Spatio-temporal analysis of COVID-19 in Västra Götaland, Sweden

Master's thesis in Mathematical Statistics

Natalia Andreeva

Abstract

Spatio-temporal analysis of COVID-19 data with the two different statistical approaches is the main objective of this thesis. The first classical approach, the Endemic-Epidemic framework (Held et al., 2005) is a class of multivariate time-series models for the incidence counts, obtained from the surveillance systems. In this formulation, the conditional mean of the number of cases is partitioned into endemic, autoregressive and spatio-temporal parts, representing different sources of infection contribution. The second approach used in the thesis is INLA (Integrated Nested Laplace Approximation (Rue and Martino, 2007)), which performs the approximate Bayesian inference for latent Gaussian models. The flexibility of the both approaches allows for various extensions of the models. As the thesis progresses, we search for the best model with different metrics used as a selection criteria.

Both frameworks allow for the inclusion of the socio-demographic covariates in the analysis, as possible drivers of the disease spread. Guided by a previous study of Söderberg et al. (2022), we chose the covariats of interest to be: Income, Foreign background, Education, Overcrowding, Square meters per person, Employed, Care workers. Also, the age factor was added as two covariates: Young and Older.

It was shown that the Endemic-Epidemic approach with a complex seasonal trend, random intercepts and the spatial weights, assigned according to the power-law principle, but without any socio-demographic covariate, achieved almost as low metric values as the best model. Given the aforementioned extensions, the best model included the following socio-demographic covariates: Education and Foreigners in the endemic part, Young and Square meters per person in the autoregressive and Overcrowded, Foreigners, Older, Income in the spatio-temporal part. All these covariates positively correlated with the number of counts.

The model with the random walk time formulation applied within INLA technique showed on average a positive correlation of the case counts with Foreign Background, Care workers, Overcrowded, Education and Income. A negative correlation with the case counts on average was shown by the Older, Young, Employed and Square meters per Person. The results suggest further research about the impact of the socio-demographics on the case counts of viral diseases.

Acknowledgements

I am deeply grateful for my supervisor, Ottmar Cronie, for giving me the opportunity to write this master thesis and for the support and assistance throughout the whole process.

<div align="right">Natalia Andreeva, 2023</div>

Contents

6 Appendix   **61**

1 Introduction

As of May 2023, the COVID-19 pandemic is over according to WHO. During the last 3 years millions of people were severely affected by the coronavirus disease, and viral outbreaks still remain a challenge for health care systems. Therefore, it is necessary to identify the factors that influence the spread of the disease, so that the strategies to prevent and to slow down the disease transmission can be undertaken efficiently by the public health authorities. It was shown that socio-demographic factors, such as age, ethnicity, population density, occupation at health care organisations, low income have significant impact on the number of outcomes or on the disease development (Andersen et al. (2022); López-Gay et al. (2022); Marković et al. (2021)). In particular, in the work of Söderberg et al. (2022) the authors found that overcrowded living conditions, low income and low education, as well as a foreign background are associated with a the higher risk of testing positive for COVID-19. This thesis will continue the study of the impact of these socio-demographic variables on the disease outcomes, which is a primary goal of this work.

As the infectious disease spreads around a geographical area, it is important to take into consideration the similarity of the disease prevalence across the neighbouring areas and at the consecutive time points when analysing the data. Spatio-temporal methods are prevailing over time-series analysis and pure spatial methods, because they allow to identify the patterns of disease transmission over space and time simultaneously. In the review Ibañez et al. (2021) of spatio-temporal models for count data, the authors analyze the daily data of positive tests of COVID-19 in the Valencian community in Spain, using both the classical frequentist approach and the Bayesian framework. They compare the performance of different approaches by computing the root mean square error (RMSE) of the predictions. In this thesis, we will apply the Endemic-Epidemic time-series model framework to model the case counts in Västra Götaland, Sweden, which is a classical frequentist approach, as well as a Bayesian spatio-temporal disease modelling/mapping approach. The comparison of the two frameworks is the secondary goal of the thesis.

In Section 2 we will present the data and its structure. Section 3 describes the general Endemic-Epidemic modelling with extensions, the search for the best models in terms of different selection criteria, concluding with the presentation of

the results. Section 4 gives an introduction to the INLA method with an example, followed by the possible models for spatio-temporal areal data and the obtained results. The thesis is concluded with a discussion.

2 Data

We have the surveillance records of the COVID-19 PCR tests consisting of the positive or negative outcomes, the date of the testing for the positive outcomes and a unique code of the DeSO where the testing procedure was conducted. DeSO zones (Swedish: DEmografiska Statistik Områden), are the demographical statistical areas, with national division considered from 1st January 2018 and is still valid to date. Sweden is divided into 5984 DeSO zones, each of which contains from 700 to 2700 residents. The Västra Götaland region is divided into 992 DeSO zones with 1734443 residents in total (estimated for year 2020). The records of PCR tests were collected on a daily basis, starting from 01-01-2020 up to 21-06-2021. Two records, which were dated to the years 2010 and 2014, were removed from the dataset. Also, 179 cases were deleted from the data, since those individuals didn't have a fixed DeSO residential status.

Spatial data is usually divided into three types: areal data, point-referenced data and spatial point patterns. Here, we have an areal data type, where the aggregated count y_i over areal unit i are modelled with a random variable Y_i. The areal units i have well-defined boundaries given by the DeSO boundaries in our case. For areal data, the neighbourhood structure of the region map defines the dependencies between the observations. Due to the ethical considerations related to the privacy protection of all residents, we don't have access to the home addresses corresponding to individual test results, instead we will analyze the numbers of counts, aggregated by the time points and the region. The counts y_{it} are modelled as random aggregated numbers of positive tests where $i = 1, \ldots, I$ are the area indices and $t = 1, \ldots, T$ are day indices. Recall that we have $I = 992$ DeSO zones in Västra Götaland and we consider a total of $T = 537$ days. The temporal dependence between the consecutive daily observations enables to use time-series analysis for the counts. Figure 1 shows daily positive tests aggregated over DeSO zones in Västra Götaland. In Figure 2 the number of cases per person for each DeSO are shown, aggregated over the study period.
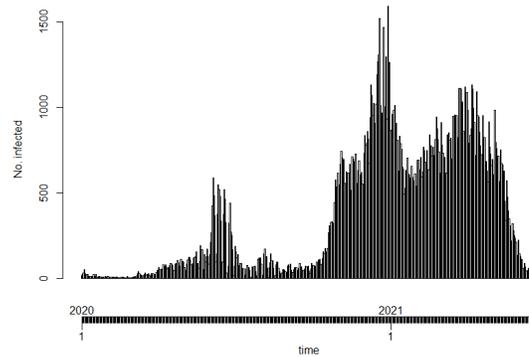
Figure 1: Time series of daily counts of positive PCR test, aggregated over the regions, from 1st Jan 2020 till 21st June 2021.

