Spatial modeling with INLA for analysis of unequal care in Skåne

Julia Wierzchoslawska

Bachelor's thesis
2021:K31

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

Acknowledgements

An immense thank you to my supervisor Johan Lindström for making this thesis possible and guiding me through it with patience.

Wilmer, Melinda, Paulina, J.B.P, Mom, Dad, Miriam and Bernard, thank you.

To my Goose.

Abstract

The objective of this thesis is to extend on a previous analysis of health care accessibility for patients diagnosed with a chronic disease in Region Skåne. The previous analysis resulted in a logistic mixed effects model having municipality as a random effect and age as a first-degree spline-function. This thesis extends on the random effects from the previous model in order to analyse the spatial dependencies on municipal and postal-code spatial levels. The models being compared are Bayesian structured additive regression models with latent Gaussian Markov Random Fields. The spatial dependencies are modeled using a Conditional Autoregressive model, and a Random Walk is used to approximate a spline-function for age in this framework. To perform approximate Bayesian inference Integrated Nested Laplace Approximation (INLA) is used. It is shown that both on a municipal and postal-code level a Random Walk of order two is preferred for approximating the spline-function. The difference lies in the spatial dependencies, where on municipal level modeling them as i.i.d. is sufficient, which is comparable to the previous analysis. Regarding spatial dependencies with more intricate geographic boarders, such as on the postal-code level, modeling using a Conditional Autoregressive model is preferred.

Contents

Introduction

There are many ethical ideas that governments stand for and proper implementation of these beliefs is crucial to the society's identity. A common example that many countries wish to achieve is equal health care for their population. A solution to effectively analysing a certain "system" and pinpointing weaknesses can be done through statistical modeling.

Equal health care for all is a concept that is deeply rooted in Swedish culture but achieving this is not straight forward. One way that Sweden seeks to solve this issue is by delegating to regions the responsibility of analysing the health care structure for possible improvements. Region Skåne has particular interest in analysing how care for those diagnosed with a chronic disease is distributed throughout the region. To do so they have resorted to a more complex analysis involving statistical modeling, compared to previous more simplistic analysis (pairwise comparison). Previous work on this topic has been done by Lundgren (2021) in partnership with Region Skåne. This thesis will expand further on Lundgrens final model.

1.1 Chronic disease & Pr1Fys Indicator

A substantial part of the population suffers from chronic illness. For these individuals, getting the health care they need can insure a longer and healthier life. Region Skåne is looking into the factors of why some individuals living in Skåne who suffer from a chronic illness are not getting the care they need. Pinpointing these factors could help the region achieve the goal of equal care for all. The approach being used to measure if individuals with a chronic disease are receiving needed care is by the Pr1Fys indicator, described as:

"Number of listed patients who received a diagnosis from one of the diagnosis groups heart failure, coronary heart disease, TIA / stroke, COPD, diabetes, dementia and / or atrial fibrillation 19-60 months ago and who were on a physical return visit to a doctor or nurse, or received a home visit, the last 18 months with a diagnosis from the same diagnostic group as 19-60 months ago." (Primärvårdskvalitet)

The data used in this analysis is specifically from 2013-01-03 to 2016-06-03. The indicator shows whether a patient in the group has had a medical care revisit within the past 18 months relative to the specific date of data collection, 2017-12-31. In addition to the indicator and data on the patients, postal-code level data describing socio-economic indicators is available, which is referred to as the Care Need Index.

This data will be used as possible explanatory variables for any differences in health care throughout Region Skåne.

1.2 Previous Results

This thesis is a continuation of Lundgren (2021) thesis in which he suggests and compares statistical models to assess variability in availability of health care for the chronically ill in Skåne. The best fitting model for the study was a Generalized Linear Mixed Model. Specifically, a logistic mixed effect model with age variable modeled as a spline function and municipalities in the region modeled as independent random effects.

1.3 Spatial Analysis with INLA

In this thesis, analysis of the the random effects in Lundgren's model is carried out. A limitation in his final model, is the assumption of independent effects of the municipalities, potentially ignoring the spatial relation between nearby municipalities.

The analysis will be done on two spatial levels. The first is on a municipal level which is the same as in previous analysis, giving insight oriented on the health care availability and providers. The second is on a postal-code level, which is a more socio-economic oriented analysis.

Here the model will be expanded to include spatial dependence of the random effects through the use of the conditional autoregressive model. To do so, the model is specified as a latent Gaussian model with assumption of conditional independence inducing a Gaussian Markov random field where this expansion can be incorporated. Additionally in the GMRF framework the random walk is used as an approximation of the Spline function for the age variable.

We will perform approximate Bayesian inference on this model, which falls into the subclass of structured additive regression models, latent gaussian models through the method of Integrated Nested Laplace Approximation (INLA) given by Rue et al. (2009).

Data

The data is provided by Region Skåne, compiled from their own database (RSV), Skånes population register (SBR), and from statistics Sweden (SCB). The data does not contain any information that could give rise to ethical concern, as all patients are anonymous.

The purpose of collection of the data is to analyze whether sufficient care is being provided by the healthcare system to persons who have been diagnosed with a chronic disease. The data contains factors that may play a role in patients receiving or not the care that is presumed to be necessary for their diagnosis. Criteria for the sample is discussed in section 2.1, and modification of the data is described in section 2.3. The manner in which "necessary care" is defined for the chronically ill by Region Skåne is based upon the Pr1Fys indicator discussed in section 2.2.

2.1 Sample criterion

The data contains diagnostic groups which are affiliated with a specific code and each patient is part of one or multiple diagnostic groups. The choice for patients being used in the sample follow the subsequent criteria:

1. Patient must have been diagnosed between 2013-01-03 and 2016-06-03 with one or more of the following diseases :
    - Heart failure
    - Coronary heart disease
    - Diabetes (diabetes mellitus type 2, other specified diabetes mellitus, unspecified diabetes mellitus)
    - Chronic Obstructive Pulmonary Disease (COPD)
    - Transient Ischemic Attack (TIA) and/or Stroke
2. Patients must be registered at a Vårdcentral (healthcare center) in Skåne on the date 2017-12-31.
3. Patient must have affiliation between their postal code and the Care Need Index (CNI) (found in Statistics Sweden's database). The CNI is a socioeconomic index composed of the weighted sum of the following variables:
    - Age of 65 or older and single

- Foreign born (Eastern Europe, Asia, Africa and South America)
- Unemployed between the age of 16-65
- Single parent with children of age 17 or less
- Individuals of age 1 or older who have moved to the area
- Individuals of age 25-64 with low level of education
- Individuals of less than 5 years of age

4. Patient must have postal code corresponding to the municipality of residence, such information should be stored both in the Listing database and in the poulation register.

2.2 The Indicator

A quality care indicator *"Pr1Fys"* (*Primärvårdskvalitet*) is being used since it provides a basis of how patients with a chronic disease are prioritized. This indicator is based upon medical revisits, including home visits of individuals who have one or more chronic disease such as, heart failure, coronary heart disease, diabetes, COPD, stroke/TIA, and/or Atrial fibrillation; who were diagnosed 19-60 months prior to a specific date.

Note: By definition the *"Pr1Fys"* indicator also includes dementia as a chronic disease, this variable is not included in the given data set used for this analysis.

In the data set that is provided the indicator shows whether a patient that has been diagnosed between 2013-01-03 and 2016-06-03 has had a medical care revisit within the past 18 months relative to the specific date of data collection, 2017-12-31.

The reason for choosing the time frame of 19-60 months is due to the nature of chronic diseases, and that patients will most definitely be needing care in that period. The reason for choosing a revisit within 18 months and not 12 is due to annual visits being rescheduled, allowing for a 6 month period for postponed visits. We assume individuals who have not had a visit within 18 month as not getting their annual medical visit, and therefore not getting the care they should be receiving for a person with their diagnosis.

What the Pr1Fys indicator considers to be a revisit for a patient is the following, a physical visit including a home visit. The visit is valid if done by a doctor or a nurse of any kind. Furthermore, the diagnostic codes and patients are linked to the indication of having a visit, as described above.

2.3 Data Modification

When handed the data it had been modified to prepare it for analysis by Lundgren (2021) who previously worked on modeling with use of this data set. The unmodified data includes 137,343 observations. A total of 1785 patients were removed from the original data set due to postal-codes being unavailable from (SBR) and/or CNI being unavailable from (SCB). Due to this we have a remaining 135,558 observations left.

Further reduction occurs when associating the data with the corresponding shapefiles describing corresponding locations. A shapefile stores nontopological geometry and attribute information for the spatial features in a data set (ESRI, 1998).

Two different shapefiles were used for the analysis, one being the municipalities of Skåne and the other being the postal-code areas of Skåne. Additional observations were omitted when joining the data to the shapefiles. Observations located either in areas which border Skåne and overlap into the neighboring regions of Halland, Småland and Blekinge or in areas which are located in these regions but overlap into Skåne are automatically omitted when using the joining function in R.

When data is joined to the municipality shapefile, there is a reduction of 15 observations, leaving a final 135,543 observations for modeling. When data is joined to the postal-code shapefile, there is a reduction of 644 observation, leaving a final 134,914 observations for modeling. The reason for there being more observations omitted on a postal-code level is due to it being divided into more intricate areas. Since individuals who are associated with in two different postal-code areas is more common than individuals being registered in two different municipalities, there are more observations omitted in the postal-code data.

