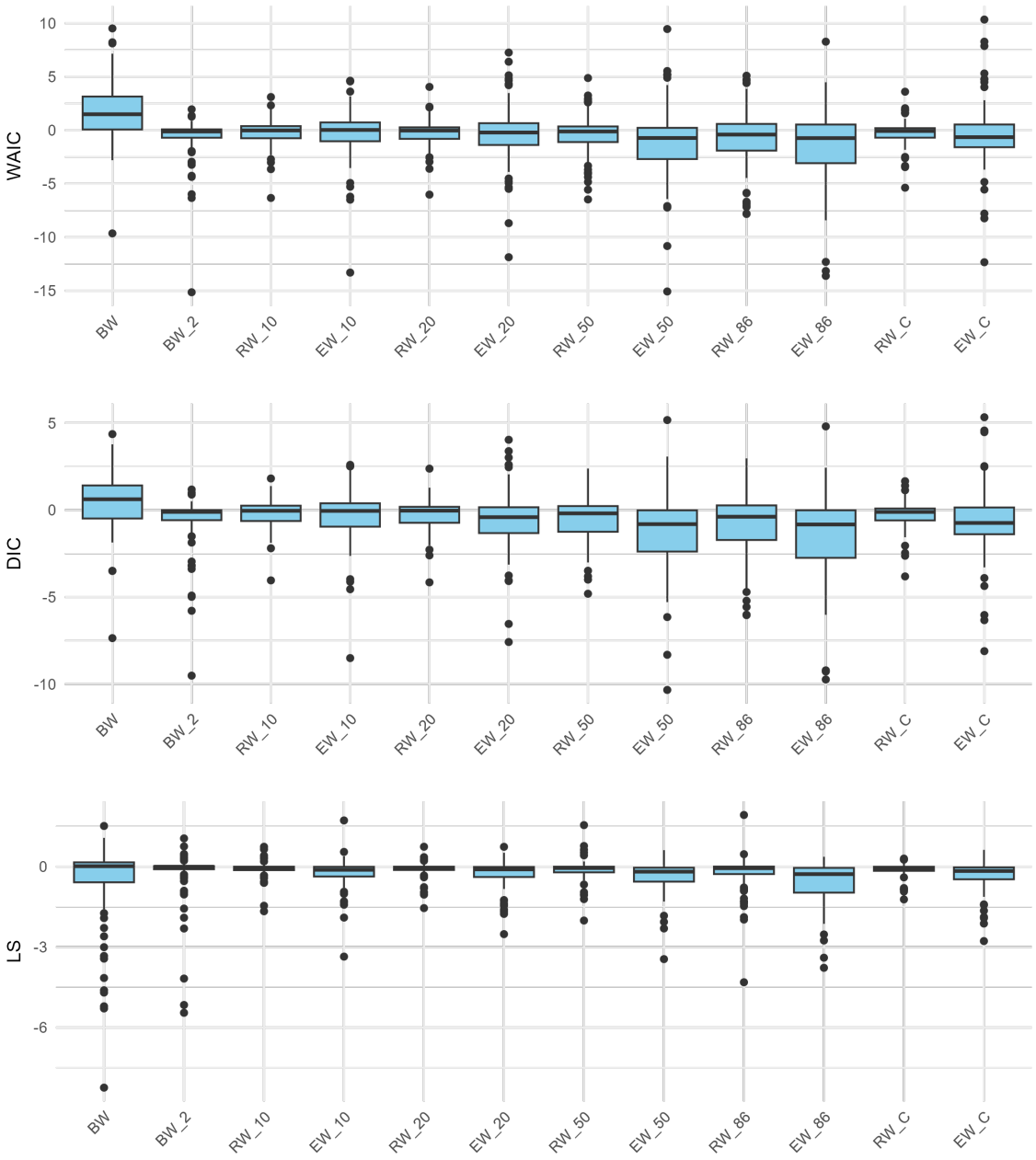
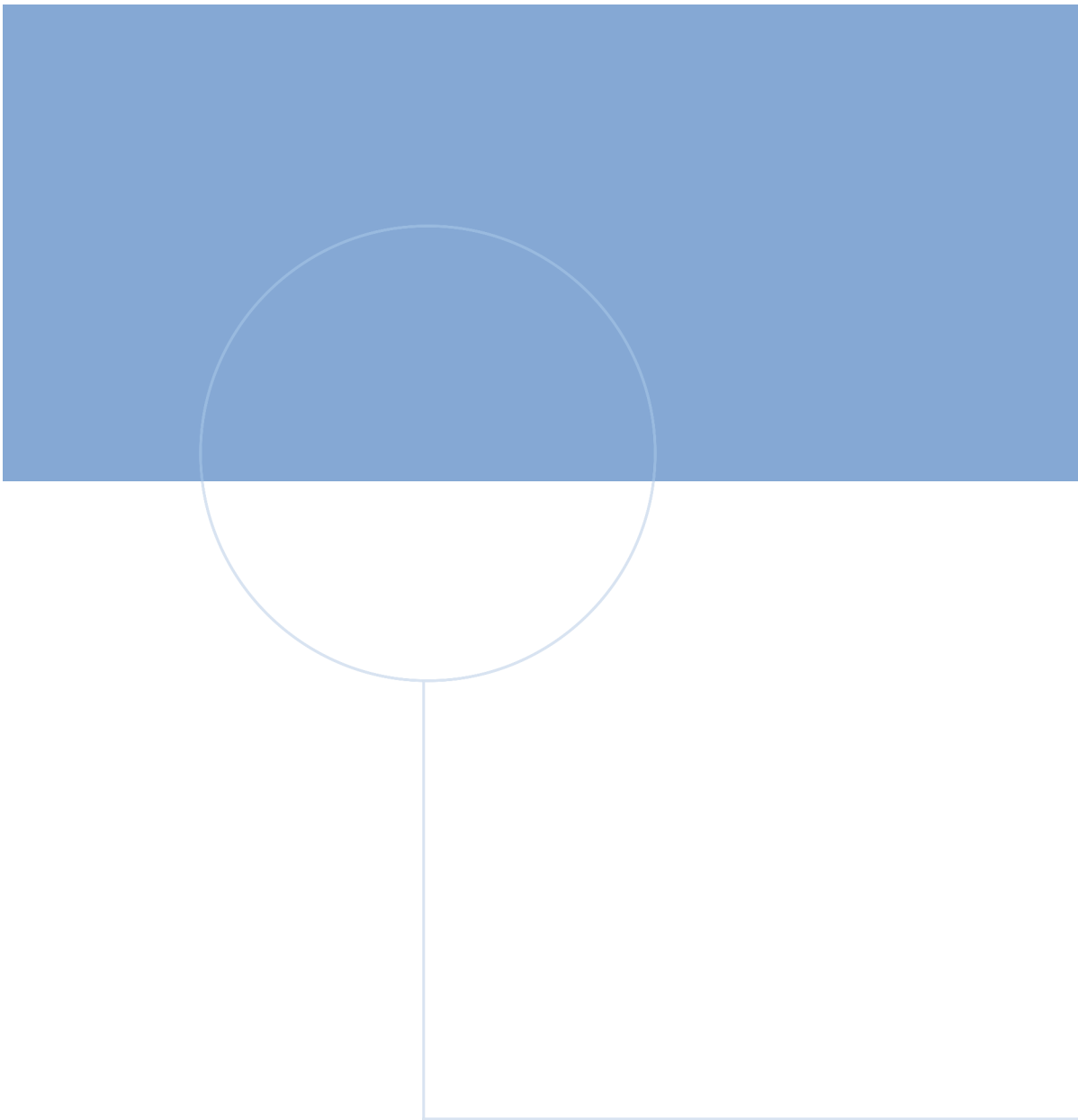


Figure E.3: Box-plots for three evaluation criteria when using the fitted matrix for the given model on the x-axis and refitting for each individual disease with the prior $\frac{1}{\sqrt{\tau}} \sim U(0, 5)$. Values lower than zero indicate that the model outperformed the ICAR, which is used as a reference. The black dots are individual outliers corresponding to a given disease.



NTNU

Norwegian University of
Science and Technology