The socio-demograpic covariates, considered in this thesis, are purely spatial and supplied per DeSO zone. They constitute aggregated counts of various individual sociodemographic factors for the residents of each DeSO. Sampled in the year 2019, these are considered representable for our study period, since demographic trends are not changing significantly over shorter time periods in Sweden. In Figures 20, 21, 22 in the Appendix we supply graphical illustrations of all covariates under consideration.

Figure 2: Incidence rates for each DeSO, aggregated over the study period.

3 Endemic-Epidemic framework
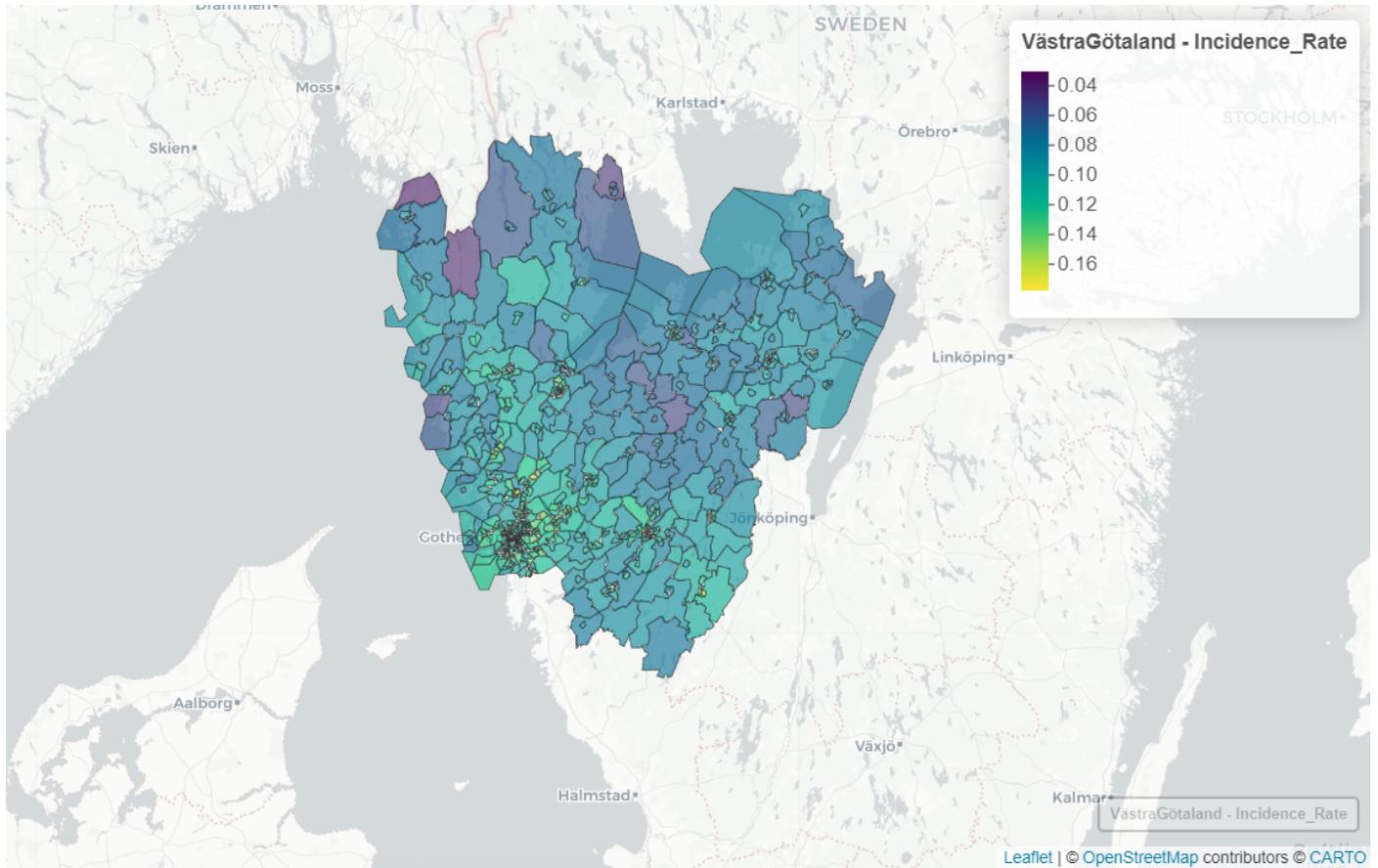
We start with a general basic framework of Endemic-Epidemic disease modelling in Section 3.1 and consider the model extensions in 3.2. In Sections 3.3-3.5 we will discuss possible specifications of the model terms. The model selection is discussed in 3.6, followed by the description of scoring rules for predictions of count data. The section is concluded with the obtained results in 3.8.

3.1 General model

Originally, this framework for analysis of spatio-temporal data was proposed by Held et al. (2005), and developed further in subsequent works (Paul et al., 2008; Paul and Held, 2011; Held and Paul, 2012; Meyer and Held, 2014). The pure parameter-driven approach, where the counts y_t are assumed to be Poisson distributed with a parameter λ, which drives the distribution of counts, is not able to describe the epidemic nature of infectious diseases well. The observed number of counts in the past y_{t-1} should be included as another governing factor, so the model is observation- and parameter-driven (Held et al. (2005)). The simple version of the model is motivated by the branching process with immigration where the conditional mean of $y_t|y_{t-1}$ is defined as

$$\mu_t = \nu + \lambda y_{t-1}, \tag{1}$$

where ν is a positive mean immigration term, or endemic term, and $\lambda \in (0,1)$ governs the branching part. In general, this branching process models the reproduction of an individual (for example, a bacteria), so that the expected number of new individuals at time t is the number of individuals at the previous time step, y_{t-1}, multiplied with an autoregression parameter λ, plus the number of the immigrated individuals ν. The appealing property of the model is that the branching process is approximated to a chain binomial model, when the number of susceptibles is unknown, which is often the case in surveillance systems.