Theory

The first section of this chapter, section 3.1, contains a brief introduction of Generalized Linear Mixed Models (GLMM), and the models that belong to this class used in previous analysis by Lundgren (2021). Further in section 3.2 an overview of GMRF is given, and theory on the spatial and temporal dependence components of these models. The INLA method which is the inference used for these models is presented in section 3.3. Lastly, model selection is described in section 3.4.

3.1 Model

The following sections 3.1.1-3.1.6 go through brief theory of GLM's, LLM's, GLMM's, Logistic regression, Poisson regression, and Negative binomial regression. These are the models and theory used in (Lundgren, 2021), which play a major part in the analysis.

3.1.1 Generalized Linear Model

General Linear Models (GLM) are an expansion of the Linear regression Model (LM). We will start with a brief introduction to LM as a basis to the GLM. In a multiple linear regression the response variable Y_i, having observations $i = 1, 2, ...n$ is continuous and restricted to having a normal distribution $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ with different means μ_i and constant variance σ^2.
In this case the linear predictor η_i is simply the expectation of the response variable $E(Y_i) = \mu_i = \eta_i$, shown below η_i is a linear combination of the independent variables x_{ij}'s and parameters β_j with $j = 1, 2, ...k$. Lastly $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the error term having zero expectation and constant variance, this term accounts for all random variation not explained by the linear model; this component will be altered in more complex models.

$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i \tag{3.1}$$

Continuing with the Generalized Linear Model, the distribution restriction on the response variable, Y_i, are changed to encompass the class of exponential family distributions which are continuous or discrete.

$$Y_i \sim \textit{exponential family}$$

Some examples are the Exponential, Poisson and Negative binomial regression. The linear prediction η_i in a GLM is related to the response variable Y_i by the link function $g(\cdot)$. The link function is typically a function of the expectation of the response variable or a function of the corresponding odds,

$$g(E(Y_i)) = g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}. \tag{3.2}$$

The Generalized linear model allows us to model discrete data, such as binary data and count data, or positive data using e.g. gamma distribution. Whereas data in linear regression is assumed to be continuous. Binary data can only take two values, which indicate whether an event has occurred or not occurred, respectively. Binary data follows a Bernoulli distribution. On the other hand count data is made up of non-negative integers $0, 1, 2, 3...$, it indicates the number of times an event has occurred in a certain time period. Count data can be modeled as e.g. Binomial, Poisson, Negative Binomial. Further details on GLM and related models can be found in McCulloch and Searle (2001).

3.1.2 Linear Mixed Model

Linear Mixed Models (LMM) are also an expansion of basic LM and can partly be expressed as in equation (3.3). LMM and LM share the characteristics of outcome variable having Normal distribution and being modeled by fixed effects, i.e. the β parameters. What differs in LMM is that certain parameters are treated as random effects to account for dependence in the data.

The LMM general model, as in the following equation is defined using matrix notation:

$$\boldsymbol{E(Y \mid u) = X\beta + Zu + \epsilon} \tag{3.3}$$

where he fixed effect component $\boldsymbol{\beta}$ is defined by a column vector of regression coefficients, and related to the response variable through matrix $\boldsymbol{X}$. The random effects component $\boldsymbol{u}$ is as a random vector which gives conditional assumption on response variable, which can be seen as $\boldsymbol{E(Y \mid U = u)}$, and it is related to the response variable through the design matrix $\boldsymbol{Z}$. The last component is the residual error $\boldsymbol{\epsilon}$ which is assigned to be normaly distributed, with expectation zero and constant variance σ^2.

The Random effects $\boldsymbol{u}$ are not parameters, rather their variance component is the parameter. In a simple case all the random effects have constant variance e.g. $\boldsymbol{u} \sim N(0, \sigma_u^2)$. In more complex cases, the random effects can be structured into groups with each group having its own variance, e.g. $\boldsymbol{u} \sim \mathcal{N}(0, \sigma_{u_j}^2)$ where u_j, $j = 1, 2, ..., p$ represents a specific group. LMM is used when our data is dependent with correlations between observations, it allows for analysis of ”groupings” in the data.

3.1.3 Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMM) are a combination of the GLM and LMM models. GLMMs allow the response variable to have a distribution from the exponential family of distributions. Thus the use of the link function $g(\cdot)$ is needed to provide a relation between the outcome variable and the linear prediction. As in LMM, the GLMM incorporates random effects to account for the correlation in the data, allowing the study of variance between these groups of dependencies. The predicted response in GLMM depends on the fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{u}$, which are treated as random variables. The parameters are the fixed effects $\boldsymbol{\beta}$ and the variance components of the random effects $\sigma^2_{w_j}$.

$$\boldsymbol{g(E(Y \mid u)) = X\beta + Zu} \tag{3.4}$$

Equation 3.4 shows the structure of a GLMM in a matrix notation. For further reading and idea of application on GLMs, LMMs and GLMMs can be found in McCulloch and Searle (2001).

3.1.4 Logistic regression

Logistic regression falls into the class of GLMs and it is the most popular model for binary data (Agresti, 2006). The response variable is assumed to be Bernoulli distributed $Y_i \sim Be(p_i)$, indicating whether an event has occurred or not by 1 and 0 respectively, with probability described as:

$$P(Y = y_i) = \begin{cases} p_i & \text{if } x = 1 \\ \\ (1 - p_i) & \text{if } x = 0. \end{cases}$$

Expectation of the outcome variable is the probability of the event occurring $E(Y_i) = p_i = \mu_i$ and the variance is the probability of the event occurring multiplied by the probability of the event not occurring $V(Y_i) = p_i(1 - p_i)$.

Logistic regression uses the natural log odds of the event happening. The link function $g(\cdot)$ is referred to as the *logit* function

$$g(\mu_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{3.5}$$

The linear predictor is given by

$$\mu_i = p_i = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 - e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}. \tag{3.6}$$

Note: The response variable is distributed as Bernoulli, before analysis the variables of use are aggregated in terms of the response variable to attain a Binomial distribution for the modeling, i.e. $Y_i \sim Be(n_i, p_i)$. For each aggregated group i, n_i is the number of observations that have the exact same covariate values, and Y_i is the total number of revisits in that aggregation group.

3.1.5 Poisson Regression

Another form of regression used to model count data is the Poisson regression. The response variable Y_i is assumed to have a Poisson distribution, with λ_i indicating the likeliness of an event occurring in a fixed time interval, s.t. $Y_i \sim Po(\lambda_i)$. The observation variables Y_i are count data, the number of events. For Poisson regression the link function is simply the natural log of the expected value of the response variable:

$$\ln(\lambda_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \qquad \lambda = e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}, \tag{3.7}$$

In the Poisson distribution there is the assumption that the expectation and the variance are equal

$$E(Y) = V(Y) = \lambda$$

It is often the case that this assumption is not fulfilled. This is the case when the data is overdispersed, meaning that the variance is larger than the expectation. To deal with this problem negative binomial regression can be used instead.

3.1.6 Negative Binomial Regression

The negative binomial is a generalization of the Poisson distribution for overdispersed data. This is accounted for by an extra parameter referred to as the dispersion parameter θ.

Allowing each observation Y_i to have its own Poisson mean $z_i\mu_i$ where z_i is a random multiplicative factor. All the means are distributed randomly around a common mean which is based on the x_i's, $\mu_i = e^{x_i\beta}$. The multiplicative factors $Z_i \sim \Gamma(\theta, \frac{1}{\theta})$ are Gamma distributed with $\theta > 0$. Consequently as θ tends to infinity the variance for Z_i tends to zero, the variance for Negative binomial as seen in (3.8) then tends to μ_i and the distribution for Y_i can be seen as Poisson since the assumption of expectation and variance being equal is satisfied.

$$E(Y_i) = \mu_i, \qquad V(Y_i) = \mu_i + \frac{1}{\theta} \cdot \mu_i^2, \qquad \theta > 0 \tag{3.8}$$

On the other hand if θ is small then the variability becomes large compared to μ_i. Here Y_i is dependant upon the Z_i, $(Y_i \mid Z_i = z_i) \sim Po(z_i\mu_i)$. the variance can freely be modeled and we have a solution to our problem. The link function for the Negative binomial, as for the Poisson, is the natural log of the expected value of the response variable. To look further into Poisson and Negative Binomial see Cameron and Trivedi (2013).

3.2 Gaussian Markov Random Fields

In section 3.2.2 GMRF are explained. This is the class of models that INLA fits. Additionally in this section the importance of the precision matrix and Markov property is explained. In section 3.2.3, the precision matrix and Markov property is used when modeling Random Walks for correlated random effects, and when modeling spatial dependence in section 3.2.5 using the Conditional Autoregressive model (CAR). In section 3.2.6 Bayesian hierarchical models are briefly introduced, and finally in section 3.2.7 the structure of the model for this analysis is specified.

3.2.1 Independent Random effects

Random effects are used to take into account variation produced by variables that are not the primary focus of the study. The Random effects will have a common Gaussian prior, with a mean $\mu = 0$ and a precision parameter τ to which a prior will be assigned.

Independent identically distributed (i.i.d.) Gaussian random effects are the simplest way to account for unstructured variability in the data. For GLMM we have the model is as shown in equation (3.4). When the random effect is said to be i.i.d. the u's will all share the same variance. Lets assume the u's are divided into groups w_i where $i = 1, 2, ..j$, depending on certain characteristics that they share. The assumption i.i.d. implies no correlation between the w_i groups, and that they have the same variance.

3.2.2 Gaussian Markov Random Fields

The types of models that INLA fits can be expressed as latent Gaussian Markov Random Fields (GMRF) with exponential family observations. Their relation to structured additive regression models will be shown in section 3.2.7.

Gaussian Random Fields consist of a random vector $\boldsymbol{x}$ having a multivariate Gaussian distribution :

$$\pi(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\sum\right|^{\frac{1}{2}}} \cdot exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T {\sum}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \quad \boldsymbol{x} \in \mathbb{R}^n \tag{3.9}$$

The density of a multivariate random variable $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sum)$ has following components, $\boldsymbol{x} = (x_1, ..., x_n)^T, n < \infty$, with mean vector $\boldsymbol{\mu}$ and symmetric positive definite covariance matrix $\sum$, which describes the dependence between the elements in $\boldsymbol{x}$.

An example that follows this process naturally is the Autoregressive process, AR(1), it is a simple temporal GRF of order 1. The joint densities can be expressed as in (3.10). Which can also be expressed as the product of conditional distributions $\prod_{i=1}^{n} \pi(x_i \mid x_{i-1})$ for $i = 1, 2, ...n$, specifying the conditional marginal distributions.

If the GRF has Markovian properties, allowing for the assumption of conditional independence, it is called a GMRF. Equation (3.11) is a result of the Markov property with the definition

$$p(x_i \mid \boldsymbol{x}_{-i}) = p(x_i \mid x_j \ ; \ j \in \mathcal{N}_i). \tag{3.10}$$

The Markov property is that the conditional distribution only depends on a small set of neighbouring points. As seen here the condition goes from $\boldsymbol{x}_{-i}$ to only neighbours x_j where $j \in \mathcal{N}_i$

This property is a key role in the efficiency of the INLA method for inference (Rue et al., 2009). The efficiency of the method is based on the sparsity of the precision matrix $\boldsymbol{Q} = \sum^{-1}$, is the inverse of the covariance matrix. The Markov property gives a sparse precision matrix. The implication of this property can be expressed in the following theorem, shows the importance of the Markov assumption in the GMRF to obtain a sparse precision matrix aiding in efficient computations.

Theorem 2.2 Let $\boldsymbol{x}$ be normally distributed with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q}$. Then for $i \neq j$,

$$j \notin \mathcal{N}_i \iff \textstyle\sum_{ij}^{-1} = Q_{ij} = 0 \tag{3.11}$$

The theorem affirms that if two points are not neighbours then, $Q_{ij} = 0$. The sparseness of Q is inherited over to the Cholesky factorization (due to the Markovian property), and is used to optimize computational efficiency. The theorem and extensive reading on the topic of GMRF can be found in Rue and Held (2005) .

3.2.3 Correlated random effects: Random Walk

The model in Lundgren (2021) has the non-linear dependence for the age variable modeled as a Spline function of order one. In this section a Random Walk approximation of the spline that uses GMRFs is shown.

Random Walk of order 1 & 2

Random Walks (RW) describe a curve in time or space, it is a non-linear function specifying temporal correlation. A Random Walk model can be used to approximate a Spline function. A Smoothing Spline (S-Spline) is defined by choosing f that minimizes the penalized least squares criterion (Wang et al., 2018),

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int (f^{(m)}(x))^2 dx. \tag{3.12}$$

In the equation above the m'th derivative of f is denoted as $f^{(m)}$, the sum measures the "closeness" to the data while the integral penalizes rapid change in the function giving a smooth fit, and the smoothing parameter λ creates a "compromise" between the sum and the integral.

To approximate the penalty function, a random walk prior on f is used. If the observations are defined as $x_1 < x_2 < .... < x_n$, and d is a constant representing the distance between each observation. Then the integral can be approximated as

$$\int (f^{(m)}(x_i))^2 dx \approx d^{-(2m-1)} \sum_{i=m+1}^{n} [\nabla^m f(x_i)]^2. \tag{3.13}$$

The penalty function is approximated with ∇^m, which is the m'th order backwards operator.

$$\nabla^1 f(x_i) = f(x_i) - f(x_{i-1}), \;\; \nabla^2 f(x_i) = f(x_i) - 2f(x_{i-1}) + f(x_{i-2}) \tag{3.14}$$

∇^1 and ∇^2 define a Random walk of order one (RW1) and two (RW2) respectively. Where each step difference is defined as normal and independent:

$$\nabla^m f(x_i) \overset{iid}{\sim} \mathcal{N}(0, \sigma_f^2), \qquad i = 1 + m, ..., n.$$

In the RW1 $f(x_i)$ is only conditionally dependant on the first order neighbours and independent of all other observation. Similarly for RW2 where $f(x_i)$ is conditionally dependant only on the first and second order neighbours, thus the random walks being a Markov property, which gives rise to a sparse precision matrix, allowing for fast Bayesian computations.

3.2.4 Spline : For temporal dependence

The Spline function can be used to obtain a better fit for the model in a regression, when there is an assumption of non-linear dependence between variables. In Lundgren (2021) paper a B-spline of order 1 is used to model the age variable. There we will instead use the RW as described above, using the random walk as a prior we can write the model as

$$Y_i \in N(f(x_i) \mid \sigma^2), \qquad \nabla^m f(x_i) \sim N(0, \sigma_f^2). \tag{3.15}$$

Which results in a log-posterior distribution.

$$\begin{aligned} &\sum_{i=1}^{n} \frac{(Y_i - f(x_i))^2}{2\sigma^2} + \frac{1}{2\sigma_f^2} \sum_{i=1}^{n} (\nabla^m f(x_i))^2 \\ &= \frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} (Y_i - f(x_i))^2 + \frac{\sigma^2}{\sigma_f^2} \sum_{i=1}^{n} (\nabla^m f(x_i))^2 \right). \end{aligned} \tag{3.16}$$

3.2.5 CAR : For spatial dependence

Conditional Autoregressive model (CAR) is used when there is first order spatial dependency, the is the case when a specified geographic area is assumed to be affected by the neighboring areas. As seen previously the AR(1) model was used for temporal dependencies of order 1, whereas the CAR(1) model is used for spatial dependencies of order 1. These two models are similar in how the Markov property is applied. The CAR(1) model with random vector $\boldsymbol{x} \sim \boldsymbol{N}(0, Q_{car}^{-1})$ is defined as Besag et al. (1991).

$$x_i \mid \{x_j : j \sim \mathcal{N}_i\} \sim \mathrm{N}\left(\frac{1}{\kappa^2 + \|\mathcal{N}_i\|} \sum_{j \in \mathcal{N}_i} x_j, \frac{1}{\tau(\kappa^2 + \|\mathcal{N}_i\|)}\right) \tag{3.17}$$

To model areal data with a CAR(1) model, the GMRF is defined as a collection of random variables $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$. Each of the random variables representing one of the n spatial areas. The set of areas that neighbour the i'th area is denoted by $\mathcal{N}_i$, while the parameter κ determines the strength of the dependencies between neighbours. As κ increases dependence weakens, and as κ decreases dependence strengthens. By the Markov property two elements, x_i and x_j, are independent if they do not share a geographic border. This conditional independence contributes to the sparsity of the precision matrix Q, which allows for efficient computations.The values in the precision matrix Q are given by:

$$Q_{ij} = \begin{cases} \kappa^2 + \|\mathcal{N}_i\| & \text{if } i = j \\ -1 & \text{if } j \in \|\mathcal{N}_i\| \\ 0 & \text{if } j \notin \|\mathcal{N}_i\| \end{cases} \tag{3.18}$$

Where the the precision value for the i'th area is defined as $\kappa^2 + \|\mathcal{N}_i\|$, a neighbouring area of i takes value of -1, and areas that are not neighbours take on the value of 0.

Alternatively the precision matrix can be expressed in matrix format as $\boldsymbol{Q} = \kappa^2 \boldsymbol{I} + \boldsymbol{G}$, where $\boldsymbol{I}$ is the Identity matrix, and $\boldsymbol{G}$ is the matrix defining the neighbourhood structure. More details on the CAR model can be found in chapter 6 of Blangiardo and Cameletti (2015). The spatial dependence between geographic areas will be modeled as a CAR(1). Specification of how it is structured into the model is given in section 3.2.7.

3.2.6 Bayesian Hierarchical Model

The main advantage of the Bayesian approach resides in it taking into account uncertainty in the estimates, and its flexibility and capability of dealing with issues like missing data (Blangiardo and Cameletti, 2015).

In hierarchical Bayesian models all unknown quantities are treated as random variables, the model is specified with multiple parameters, used to represent complex structures with multiple dependencies. The unknown parameters and latent variables are modeled using prior distributions.