We start with a brief introduction to the Reed-Frost chain binomial model for epidemic diseases (Abbey, 1952), then we derive the approximation for the conditional mean μ_t (Wakefield et al., 2019). The whole population is divided into three compartments: Susceptibles, Infectives and Recovered. The discrete time intervals, during which infection is transmitted to susceptible individuals, are called

13

the generation times, and at each generation time new infectives emerge. The Reed-Frost model assumes homogeneous mixing, that is a contact between two individuals in a population occurs randomly with equal probability. Also, we assume that the infectious period lasts only one generation time, after which the individuals are removed (recovered). Denote the number of susceptibles, infectives, and recovered at generation time t by x_t, y_t, z_t, respectively, and the total population size by N. Under the assumption of homogeneous mixing and frequency dependent contact rate, we can express the infection rate as $\beta = c_{FD}p_I$, where c_{FD} is the frequency of contacts and p_I is the probability of infection transmission from an infective to a susceptible individual (Wakefield et al., 2019). Then during the time interval $[t-1, t)$ the probability that a susceptible meets an infective individual and gets infected is β/N. It implies that the probability that this individual avoids becoming infected is $1 - \beta/N$, which for a large population N is approximated by a Taylor expansion given by $\exp\{-\beta/N\}$. Then, if all contacts are independent, it follows that the probability that one susceptible avoids becoming infected, after having contacted all the infectives y_{t-1} at generation time $t-1$, is equal to $\exp\{-\beta y_{t-1}/N\}$. Therefore, the event that one susceptible became infected in one time step is Bernoulli distributed, and for independent contacts the number of new infectives at generation time t is Binomially distributed according to the probability

$$Y_t | y_{t-1}, x_{t-1} \sim \text{Binomial}\left(x_{t-1}, 1 - e^{-\frac{\beta y_{t-1}}{N}}\right),$$

$$X_t = X_{t-1} - Y_t = X_0 - \sum_{i=1}^{t} Y_i.$$

Assuming that the population size N is large and the quantity $\frac{\beta y_t - 1}{N}$ is small, then $1 - e^{-\frac{\beta y_{t-1}}{N}} \approx \frac{\beta y_{t-1}}{N}$, therefore Poisson approximation of the Binomial distribution yields that we approximately have $Y_t | y_{t-1} \sim \text{Poisson}(\frac{\beta x_{t-1} y_{t-1}}{N})$. Approximating the number of susceptibles x_{t-1} with the total population size N we get

$$Y_t | y_{t-1} \sim \text{Poisson}(\beta y_{t-1}). \tag{2}$$

Here we arrived at the approximation of the branching process to the chain binomial model. The term ν in (1) is included in the model to represent the endemic part of the outbreak, and it ensures that the process will not die out (Held et al.

(2005)). The Poisson distribution (2) assumes equality between the variance and the expected value of a random variable, which is often not the case for disease outcomes. Hence, it is usually replaced by a more flexible Negative Binomial distribution to account for overdispersion. We achieve this by letting

$$\mu_t = \beta y_{t-1},$$

$$\sigma_t^2 = \mu_t + \frac{\mu_t^2}{\psi},$$

where the parameter $\psi > 0$ is estimated from the data. The variance σ_t^2 equals the mean when $\psi \to \infty$. Otherwise the variance is assumed to be larger than the expected value, which is called overdispersion. We observe that the above formulation has an intuitive explanation that on average we can expect the number of infected individuals in the next time step to be equal to the number of infected individuals from the previous time step multiplied by the infection rate, which is also interpreted as the basic reproduction number.

3.2 Model extensions

The simple model (1) has to be extended in order to capture more complicated dynamics of disease spread, adjusting for the temporal trends, seasonal patterns, spatio-temporal dependence, as well as the other covariates. The general model formulation was presented in Meyer et al. (2014) with implementation in the R package **surveillance** as hhh4 models (source). Y_{it} has a Negative Binomial distribution with an additive conditional mean

$$\mu_{it} = \underbrace{\nu_{it} e_{it}}_{\text{endemic}} + \underbrace{\lambda_{it} Y_{i,t-1}}_{\text{autoregressive}} + \phi_{it} \underbrace{\sum_{j \neq i} w_{ji} Y_{j,t-1}}_{\text{spatio-temporal}} \tag{3}$$

and the conditional variance of Y_{it} is $\mu_{it}(1 + \mu_{it}\psi_i)$. If the overdispersion parameter $\psi_i = 0$, the counts have Poisson distributions. The first component of the mean (3) is an endemic part, scaled with the population size (or fraction) e_{it} and, possibly, other socio-demographic covariates. In our case, the population sizes of DeSO zones and other DeSO characteristics are constants over the study period, hence we replace e_{it} with e_i. As in the original formulation, the parameter ν_{it} is an immigration term. Usually, it models the cases emerged from an environmental

reservoir within the region or imported cases. As COVID-19 did not emerge in Västra Götaland, the ν_{it} term represents the immigrated cases from outside of the Västra Götaland region.

The epidemic part is split into two parts: the autoregressive component $\lambda_{it} Y_{i,t-1}$, which quantifies the infection spread within the unit i from the previous day, and the term $\phi_{it} \sum_{j \neq i} w_{ji} Y_{j,t-1}$ which accounts for the adjacent regions' contribution to the infection transmission from the previous day, i.e. a spatio-temporal component. We denote the components by **end** (endemic), **ar** (autoregressive), and **ne** (neighbours). Each component is log-linked to the covariates $z_{it}^{(\cdot)}$:

$$\log(\nu_{it}) = \alpha_i^{(\nu)} + \boldsymbol{\beta}^{(\nu)\top} \boldsymbol{z}_{it}^{(\nu)}, \tag{4}$$

$$\log(\lambda_{it}) = \alpha_i^{(\lambda)} + \boldsymbol{\beta}^{(\lambda)\top} \boldsymbol{z}_{it}^{(\lambda)}, \tag{5}$$

$$\log(\phi_{it}) = \alpha_i^{(\phi)} + \boldsymbol{\beta}^{(\phi)\top} \boldsymbol{z}_{it}^{(\phi)}, \tag{6}$$

where $\alpha_i^{(\cdot)}$ is an intercept, z_{it} are the covariates varying in space in time with the corresponding effects $\boldsymbol{\beta}^{(\cdot)}$. We see that the framework allows to add covariates in each compartment and in the following section we will explain in detail the choice of covariates for each component and possible model extensions. The appealing property of the additive risk modelling is that it allows us to estimate the relative contribution of the 3 sources of the infection: the external, the autoregressive and from the neighbours.

We start with a more simple model with common intercepts $\alpha^{(\nu)}, \alpha^{(\lambda)}, \alpha^{(\phi)}$ in each component (formulas (4), (5), (6)) and then move to model with random area-specific intercepts $\alpha_i^{(\cdot)}$ in each component. The model with area-specific intercepts captures the heterogeneity across the regions, not explained by the other covariates. Random area-specific intercepts are assumed to be correlated or independent. The independent intercepts are assumed Normally distributed:

$$\alpha_i^{(\nu)} \sim \mathrm{N}(\alpha^{(\nu)}, \sigma_\nu^2), \tag{7}$$

$$\alpha_i^{(\lambda)} \sim \mathrm{N}(\alpha^{(\lambda)}, \sigma_\lambda^2), \tag{8}$$

$$\alpha_i^{(\phi)} \sim \mathrm{N}(\alpha^{(\phi)}, \sigma_\phi^2). \tag{9}$$

It is also possible to set a distric-specific overdispersion parameter ψ_i or a common

overdispersion ψ over all districts. But we will not attempt further modelling with correlated intercepts and nor with district-specific overdispersion parameters due to the associated computational burden.

3.3 Spatial weights

The weights w_{ji} in the neighbour components in (3) define the distribution of neighbour influence on the unit i. These weights can be chosen as an indicator function: set 1 if two units are neighbours and 0 otherwise. This simplistic assumption gives worse fit than the weights set according to the power law principle:

$$w_{ji} = \begin{cases} o_{ji}^{-d} & \text{if } j \neq i \,, \\ 0 & \text{if } j = i, \end{cases} \tag{10}$$

where o is the adjacency order of neighbour, so if $o_{ji} = k$ then region i is the k'th order neighbour of region j. It means, that one has to cross the regions' borders k times as the shortest way from region i to region j. The weights are further normalized to

$$w_{ji} = \frac{o_{ji}^{-d}}{\sum_{k=1}^{I} o_{jk}^{-d}}$$

to ensure that for each region j the sum of the weights of all its neighbours is equal to one. The decay parameter d is estimated from the data.

Meyer and Held (2014) demonstrate that the power-law weight model extension, based on the neighbourhood order, improves both the fit and prediction. It was shown that the human travel behaviour is described by a decaying function $f(x) \propto \frac{1}{x^d}$, where x is distance and d is the decay parameter (Brockmann et al., 2006). In the definition in (10) we see that if region i is a distant neighbour of order k of region j, then its weight should be small. Also, a large d implies a quick decay, so that the influence of the distant neighbours is negligible. The power law adjusts for the heavy tail of the human travel distance distribution, and, given the fine resolution of the DeSO units and large fraction of commuters in the region (SCB, data 03/2020), we should definitely include the spatial dependence even on the most distant neighbouring units in our model. Thus, we apply this by setting the upper bound for the adjacency order to maximum equal to 20: `weights =`

`W_powerlaw(maxlag = 20)` in the function `hhh4` in the R package surveil-
lance. The neighbourhood structure of the regions is supplied to the model in
form of an adjacency matrix, computed with the functions `poly2adjmat` and
`nbOrder`.

3.4   Seasonal effects

It is possible to add seasonality in each of the components (4), (5), (6), using a
Fourier series expansion (Held and Paul, 2012):

$$f(t) = A_0 + A_1 \sin(2\pi wt) + B_1 \cos(2\pi wt) + A_2 \sin(4\pi wt) + B_2 \cos(4\pi wt) + ...$$

$$= A_0 + \sum_{n=1}^{\infty} (A_n \sin(2\pi nwt) + B_n \cos(2\pi nwt)).$$

We set the frequency $w = \frac{2\pi}{365}$, as we have daily counts. If we add one sinusoidal
wave and a temporal trend t in the **end** component, then we rewrite the endemic
formulation (4) as

$$\log(\nu_{it}) = \alpha_i^{(\nu)} + \beta_t t + \gamma \sin(wt) + \delta \cos(wt).$$

The model with one sinusoidal wave in **end** component will be a basic reference
model. Adding more waves in other components enables to capture more compli-
cated patterns of seasonal effects:

$$\log(\nu_{it}) = \alpha_i^{(\nu)} + \boldsymbol{\beta}^{(\nu)\top} \boldsymbol{z}_{it}^{(\nu)} + \sum_{s=1}^{S} \left(\gamma_s^{(\nu)} \sin(w_s t) + \delta_s^{(\mu)} \cos(w_s t) \right),$$

$$\log(\lambda_{it}) = \alpha_i^{(\lambda)} + \boldsymbol{\beta}^{(\lambda)\top} \boldsymbol{z}_{it}^{(\lambda)} + \sum_{s=1}^{S} \left(\gamma_s^{(\lambda)} \sin(w_s t) + \delta_s^{(\lambda)} \cos(w_s t) \right),$$

$$\log(\phi_{it}) = \alpha_i^{(\phi)} + \boldsymbol{\beta}^{(\phi)\top} \boldsymbol{z}_{it}^{(\phi)} + \sum_{s=1}^{S} \left(\gamma_s^{(\phi)} \sin(w_s t) + \delta_s^{(\phi)} \cos(w_s t) \right).$$

Here, the S is the number of harmonic waves and the Fourier frequencies are
$w_s = \frac{2\pi s}{365}$. In R we set `addSeason2formula` in the `hhh4` formulation. We

will find the best seasonality pattern by experimenting with different numbers of waves in each compartment and then apply the model selection criteria AIC and BIC (see Section 3.6).