Given a vector $y = (y_1, y_2, ...y_n)$ of n observations with x defined as a latent random variable. Latent means it is not observed and that it is inferred from other observed variables which are given. In the spatial case $\boldsymbol{x}$ has high dimension and is therefore referred to as a latent field. The Bayesian hierarchical model has the following components

$\pi(y \mid x, \theta)$	The likelihood of the observations.
$\pi(x \mid \theta)$	The latent field.
$\pi(\theta)$	The prior distribution, can be thought of as prior beliefs about the data.

The hyperparameters θ for outcome variable y and for latent field $\boldsymbol{x}$ are treated as random variables. It can be thought of as prior beliefs about the model and data. The goal is to make inference on the latent field $\boldsymbol{x}$. This is done using observations y, prior $\pi(\theta)$ and Bayes theorem, to find the posterior distribution, $\pi(x, \theta \mid y)$.

3.2.7 The model

The model structure for this analysis consists of a linear structure for fixed effects, correlated random effects for age as a RW1 or RW2, and the spatial dependencies taken into account with a CAR(1) model.

$$\begin{aligned}\boldsymbol{\theta} = [\theta_1, \theta_2]^T &\sim \pi(\boldsymbol{\theta}) \\ \boldsymbol{v} &\sim \mathcal{N}(0, Q_{rw}^{-1}(\theta_1)) \\ \boldsymbol{u} &\sim \mathcal{N}(0, Q_{car}^{-1}(\theta_2)) \\ \boldsymbol{\beta} &\sim \mathcal{N}(0, 10^3 \cdot \mathrm{I})\end{aligned} \tag{3.19}$$

$\boldsymbol{\beta}$ are the fixed effects and take on a linear format with prior covariance matrix value $10^3 \cdot \mathrm{I}$ being the default value given by the INLA package. The $\boldsymbol{v}$ component is the vector containing the age variables which is modeled by a RW1 or RW2. $\boldsymbol{u}$ is the random vector which is used to model the spatial dependency by a CAR model, and the hyperparameters specified as θ_2 and θ_2 are distributed according to the prior $\pi(\boldsymbol{\theta})$. All the latent parameters are collected in the set $\boldsymbol{x}$, referred to as a latent field.

$$\boldsymbol{x} = [\beta_0, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{u}] \sim \mathcal{N}(0, Q^{-1}), \qquad Q = \begin{bmatrix} 10^{-3} \cdot 1 & 0 & 0 \\ 0 & Q_{rw} & 0 \\ 0 & 0 & Q_{car} \end{bmatrix} \tag{3.20}$$

The components of $\boldsymbol{x}$ are what construct the latent spatial field $\boldsymbol{\eta}$ expressed in the equation below. Since the latent field $\boldsymbol{x}$ is assumed to have a joint Gaussian distribution, it can be referred to as the Latent Gaussian Model (LGM).

$$\eta_i = \underbrace{\beta_0 + z_i \boldsymbol{\beta}}_{\textit{fixed}} + \underbrace{v_{a(i)}}_{\textit{Age}} + \underbrace{u_{s(i)}}_{\textit{Spatial}} \tag{3.21}$$

η_i is the linear predictor which is related to the outcome variable $\boldsymbol{y}$ through a logit link function $g(\cdot)$. The spatial component is denoted by $u_{s(i)}$ with $s(i)$ being the geographic area of the i'th observation. Similarly the age component is denoted by $v_{a(i)}$ with $a(i)$ linking observation i to the correct age. With all the implications from above the posterior for the joint posterior distribution of the latent effects and hyperparameters is

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \pi(\boldsymbol{\theta}) \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \prod_{i=1}^{N} \pi(y_i \mid x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta}) |Q(\boldsymbol{\theta})|^{\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{x}^T Q(\boldsymbol{\theta}) \boldsymbol{x} + \sum_{i=1}^{N} \log\{\pi(y_i \mid x_i, \boldsymbol{\theta})\}\right]\end{aligned} \tag{3.22}$$

The main objective is to approximate the posterior marginal distributions:

$$\pi(x_i \mid \boldsymbol{y}) \quad \pi(\theta_k \mid \boldsymbol{y})$$

Additionally there are two assumptions for LGM that are important to produce fast inference (Rue et al., 2009). The first is that the latent field $\boldsymbol{x}$ is Markov, which makes the latent field a GRMF with sparse precision matrix, and the second is that the number of hyperparameters is small. In chapter 3.3 it is discussed how the INLA method estimates on these types of models.

3.2.8 Optimizing Estimation

To speed up the estimation, good initial values for the hyperparameters $\boldsymbol{\theta}$ for the RW and for the CAR are specified. As in section 3.1.4 the variables of use are grouped in terms of the response variable ”revisit”, to attain a binomial distribution. This aggregated data is then used to specify four models which correspond to the final models, but where fixed effects are excluded, and only spatial and temporal effects are included. Excluding the fixed effects gives few aggregation groups and allows for very fast estimation of reasonable initial values. From these resulting models the temporal and spatial parameters are estimated and used as initial values for the final models as seen in equation 3.20. This prior estimation aids in optimizing the final model by reducing out unnecessary calculations.

3.3 How to estimate : INLA

Integrated Nested Laplace Approximation (INLA) is a method for computing Bayesian inference, used specifically for models that can be expressed as latent Gaussian Markov Random Fields (GMRF). Large data sets which can be modeled to take into account spatial and temporal structures are very complex. These data sets have been shown to be computationally costly to model using the ”traditional” Bayesian inference method of Markov Chain Monte Carlo (MCMC). This Chapter gives a brief layout of the INLA method.

3.3.1 INLA procedure

The Goal of the INLA approach is to approximate the posterior marginals of the latent Gaussian field. INLA consists of three steps proposed by Rue et al. (2009). Step one approximates the posterior marginal of $\boldsymbol{\theta}$, $\pi(\theta_k \mid \boldsymbol{y})$ using Gaussian and Laplace approximation. The second step entails computing the Laplace approximation or simplified Laplace approximation for $\pi(x_i \mid \boldsymbol{y}, \boldsymbol{\theta})$. The third step combines steps one and two using numerical integration.

The components of the Latent Gaussian model of interest are the regression parameters, also known as the marginals for the latent field, and elements from the hyperprior distribution, which can, for example represent the variance in random effects or correlation parameters in the autoregressive models.

The posterior marginals of focus are given below, these are obtained from the joint posterior in (3.23). $\tilde{\pi}(\cdot \mid \cdot)$ denotes the approximated conditional densities, i.e. $\pi(x_i \mid \boldsymbol{y}) \approx \tilde{\pi}(x_i \mid \boldsymbol{y})$ and $\pi(\theta_k \mid \boldsymbol{y}) \approx \tilde{\pi}(\theta_k \mid \boldsymbol{y})$. Idealy we would like to obtain,

$$\pi(x_i \mid \boldsymbol{y}) = \int \pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}, \qquad \pi(\theta_k \mid \boldsymbol{y}) = \int \pi(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}_{-k} \tag{3.23}$$

but these integrals are hard (or imposible), and we instead focus on the approximations

$$\tilde{\pi}(x_i \mid \boldsymbol{y}) = \int \tilde{\pi}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}, \qquad \tilde{\pi}(\theta_k \mid \boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}_{-k}. \tag{3.24}$$

As seen above the terms we need to approximate are $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ and $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$.

The first step is the approximation of $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ and estimation of $\hat{\sigma}$ which is necessary for estimating the marginals for the latent field of $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Firstly the joint posterior of the hyperparameters is rewritten as

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{y} \mid \boldsymbol{\theta})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x} \mid \boldsymbol{y}, \theta)\pi(\boldsymbol{y})} \propto \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})} \tag{3.25}$$

where the following equality is used:

$$\begin{aligned}\pi(\boldsymbol{\theta} \mid \boldsymbol{y})\pi(\boldsymbol{y} \mid \boldsymbol{\theta}) =& \pi(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}) = \pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \\ \longrightarrow \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) =& \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x} \mid \boldsymbol{y}, \theta)}\end{aligned} \tag{3.26}$$

Note that the complicated term in $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ is $\pi(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})$, the other parts are related to the model definition: $\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ are the Bernoulli observations, and $\pi(\boldsymbol{x} \mid \boldsymbol{\theta})$ the latent fields. Approximating $\pi(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})$ gives

$$\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) \underset{\sim}{\propto} \left. \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \tag{3.27}$$

where $\tilde{\pi}_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ is a Gaussian approximation of $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$. Computed by the Laplace method and at the mode $\boldsymbol{x}^*$ for the latent field given parameters $\boldsymbol{\theta}$. The parameter estimates can be found by maximizing $\approx \pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ and posterior marginal $\pi(\theta_k \mid \boldsymbol{y})$ can be approximated by numerically integrating out $\boldsymbol{\theta_{-k}}$ from $\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})$. Further details can be found in Rue et al. (2009).

The second step is to approximate $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ which is more difficult due to the latent field $\boldsymbol{x}$ having large dimensions. The previous approximation was simpler due to $\boldsymbol{\theta}$ having smaller dimensions. This results in a more computationally costly calculations and inaccuracies, to remedy this Rue et al. (2009) suggest the use of *simplified Laplace approximation*, which uses Taylor expansion to a specified order.

There are three ways in which INLA approximates the posterior: Gaussian approximation, Laplace approximation, and simplified Laplace approximation. The simplest is to use the Gaussian approximation for $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$. The Gaussian approximation was computed in (3.27) followed by a few additional computations. This method is fast, but the approximation is not sufficiently accurate.