3.5   Socio-demographic covariates

In the study "The influence of overcrowding and socioeconomy on the spatio-temporal spread of COVID-19" (Söderberg et al., 2022) the spatial and spatio-temporal analyses were conducted in order to identify the most influential co-variates amongst the DeSOs demographic characteristics. By using elastic net regularised Poisson regression, the authors found that out of 44 covariates (those included also interaction terms) 21 were chosen as most important for the prediction performance, such as overcrowding, the number of square meters per person, gainfully employed, health care workers, education level, foreign background, personal cars, income. Guided by this set of covariates, we will perform model selection, based on the AIC and BIC criteria. The component structure of the model framework allows adding the covariates into each compartment. We will start with the richest model, with all the covariates included into **end**, **ar** and **ne** and then gradually removing the abundant covariates, aiming to find the most parsimonious model. We excluded the number of personal cars, as it will confound the interpretation. Also, we believe that the age of the inhabitants is an important predictor for the number of cases, hence we include two additional covariates: the number of children and adolescents (0-19 y.o.) in a DeSO, call them Young, and the number of senior people (65+ y.o.) in a DeSO, give them the name Older. Thus, the full set of interesting covariates are: Young, Older, Income, M2 per person, Overcrowded, Education, Employed, Foreigners, Care workers. It was found that the fraction of health care workers worsened the fit, and thus was excluded from the further investigation. The covariate Education is the number of individuals who have at least 3 years of post-secondary education. The covariate Foreigners represents the number of people of non-Swedish background, Employed is the gainfully employed number of people, and Income is the median income per DeSO. According to the study, the variable Overcrowded is defined as more than one person per room in a household, with the exceptions of adult couples in a relationship, children under a certain age or anyone living in a villa, detached or semi-detached house. We

included the covariates on the logarithmic scale.

3.6   Model selection

We will use the AIC and BIC criteria for model comparison:

$$AIC = -2\log(L(\hat{\theta}|\text{data})) + 2p, \tag{11}$$

$$BIC = -2\log(L(\hat{\theta}|\text{data})) + p\log(n), \tag{12}$$

where $\log(L(\hat{\theta}|\text{data}))$ is the log-likelihood and p is the number of estimable pa-
rameters.

We start with a simplifying assumption of identical intercepts across all regions
for each of the component. The first basic model contains time trend and one
sinusoidal wave in **end** with the population fraction as an offset, an intercept term
in **ar** and population fraction as covariate in **ne**. This is achieved by means of the
following R commands:

```
basic <- list(end = list(f = addSeason2formula(~1 + t,
                         period = obj@freq, S=1),
                         offset = population(obj)),
              ar = list(f = ~1),
              ne = list(f = ~ 1+log(pop),
              weights = W_powerlaw(maxlag = 20)),
              family = "NegBin1",
              data = list(pop = population(obj)))

basic_model <- hhh4(stsObj = obj, control = basic)
```

The obtained estimates are the MLEs (maximum likelihood estimates) com-
puted with the quasi-Newton algorithm. For the **basic** model we obtained $AIC =$
651312.8 and $BIC = 651413.5$, which can serve as reference values. Adding one
extra sinusoidal wave so that $S = 2$ in the endemic component, we got slightly
improved results, with $AIC = 651179.4$ and $BIC = 651302.4$. Assuming that
seasonality has an impact on the epidemic components as well as on the endemic
one, we experiment with allowing for sinusoidal waves in the **ne** and **ar** parts.

20

The best model with $S = 2$ in each 3 components yielded $AIC = 650065$ and $BIC = 650289.6$. Modelling seasonality with $S = 3$ resulted in the worsening of both criteria as compared with $S = 2$, which is possibly due to the fact that AIC and BIC penalize for larger numbers of parameters. The higher order waves $S = 4$ and $S = 5$ did improve both AIC and BIC, but not by a lot. We keep the model with $S = 2$ in each component, avoiding overfitting with too many waves, and call this model **seasonal**. In Figure 3 the multiplicative effect of $S = 2$ sinusoidal waves for each component are shown for a one-year period. We observe that the amplitude for the epidemic curves **ar** and **ne** is small with the value of the multiplicative effect being close to 1, whereas the endemic curve (the solid line) has larger amplitude and significant multiplicative effect. The possible interpretation for the slight drop during summer days for the epidemic curves is likely due to people being on vacation. This results in people travelling outside to a larger extent and a reduced number of close contacts, e.g. in schools. The endemic curve shows two peaks around June and November, so that the influence of the imported cases is large during these seasons.

We have 9 covariates which we can add into each component, which yields the model with $AIC = 648256.7$ and $BIC = 648692.7$. Initially, a model with all covariates only in the endemic part was considered. The socio-demographic characteristics were believed to be the driving factors explaining the endemic component, as the endemic component represents the imported cases and the cases not attributed to the epidemic spread. But it turned out that better fits were achieved with the inclusion of all or several covariates in each component. However, nine covariates in each component complicates the interpretation. We should sequentially exclude the abundant covariate from different components. Then we obtain the best possible model, with $AIC = 648290.2$ and $BIC = 648636$, which has the following structure: **end** : Young, Education, Foreigners, **ar**: M2 per person, Young, Education, **ne**: Overcrowded, M2 per person, Older, Income, Education, Foreigners. We call this model **rich** and keep it for further comparison in terms of prediction quality. Now we aim to find the most simple and parsimonious model, yet with AIC and BIC as close as possible to those of the **rich** model. To simplify the interpretation, we will exclude the repeated covariate from the compartment, if it gives the smallest impact. Thus, we arrive at the **greedy** model with: **end**: Ed-

ucation, Foreigners, **ar**: M2 per person, Young, **ne**: Overcrowded, Older, Income, Foreigners, where AIC = 648598.8, BIC = 648900.8.
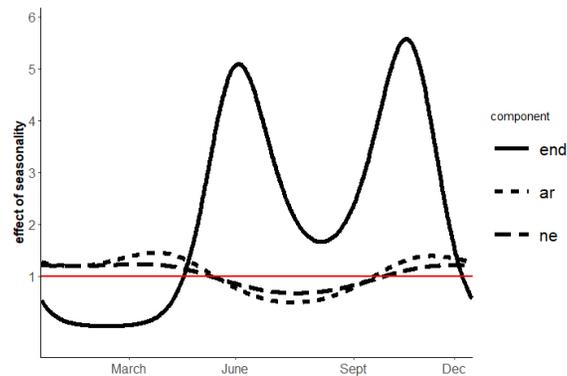

Figure 3: The multiplicative effect of two sinusoidal waves on each component. The effect is estimated for a one year period.