The next approach is to use the Laplace approximation

$$\tilde{\pi}_{LA}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{\pi(x, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_{GG}(\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})} \tag{3.28}$$

As shown in Rue et al. (2009), where $\tilde{\pi}_{GG}$ is the Gaussian approximation of the conditional $(\boldsymbol{x}_{-1} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})$ and $\boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})$ is the mode. The approximation resulting from this method is very good, however $\tilde{\pi}_{GG}$ must be recomputed for each value of x_i and each $\boldsymbol{\theta}$, since the precision matrix depends on these values. This results in costly computations which are not feasible.

Finally, the posterior can also be approximated by the simplified Laplace approximation, which can be seen as a combination of the two previous approaches, this method allows for efficient, cheap computations and good approximations. This method is based upon the Taylor expansion to the third order of the Laplace approximation of $\tilde{\pi}_{LA}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$. After attaining the approximations $\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})$ and

$\tilde{\pi}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ the marginal posterior distributions of interest $\pi(x_i \mid \boldsymbol{y})$ and $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ can be approximated by solving the integrals from equation (3.25). INLA solves the integrals numerically by the following approximation

$$\tilde{\pi}(x_i \mid \boldsymbol{y}) \approx \sum_j \tilde{\pi}(x_i \mid \boldsymbol{y}, \boldsymbol{\theta}^{(j)})\tilde{\pi}(\boldsymbol{\theta}^{(j)} \mid \boldsymbol{y})\Delta_j \tag{3.29}$$

Summing over the set of suitable integration points $\boldsymbol{\theta}^j$ selected based on the Gaussian approximation in equation (3.26), with each point being associated to a weight Δ_j, which is the distance between the integration points. A final note on INLA given by Morrison (2017): The term Integrated comes from the use of numerical integration, the term Nested comes from needing $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ to obtain $\pi(x_i \mid \boldsymbol{y})$, and the term Laplace Approximation comes from it being the method to obtain the parameters for the approximation.

More in depth specifications for the INLA procedure are found in Rue et al. (2009). To implement approximate Bayesian inference using the INLA method an R package is available at https://www.r-inla.org/, as well as many modeling examples, theory, questions and answers, on the applications of INLA can be found.

3.4 Model selection

In section 3.4.1 the use of Widely Applicable Information Criterion (WAIC) is explained to select the best model and to see whether adding a component to the model to account for spatial correlation on a municipal and postal-code level improves the modeling significantly. In section 3.4.2 the variable selection performed by Lundgren (2021) is explained, to provide an overview for completeness.

3.4.1 Widely Applicable Information Criterion

The Widely Applicable Information Criterion (WAIC) is a method for estimating pointwise out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values (Vehtari et al., 2017). WAIC assess how accurate Bayesian models are at representing the data they are modeling, using an approach of estimating out of sample predictive accuracy using within sample fits. WAIC can be thought of as an improved version of the deviance information criterion (DIC) which is commonly used when assessing geographic models, the reason WAIC is not as commonly used as DIC or as Akaike Information Criterion (AIC) is due to it having additional computational steps, such as the effective number of parameters being computed differently. More information on different criterion methods used for Bayesian model selection can be found in Gelman et al. (2014).

The goal is to obtain the expected log pointwise predictive density (elppd). To begin, the log pointwise predictive density (lppd) of the data is calculated in practice by drawing from the posterior distribution, where $\theta^s, s = 1, ..., S$ denoted as $p_{post}(\theta)$ in Gelman et al. (2014).

$$\text{lppd} = \sum_{i=1}^{n} \log(\frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta^s)) \tag{3.30}$$

The calculated lppd for the observed data y is an overestimate of the elppd for future data. To remedy this, a correction for the effective number of parameters to prevent overfitting is applied to equation 3.30 to attain a reasonable estimate of the elppd. The computation for the effective number of parameters p_{waic} is as follows:

$$\text{p}_{waic} = \sum_{i=1}^{n} V_{s=1}^{S}(\log p(y_i \mid \theta)) \tag{3.31}$$

where the posterior variance of the log predictive density for each data point is computed. Summing over all the points y_i is done to obtain the effective number of parameters, and where $V_{s=1}^{S}$ is the sample variance. The bias correction is done by subtracted the p_{waic} from the lppd to attain the expected log pointwise predictive density,

$$\widehat{\text{elppd}}_{waic} = \text{lppd} - \text{p}_{waic}. \tag{3.32}$$

The expression above is multiplied by a factor of -2 to make it comparable with AIC and DIC. In Watanabe's original definition, WAIC is the negative of the average log pointwise predictive density (assuming the prediction of a single new data point) and thus is divided by n and does not have the factor 2 (Gelman et al., 2014). Where we use:

$$\text{WAIC} = -2\text{lppd} + 2\text{p}_{waic} \tag{3.33}$$

AIC is defined as: $\text{AIC} = -2\log p(y \mid \hat{\theta}_{mle}) + 2k$ where the log predictive density given the maximum likelihood estimate minus the number of parameters estimated in the model k gives a correction for how much these parameters increase predictive accuracy. AIC is best for linear models. For models with informative priors or hierarchical structure, the effective number of parameters strongly depends on the variance of the group-level parameters (Gelman et al., 2014). Thus WAIC is best for the analysis, comparison between AIC, DIC and WAIC is found in Gelman et al. (2014).

The WAIC is implemented in R-INLA, and obtained as part of the model fitting. All the models must be fitted using the same data when comparing multiple Bayesian models with this method. The WAIC value itself cannot be directly interpreted but rather the different models WAIC should be compared to each other, and the preferred model is that with the lowest value.

3.4.2 Variable selection

Here we will use the variables found in Lundgren (2021), and this section is included for completeness. In previous analysis the diagnostic group for coronary artery disease and variables gender, age, and CNI were deemed significant, those variables were considered when performing the analysis. Significant variables were determined by:

1. Uni-variate Logistic regression is performed on gender, age and CNI to evaluate their relationship with the outcome variable, followed by a multivariate regressions consisting of all three variables. The age variable was treated in three different ways: as continuous, as a factor, and as a Spline function. The latter was the final choice and modeled using B-Spline of order one. Different alterations of internal node choice for the B-Spline were considered and compared using AIC, and significance of individual parameters. The selected model was compared to the best of the univariate models.

   The random effect was studied on the intercept at the municipal level by using the lme4 package in R to perform analysis of variance between the municipalities. Each municipality was assigned one random effect estimate depending on the municipal affiliation, details are shown in Lundgren (2021).

   From part 2 to part 5 model selection is discussed after having chosen covariates with AIC and significance for nested models tested with LTR. Then specificity and sensitivity were used determine useful models.

2. Fixed and mixed effect regression model was adapted to the diagnostic group coronary artery disease. Initially the same model as in part one was used, but with the use of backwards elimination any insignificant parameters were removed.

3. Analysis of the CNI index was done by evaluating the weighting and individual variables significance. The CNI variables were treated as independent and separately added to the model instead of the full index, and significance tests were performed. From the model Estimated parameter β_{CNI_i} ware compared with their weighting constant w and estimates of β's of entire CNI index β_{CNI}. Specific details are shown in (Lundgren, 2021).

4. The elastic net was applied using the glmnet package in R for choosing between algorithmic models. For weighting penalty functions between models, the error was set to 0.5, both λ values to test the effect of the penalty function on the variable selection. Since the Spline cannot be combined with an elastic net, an offset in the linear predictor representing the splines contribution to the GLM was used.

Note: If there had not been a previous analysis done by Lundgren (2021) an alternative approach to selecting the co-variates as applied in (Tufvessson, 2017) would be used. This is similar to the approach in Lundgren (2021), but modeling the GMRF and using INLA.

Results

In this chapter the results are presented for the models on a municipal spatial level and on a postal-code spatial level. The outline of the analysis performed on each group of models partly stems on the ”suicides in London” example, in chapter 6.1.2 of Blangiardo and Cameletti (2015). An overview of the data is given in Lundgren (2021), and here a table of the covariates description can be found in Table A.1 in the appendix.

4.1 Results

There are eight models fitted having the form as specified in equation (3.22). The models specified in Table 4.1 have municipalities as the spatial boundaries, and models spefified in Table 4.2 have postalcode areas as spatial boundaries.

Table 4.1: Model specification on municipal spatial level, with corresponding linear predictor.

Model	Linear predictor
mod.iid.rw1.kom	$\eta_i = \beta_0 + z_i\boldsymbol{\beta} + v_{rw1} + u_{iid}$
mod.iid.rw2.kom	$\eta_i = \beta_0 + z_i\boldsymbol{\beta} + v_{rw2} + u_{iid}$
mod.besag.rw1.kom	$\eta_i = \beta_0 + z_i\boldsymbol{\beta} + v_{rw1} + u_{car}$
mod.besag.rw2.kom	$\eta_i = \beta_0 + z_i\boldsymbol{\beta} + v_{rw2} + u_{car}$

Table 4.2: Model specification on postal-code spatial level, with corresponding linear predictor.