3.7 Goodness of fit

The previous models didn't contain any random effects. It is possible to include random effects as random intercepts according to the distributions (7), (8) and (9) in each compartment. Denote random effects as **ri**. We will consider the following models for comparison,

1. **basic** : a trend + one sinusoidal wave in **end**

2. **basic + ri** : a trend + one sinusoidal wave in **end** + a random effect in each component

3. **seasonal** : a trend in **end** + $S = 2$ waves in each component

4. **seasonal + ri** : a trend in **end** + $S = 2$ waves in each component + a random effect in each component

5. **rich**: a trend + $S = 2$ in each component + rich covariates set

6. **rich + ri**: a trend + S=2 in each component + rich covariates set + a random effect in each component

7. **greedy**: a trend + $S = 2$ in each component + few covariates

8. **greedy + ri**: a trend + $S = 2$ in each component + few covariates + a random effect in each component

With random effects included, the criteria AIC and BIC are not applicable anymore. As the AIC and BIC are computed based on the number of parameters, the inclusion of random effects complicates the estimation of the effective number of parameters in the model. Instead, the fit of the model can be evaluated by scoring rules, computed from the predictions. The scoring rules for count data are the metrics used to measure the discrepancy between the predictive distribution P and the observed count x (Paul and Held, 2011). A smaller scoring rule indicates a better model.

Before we compute the scoring rules for our predictions, we will define the proper scoring rules. Gneiting and Raftery (2007) state that the probabilistic forecasts, formed as predictive distributions, should aim to maximize sharpness of the predictive distributions subject to calibration. Let P be the predictive distribution, and x the value that materializes, then the score $S(P, x)$ is assigned to the forecast P and is seen as a penalty if negatively oriented (the smaller the score, the better) and is subject to minimization. Denote by Q the true probability distribution of the observed values, and by $S(P, Q)$ the expected value of $S(P, \cdot)$ under the distribution Q. A score is proper if

$$S(Q, Q) \leq S(P, Q)$$

for all P and Q. A strictly proper score is one where the equality above is achieved if and only if $P = Q$. With proper score metrics a forecaster is encouraged to report her true belief $P = Q$. Strict proper scores address both sharpness and calibration by penalizing broader prediction intervals (Gneiting and Raftery, 2007). We report the mean scores as a summary measure of the predictions. Before we move to the chosen scoring rules of our predictions, we will visually inspect the calibration of our models by PIT diagrams.

Probability integral transform

The Probability integral transform (PIT) is a measure for the model calibration (proposed by Dawid (1984)), which can be used as a test of whether the given data can be reasonably assumed to be generated from the given distribution. Assume that a continuous random variable X has a cumulative distribution function

$F_X(X)$, then the random variable $Y = F_X(X)$ has a standard uniform distribution. For the predictive distribution P denote its probability mass function as $(p_k)_{k=0}^{\infty}$ and its cumulative distribution function as $(P_k)_{k=0}^{\infty}$ (Czado et al., 2009). If we obtain a predictive cumulative distribution for a given model, then the values of this cumulative distribution, attained at the observations, are PIT values. For the observations generated by the given model the PIT values should follow the standard uniform distribution. If the obtained PIT values plotted as a histogram, exhibit uniformity, it is an indicator of a well-calibrated model. For count data, Czado et al. (2009) proposed a uniform non-randomized version of the PIT histogram. Let x be an observed count, P_x is a predictive distribution for the observed count x and u is a standard uniform. Then define

$$F(u) = \begin{cases} 0, & u \leq P_{x-1}, \\ (u - P_{x-1})/(P_x - P_{x-1}), & P_{x-1} \leq u \leq P_x, \\ 1, & u \geq P_x, \end{cases} \tag{13}$$

for $x \geq 1$, and

$$F(u) = \begin{cases} u/P_0, & u \leq P_0, \\ 1, & u \geq P_0, \end{cases} \tag{14}$$

for $x = 0$. $F(u)$ is a cumulative distribution function, given the observed count x. For each data point $i \in (1, ..., n)$ we compute the predictive distribution $P^{(i)}$. To plot a PIT histogram, choose a number of bins for the histogram equal to J. Having observed the materialized counts x for each i, we compute $F^{(i)}(u)$ according to (13) and (14). Compute the heights of the J histogram bins as $f_j = \bar{F}(\frac{j}{J}) - \bar{F}(\frac{j-1}{J})$, where the mean PIT, aggregated over predictions is $\bar{F}(u) = \frac{1}{n} \sum_1^n F^{(i)}(u)$ for $i = 1, \ldots, n$.

For each of our models, we plot the empirical cdfs for the probabilistic forecasts and visually examine the deviation from the standard uniform distribution (Figure 4). We see that each model shows a PIT histogram being very close to the standard uniform distribution, indicating a good calibration fit. There is, however, a slight decay in the right part of all the histograms, which indicates that the data actually has smaller values than the models tend to predict. Still, the deviation is arguably negligible, and we conclude that there is no apparent miscalibration in any of the models. In particular, we note that the **basic** and the **basic + ri** models are a bit worse calibrated than the others. The models with visibly better calibrations are
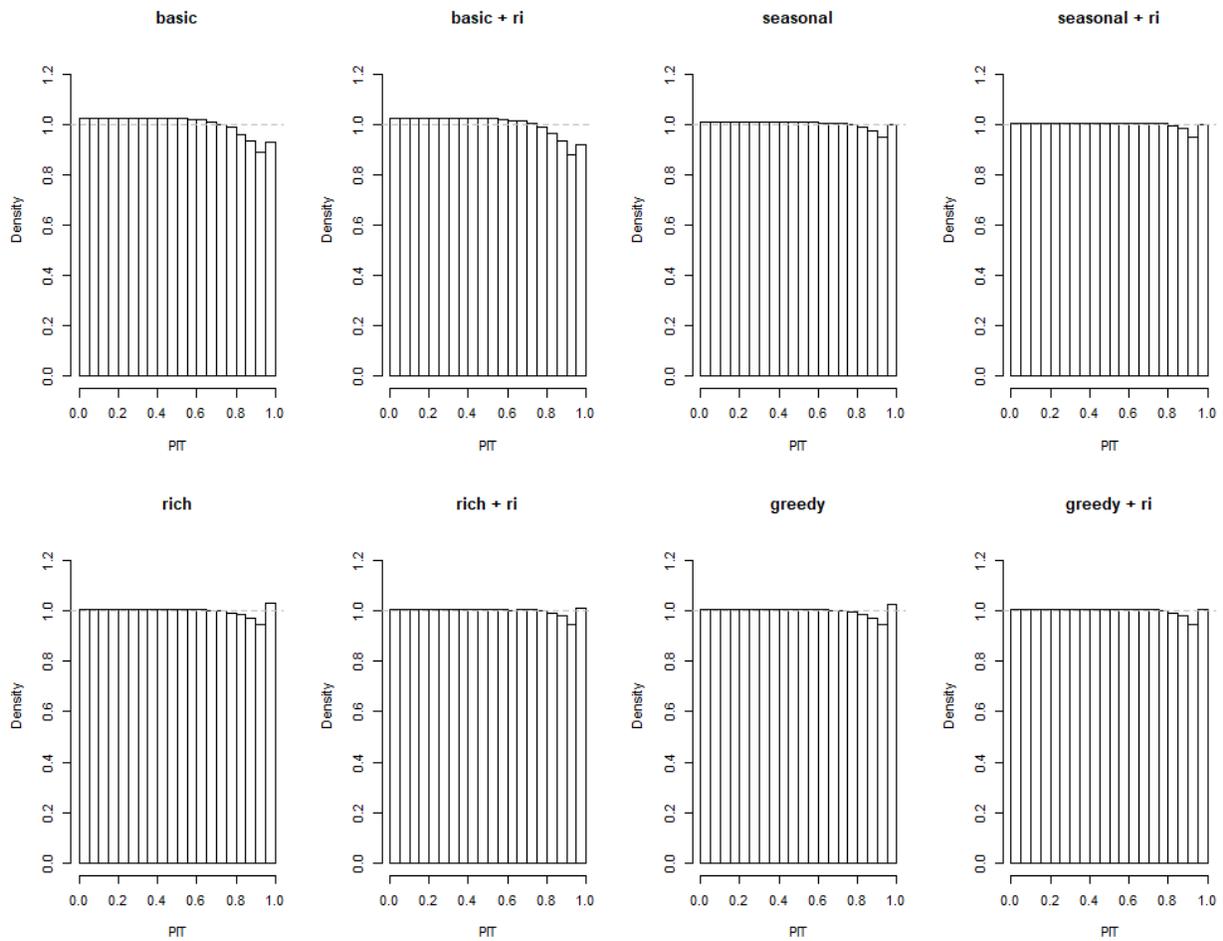
Figure 4: PIT histograms for calibration check for each competing model of endemic-epidemic approach.

seasonal + ri, **rich + ri**, **greedy + ri**.

Scoring rules for count data

We will consider four scoring rules.

The proper scoring rule is a logarithmic score (logs):

$$\log S(P, x) = -\log(p_x)$$

where p_x is the probability mass function at the observation.
The ranked probability score is strictly proper (rps) :

$$\mathrm{rps}(P, x) = \sum_{k=0}^{\infty} \{P_k - 1(x \le k)\}^2.$$

Assume that μ_P is the mean and σ_P is the variance of predictive distribution P. The Dawid-Sebastiani score (Dawid and Sebastiani, 1999) is based on the first two moments of the predictive distribution and is given by

$$\mathrm{dss}(P, x) = \left(\frac{x - \mu_P}{\sigma_P}\right)^2 + 2\log(\sigma_P).$$

We further have the traditional squared error score

$$\mathrm{ses}(P, x) = (x - \mu_P)^2.$$

In the **surveillance** package, with the function `oneStepAhead` we can compute one-day-ahead predictions. The function refits the model for each new prediction, which would be infeasible even for 10 predictions. Instead, we can use the fitted values to compute the scoring rules, hence assessing the goodness-of-fit. The scores are computed as averages across the regions and days.

In the Table 1 we observe that the inclusion of random effects improves the fit for all models. The best scores **logs**, **rps**, **dss**, **ses** are achieved by **seasonal + ri**, **greedy + ri**, **basic + ri** and **greedy + ri** correspondingly. The second best are shown by **greedy + ri**, **greedy +ri**, **seasonal + ri**, **rich + ri**. Thus, we can choose **greedy + ri** as the best performing model in terms of scoring rules. This model

Model	logs	rps	dss	ses
basic	0.6124507	0.2295692	-0.7761596	0.4502016
basic + ri	0.6038401	0.2254187	-0.8568488	0.4360547
seasonal	0.6112801	0.2290956	-0.7839494	0.4495570
seasonal + ri	0.6024040	0.2247373	-0.8466738	0.4347323
rich	0.6095670	0.2281738	-0.7914575	0.4464474
rich + ri	0.6024976	0.2247419	-0.8403673	0.4347016
greedy	0.6098648	0.2283480	-0.7870066	0.4470991
greedy + ri	0.6024615	0.2247151	-0.8393824	0.4345951

Table 1: Averaged scoring rules

incorporates the covariates, with the seasonality and random effects structure being the same as that of **seasonal + ri** model.

3.8 Results

All estimated fixed effects of the **greedy + ri** model are shown in the Appendix in Table 5 (For comparison, all estimated fixed effects of **rich + ri** are shown in Table 6 in the Appendix). Figure 5 represents the estimates $\hat{\beta}_k$ of the fixed effects and their 95% -confidence intervals (the seasonal terms are not shown) on the logarithmic scale. No interval contains zero. For interpretation purposes, we present $2^{\hat{\beta}_k}$ and their confidence intervals in the Table 2.

	Estimate ($2^{\hat{\beta}_k}$)	2.5 %	97.5 %
ar.M2	2.280	1.855	2.802
ar.Young	1.517	1.386	1.660
ne.population	1.698	1.606	1.794
ne.Overcrowded	1.032	1.019	1.045
ne.Foreigners	1.097	1.068	1.126
ne.Older	1.052	1.020	1.086
ne.Income	1.242	1.165	1.324
end.t	0.998	0.997	0.998
end.Education	1.243	1.184	1.304
end.Foreigners	1.246	1.194	1.299

Table 2: Estimates of fixed effects with CIs on the natural scale, **greedy + ri**