Model	Linear predictor
mod.iid.rw1.post	$\eta_i = \beta_0 + z_i\boldsymbol{\beta} + v_{rw1} + u_{iid}$
mod.iid.rw2.post	$\eta_i = \beta_0 + z_i\boldsymbol{\beta} + v_{rw2} + u_{iid}$
mod.besag.rw1.post	$\eta_i = \beta_0 + z_i\boldsymbol{\beta} + v_{rw1} + u_{car}$
mod.besag.rw2.post	$\eta_i = \beta_0 + z_i\boldsymbol{\beta} + v_{rw2} + u_{car}$

As shown in the tables above the co-variate fixed effects $z_i\boldsymbol{\beta}$ in all eight models is the same, the difference lies in the specification of the temporal and spatial effects, v and u respectively. Model mod.iid.rw1.kom, is an approximation of the final model by Lundgren (2021). Model mod.iid.rw2.kom, is similar to the first but, with temporal effects modeled as a RW2. Model mod.besag.rw1.kom, conditional independents regarding the spatial effects is incorporated, with the temporal effect being a RW1. Model mod.besag.iid.rw2.kom, also has the incorporation of conditional independents regarding the spatial effects but, with temporal effect being specified as a RW2. The models in Table 4.2 have the same specifications as the four previous models except for having postalcode areas as the geographic boundaries instead of municipalities.

Before running any of the models, two graphs which assigns the set of neighbors for each municipality and for postalcode area is specified. The graphs are produced from two shapefiles, each containing one of the erea boundaries. With use of certain function in R, both shapefiles are transformed into adjacency matrices to make them compatible with the R-INLA format, it can then be visualized for the municipalities as Figure 4.1 and for postalcode areas as Figure 4.2. As an example we can see that municipality number 23 being Lund has 8 neighbouring municipalities: 2:Staffanstorp, 8:Kävlinge, 9:Lomma, 10:Svedala, 11:Skurup, 12:Sjöbo, 23:Lund, and 27:Eslöv.

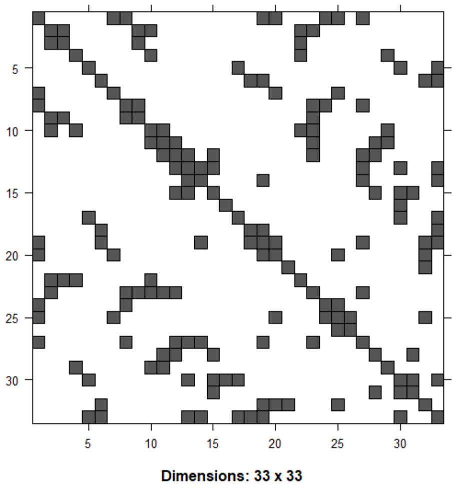

Figure 4.1: Adjacency matrix for municipalities: rows and columns identify municipality areas; squares identify neighboring municipalities

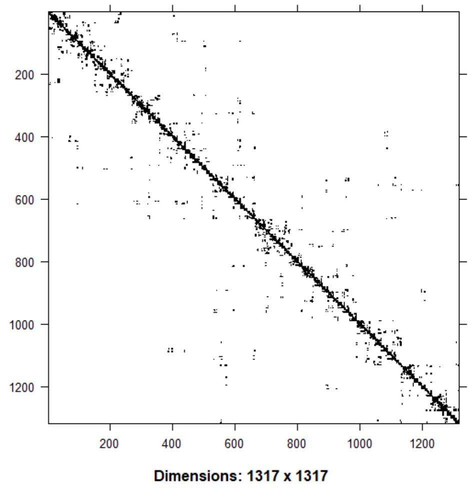

Figure 4.2: Adjacency matrix for postalcode areas: rows and columns identify postal-code areas; squares identify neighboring areas

Next, data modification is needed where reduction of the observations is specified in Chapter 2.3. The data and the usable shapefiles are joined by a municipality ID and postalcode ID, rendering two separate data sets. To optimize estimation of the models, initial values for the hyperparameters $\boldsymbol{\theta}$ are specified as in section 3.2.8. Once model specification is done for the eight models, they are computed using the INLA command in R.

The WAIC values are computed to perform comparison of each model based on their predictive accuracy. Table 4.2 shows these results, where the ”best” models on a municipal spatial level based on the WAIC value are those with the spline-function for age approximated as a RW2, i.e. mod.iid.rw2.kom and mod.besag.rw2.kom.

Table 4.3: The WAIC for each model, WAIC and standard error for each model wrt mod.iid.rw1.kom set as reference model.

	WAIC	WAIC wrt ref.model	s.e. wrt ref.model
mod.iid.rw1.kom	134051		
mod.iid.rw2.kom	**134024**	**-27**	**59**
mod.besag.rw1.kom	134056	-5	28
mod.besag.rw2.kom	**134024**	**-27**	**59**

Table 4.5 shows the ”best” model on a postal-code level based on the WAIC value is that with the spline-function for age approximated as a RW2 and spatial dependencies as an Conditional Autoregressive model, i.e. mod.besag.rw2.kom.

Table 4.4: The WAIC for each model, the WAIC and standard error for each model wrt mod.iid.rw1.post set as reference

	WAIC	WAIC wrt ref.model	s.e. wrt ref.model
mod.iid.rw1.post	133181		
mod.iid.rw2.post	133154	-27	59
mod.besag.rw1.post	132904	-277	101
mod.besag.rw2.post	**132877**	**-304**	**103**

The variables used for all models is listed in Table 4.5, the only difference lies in spatial identification for the the two sets of models by using **kommun_ID** (33 municipalities) and **postalcode_ID** (1317 postal-code areas). The CNI has a range between 0.7241 - 6.9864 where the mean for individuals in the data set is 2.569. There are four diagnoses being used, Diabetes, COPD, Coronary heart disease, and Strokr/TIA, and multiple diagnoses, where a value of 1 indicates the individual is part of the diagnosis group and 0 indicating that they are not. The gender variable represents gender of the individual, 0 for female and 1 for male. The age variable indicates age of individual, going from 1 year of age to 106.

Table 4.5: Variable interpretations and their value.

Variables	Interpretation	Value
CNI_per_person	CNI per person	0.7241 - 6.9864
diag_diab	Diabetes diagnosis group	0 or 1
diag_kol	COPD diagnosis group	0 or 1
diag_kr	Coronary heart disease diagnosis group	0 or 1
diag_str	Stroke/TIA diagnosis group	0 or 1
multisjuk	Individual with more than one diagnosis	0 or 1
male	Indication of gender 1 for males 0 for females	0 or 1
alder_20171231	Age of individual	1- 106
kommun_ID	municipality ID linking data to shapefile	1 - 33
postalcode_ID	postalcode ID linking data to shapefile	1- 1317

For the $\boldsymbol{\beta}$-estimates for each model, Figure 4.3 is given. The estimates are very similar for all of the models, where **diag_diab** has a high impact on revisits.

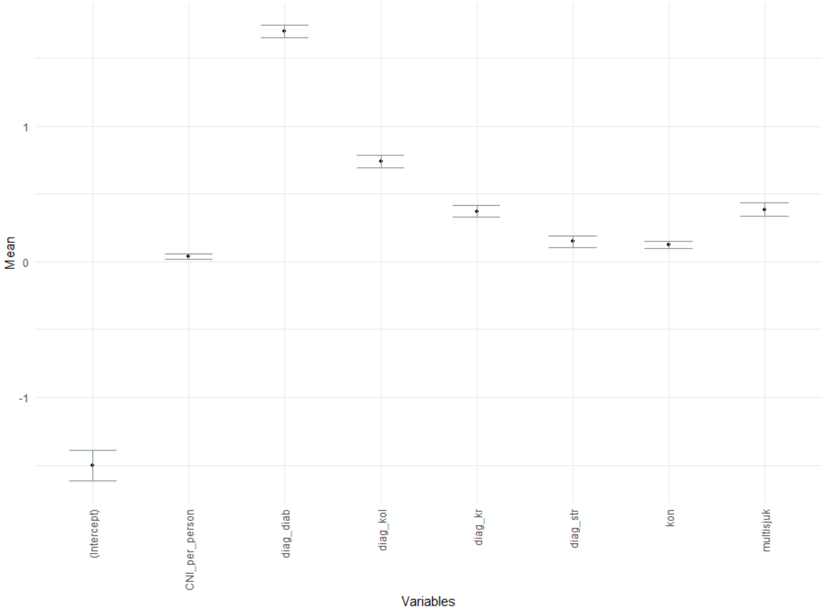

Figure 4.3: β-parameters and their 95% Confidence Intervals for each model.

The temporal effect v defined by the **age** variable, modeled as a RW1 for models mod.iid.rw1.kom, mod.besag.rw1.kom, mod.iid.rw1.post, mod.besag.rw1.post, and as RW2 for mod.iid.rw2.kom, mod.besag.rw2, mod.iid.rw2.post, mod.besag.rw2.post. All models defined with a RW1 produce the same temporal effect, likewise with all models defined as a RW2, their plots are shown in Figure 4.4. We see that the probability of a revisit increases with age up until 84 years for the RW1 and 86 for the RW2, then slight decreases occurs for all models as the age increases. Additionally we see the that the models with a RW2 are smoother, and based on the WAIC of the corresponding models they also have better predictive accuracy.

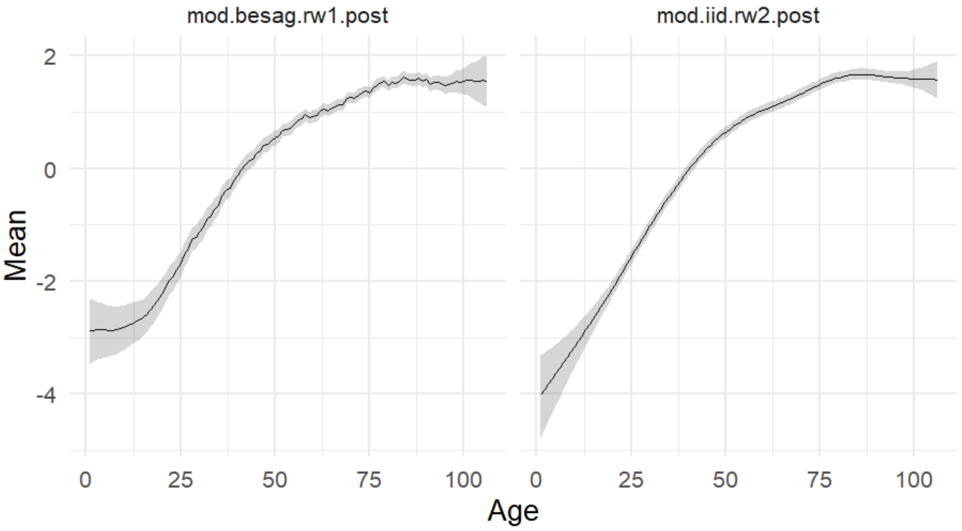

Figure 4.4: Plots of the **age** variable modeled as RW1 and RW2.

For the random effects v and the spatial component of our model, we compute the posterior mean by extracting the marginal posterior distribution for each area, and apply the inverse logit transformation.

In Figure 4.5 the map of the posterior mean for each municipality is shown, all four of the models are very similar, which is supported by the WAIC results from Table 4.3, where values differ no more than by 27. The municipalities in blue indicate higher probability of a revisit whereas the ones in red indicate areas where there is a deficiency in revisits. The areas in white, tending to 0.5 signify that there is no effect on revisits based on the municipality.

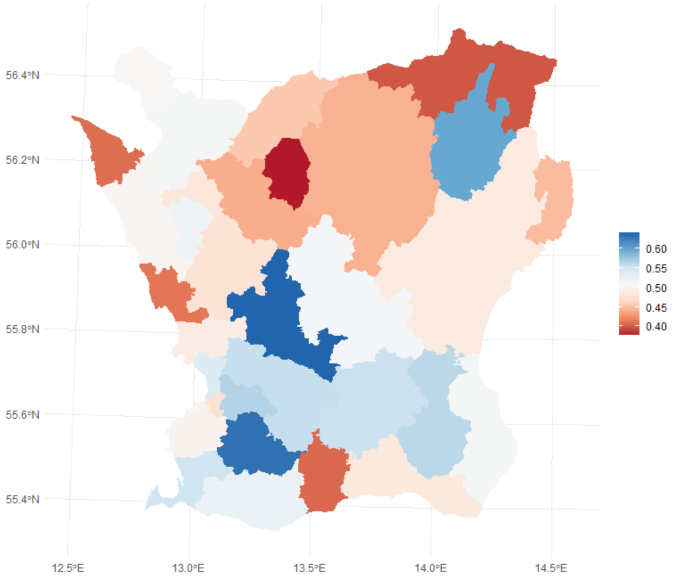

Figure 4.5: Map of posterior mean specified for each municipality.

In Figure 4.6 the map of the posterior mean for each postal-code area is shown. Where it was calculated by extracting the marginal posterior distribution for each area, and applying the inverse logit transformation. From these models we can see that the upper areas of Skåne, and a few in the south have a negative effect on the probability of a revisit. For a more detailed close up of the map, Figure A.2 can be referred to. For the models with i.i.d for spatial effect there are a few postal-code areas with no predicted value, this is due to some postal-codes having no observations. In the models with CAR as spatial effect these empty values are inferred from the neighbouring areas.

Note: Grey areas on the maps seen in Figures 4.6, 4.7, and 4.8 are due to postal-code areas not having observations. A closeup map can be seen in Figure A.2 in the appendix.

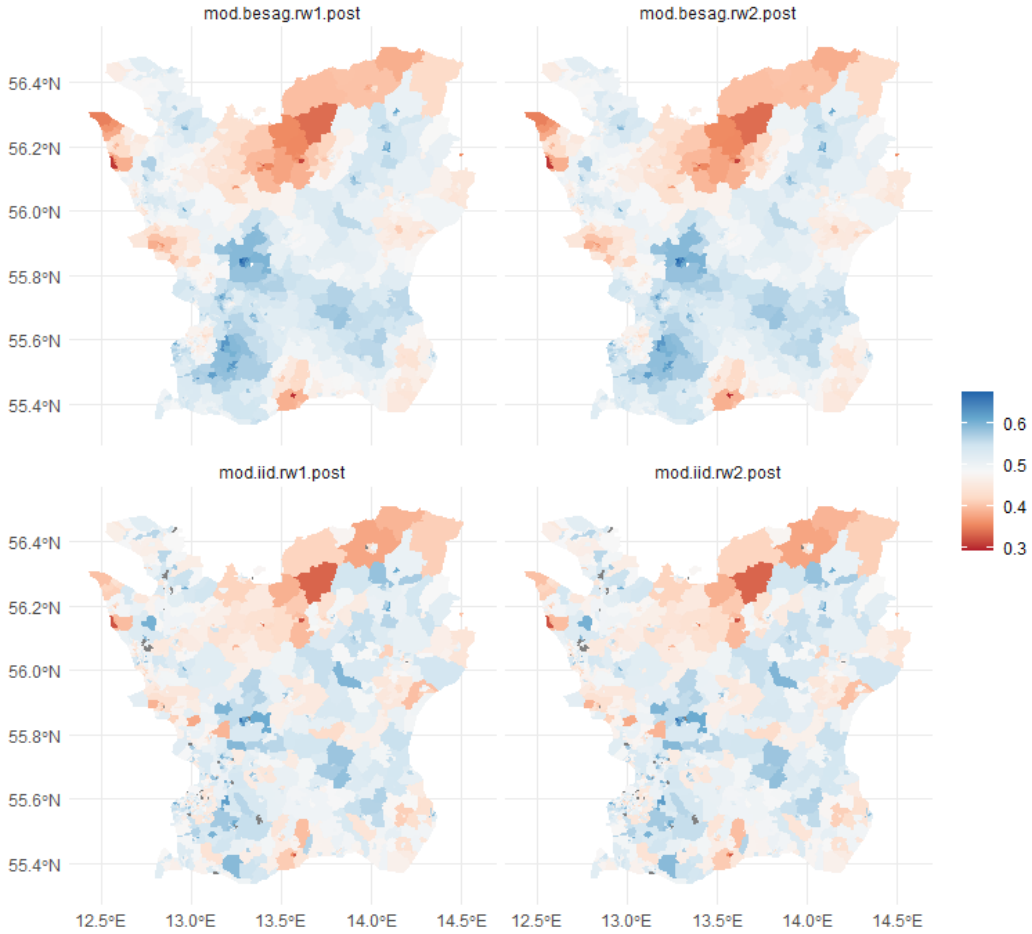

Figure 4.6: Map of posterior mean for all four models specified for each postal-code area.

Looking specifically at the **CNI_per_person** variable which has a range between 0.7241 and 6.9864 given in Table 4.3 to see its effect on revisits in different geographic areas. A map of the average CNI value per postal-code area is shown in Figure 4.7. The average mean value is 2.572, this is seen when looking at the map and legend, where the lighter color represents a lower CNI value, and darker a higher value.

In Figure 4.8 the map of the posterior mean for the variable **CNI_per_person** in each postal-code area is shown. Similarly to Figure 4.6 the posterior mean is calculated by multiplying the **CNI_per_person** for each observation by the fixed β value from mod.besag.rw2.post, and then applying the inverse logit transformation. Here we see that the range of the CNI value is between 0.52 and 0.50. This indicates that **CNI_per_person** has a weak effect on revisits. Note: The model used for both maps in Figure 4.7 and Figure 4.8 is mod.besag.rw2.post.

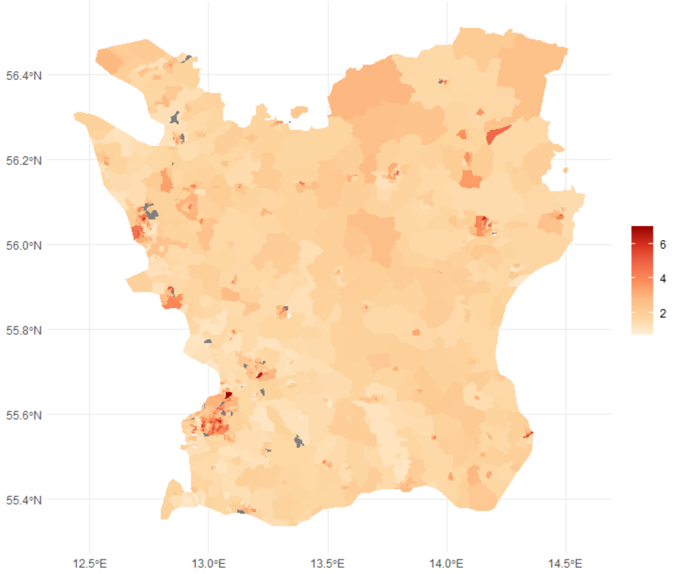

Figure 4.7: Map of Average CNI value for each postal-code area.

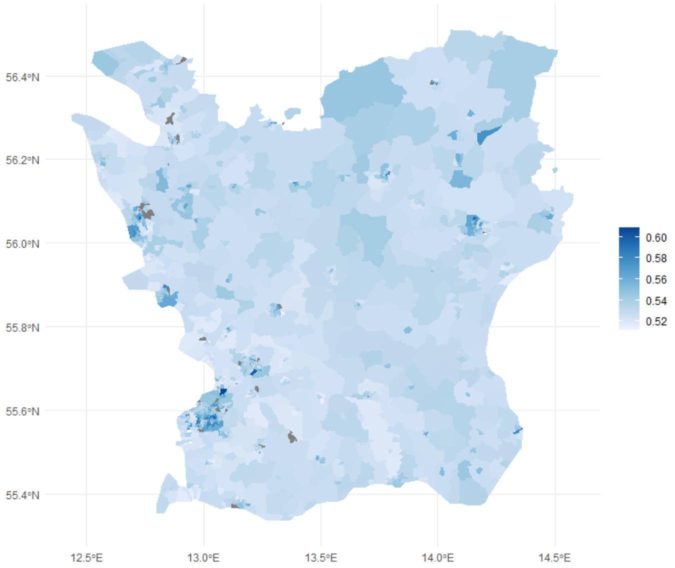

Figure 4.8: Map of posterior mean for the variable CNI in each postal-code area.

Conclusion & Discussion

In section 5.1 the conclusion is presented. In section 5.2 possible ideas for further work regarding this analysis is discussed.

5.1 Conclusion

To conclude on all the models comparison from results in section 4.1, and of previous analysis is discussed. Note that all the models in this analysis fall into the class of GMRFs, where the INLA method is applied numerically for efficient estimation of these models, which otherwise would be computationally costly. The specification of the initial values for the hyperparameters $\boldsymbol{\theta}$ improves the approximation of the models.

Regarding the fixed effects for all models on a municipal and postal-code level the estimated $\boldsymbol{\beta}$-parameters give the same results, which are also the same as in previous analysis by Lundgren (2021).

For the splines modeled as RW1s and RW2s we see that probability of a revisit increases with age up until 84 and 86 years, for all RW1 and RW2 respectively, past those years revisits begin to decrease with age. For the municipality models the ones with a RW2 model outperform in terms of predictive accuracy i.e. mod.iid.rw2.kom and mod.besag.rw2.kom. The model mod.iid.rw2.kom is comparable to the final model of Lundgren (2021) being a logistic mixed effect model with age variable modeled as a spline function and municipalities as independent random effects, but with a spline of degree 2.

For the postal-code models, the model mod.iid.rw1.post is comparable to Lundgren (2021)'s model but on a more elaborate geographic level, it performed the least well in comparison to the other models. Model mod.besag.rw2.post is superior based the predictive accuracy. We see from the maps that in some of the more northern areas in Skåne there is higher probability of not having a revisit. This could be due to these regions being rural and distances being greater to health care centers making them less accessible for individuals. Additionally taking a look at the **CNI_per_person** we see it does not seem to greatly effect the results for revisits.

To conclude, for the municipal models specifying the spatial effects as i.i.d. or CAR does not seem to influence the models predictive accuracy, however the for best results the temporal effect must be modeled as RW2. The latter concerning the RW2 also applies for the postal-code models, where in addition the spatial effects modeled as CAR rather than i.i.d is best.

5.2 Further work

There is further work that can be done regarding this analysis:

- Regarding the model mod.besag.rw1.post, where comparison of an identical model but with separated CNI variables in the fixed effects instead of the full CNI index is worth looking into.
- Continuing the analysis on specific postal-code areas where indication of high or low effect on revisits are present.

Appendix

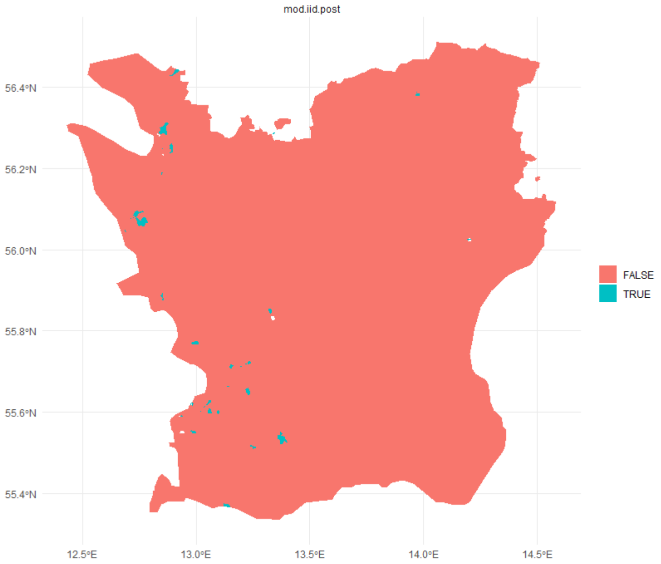

Figure A.1: Areas with no observations for the models with spatial effect modeled as i.i.d. where TRUE: no value & FALSE: yes value.

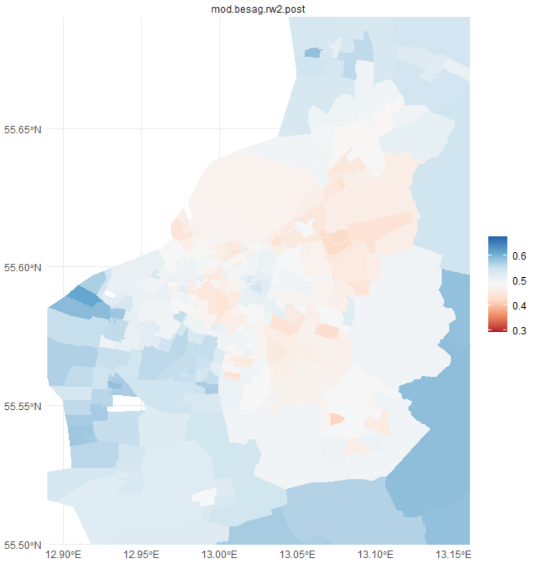

Figure A.2: Map of posterior mean from mod.iid.rw2.kom of Malmö.

Table A.1: Variable description

Indicator variables	Interpretation	# of 0's	# of 1's	Total
male	1 = males 0 = females	63404	73939	137343
diag_diab	Diabetes	75115	62228	137343
diag_kol	COPD	112892	24451	137343
diag_kr	Coronary heart disease	91301	46042	137343
diag_str	Stroke/TIA	104804	32539	137343
multisjuk	Multiple diagnoses	99879	37464	137343
Continuous variables	Interpretation	Minimum	Maximum	Mean
CNI_per_person	CNI per person	0.00	6.99	2.57
alder_20171231	Age of individual	1.00	106.00	70.59
Identification variables	Interpretation	# of areas		
kommun_ID	municipality ID	33		
postalcode_ID	postalcode ID	1317		

References

A. Agresti. *An introduction to categorical data analysis.* Hoboken, New Jersey ; Chichester : Wiley, cop., 2006. ISBN 9780470114759.

J. Besag, J. York, and A. Mollie. Bayesian image restoration with two applications in spatial statistics. 43(1):1–20, 1991.

M. Blangiardo and M. Cameletti. *Spatial and Spatio-temporal Bayesian Models with R-INLA, First Edition.* John Wiley Sons, 2015.

A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data, Second Edition.* Cambridge University Press, 2013. ISBN 9781139013567.

E. ESRI. Esri shapefile technical description (an esri white paper). 1998.

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

S. Lundgren. Regression analysis för utredning av omotiverade skillnader I tillgang till uppföljande vård för kroniskt sjuka I Region Skåne. 2021.

C. E. McCulloch and S. R. Searle. *Generalized, Linear, and Mixed Models.* John Wiley Sons, 2001. ISBN 9780470073711.

K. Morrison. A gentle inla tutorial. *Precision Analytics*, 2017.

S. K. o. R. Primärvårdskvalitet. Pvq prioritering pr1fys. `https://kvalitetsindikatorkatalog.se/#/measures/edit/9cee4daa-dbc7-47b1-aa31-6af393791175`. Online; accessed: 01.09.2016.

H. Rue and L. Held. *Gaussian Markov Random Fields (Theory and Applications).* Taylor Francis Group, LLC, 2005. ISBN 9780429208829.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Royal Statistical Society*, 71(2):319–392, 2009. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x.

O. Tufvessson. An epidemiologist approach to improved car insurance. 2017.

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

X. Wang, Y. Yue, and J. J. Faraway. *Bayesian Regression Modeling with INLA.* CRC Press Taylor Francis Group, 2018. ISBN 9781351165761.

Bachelor’s Theses in Mathematical Sciences 2021:K31
ISSN 1654-6229

LUNFMS-4059-2021

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/