This means, that if the value of the covariate k is doubled, then we expect a $2^{\hat{\beta}_k}$ multiplicative effect on the number of cases. For example, the number of endemic incidences are estimated to be multiplied by 1.246 on average if the number of foreigners is doubled. Surprisingly, if the number of people with 3 or more years
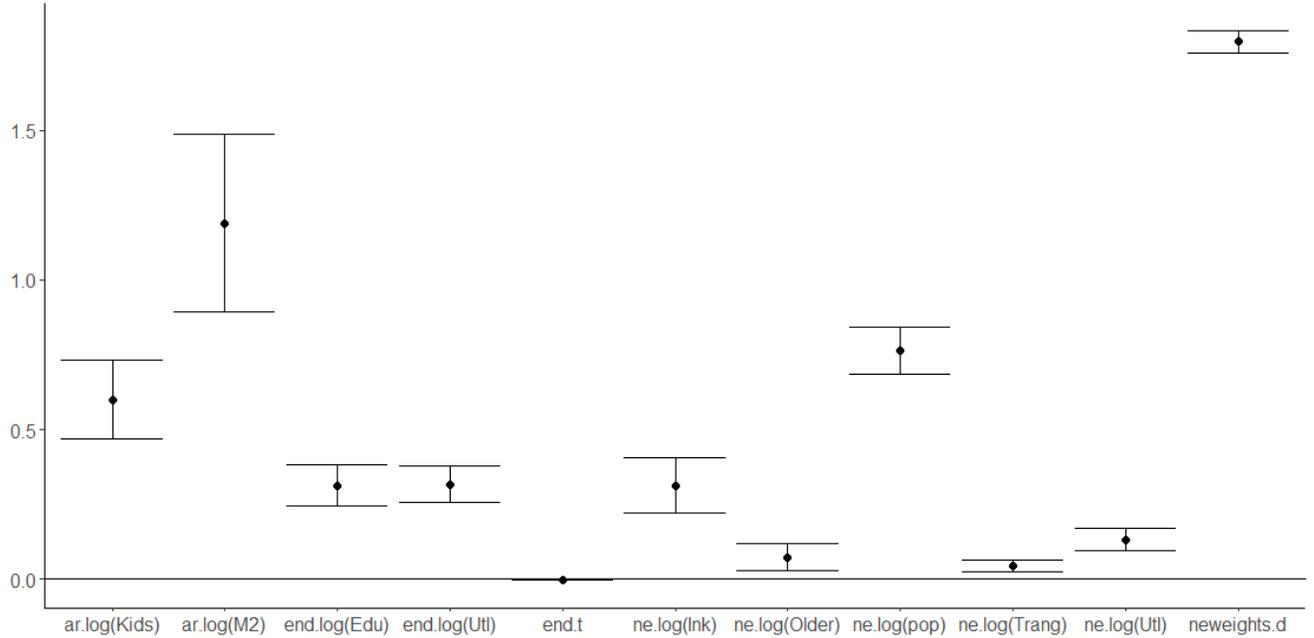
Figure 5: Confidence intervals for the fixed effects, **greedy + ri** model.

of post-secondary education is doubled, then the endemic incidence is also increased by 1.243 on average. For the epidemic part (**ar** and **ne**), if the number of people living in overcrowded conditions, as well as foreigners and older people are doubled, then we expect on average the multiplicative effect of 1.032, 1.097, 1.052 correspondingly. Also, the doubled number of square meters per person is associated with the average multiplicative effect of 2.280 on the autoregressive component, which was not expected. Additionally, the decay parameter d was estimated to be $1.80(1.76, 1.83)$, which is interpreted as a weak decay of spatial influence from the distant neighbours.

The fitted mean components, aggregated over all regions, together with the observed counts as dots are shown in Figure 6. We see that the largest contributor of disease cases is the neighbour or spatio-temporal (orange) component. Also, in the second wave of disease spread, which started in the autumn 2020, the (blue) autoregressive component is more apparent than in the first wave. The plot 6 returns also an object, containing the fitted mean values together with their estimated components for each day. For example, for the 1st of May 2020 we have
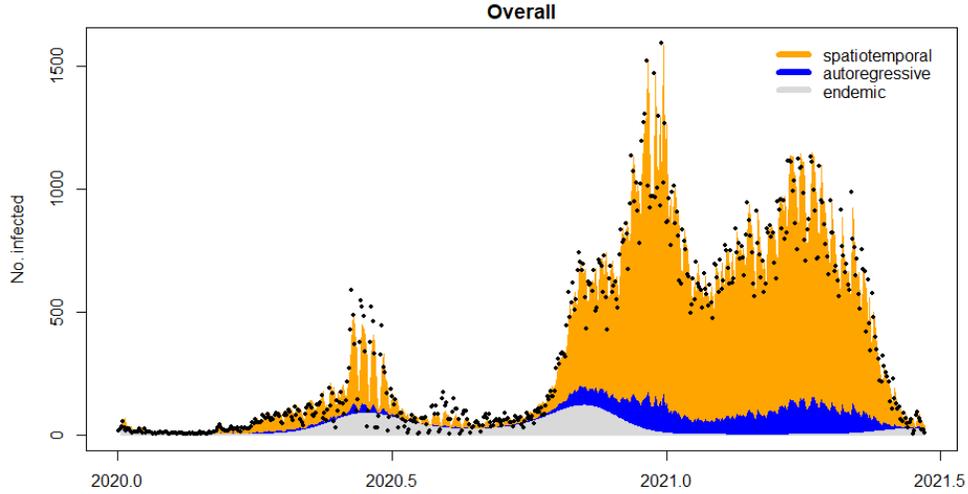
Figure 6: The fitted mean components for the **greedy + ri** model, aggregated over all regions.

mean	epidemic	endemic	epi.own	epi.neighbours
76.471412	58.806892	17.664520	6.762425	52.044466

Then we can compute the proportion of each component contribution to the mean for each day. The average proportions over all days are : **end** = 0.080782, **ar** = 0.104965, **ne** = 0.814253. 92 % of the cases are explained through the epidemic components, of which 81.4% refers to the spatio-temporal component. This reflects the highly viral nature of COVID-19. We present the similar plots of the fitted mean components for each of the considered models in the Appendix (Figures 23, 24, 25, 26, 27, 28, 29). Also, we plot the map of each component contribution to the mean in Figure 7. We see no regions with a spatio-temporal contribution less than 60 % into the mean number of cases, whereas the autoregressive proportion is never higher than 0.12 %. This means, that the reproduction of the disease within the DeSO is very small.

We included random intercepts in our model to account for the unobserved heterogeneity across the regions. If $\alpha^{(\cdot)}$ is the mean intercept of the component, and $\alpha_i^{(\cdot)}$ is a region-specific intercept, we can plot the estimated exponentially transformed deviations $\alpha_i^{(\cdot)} - \alpha^{(\cdot)}$ for each region in Figure 8. These deviations
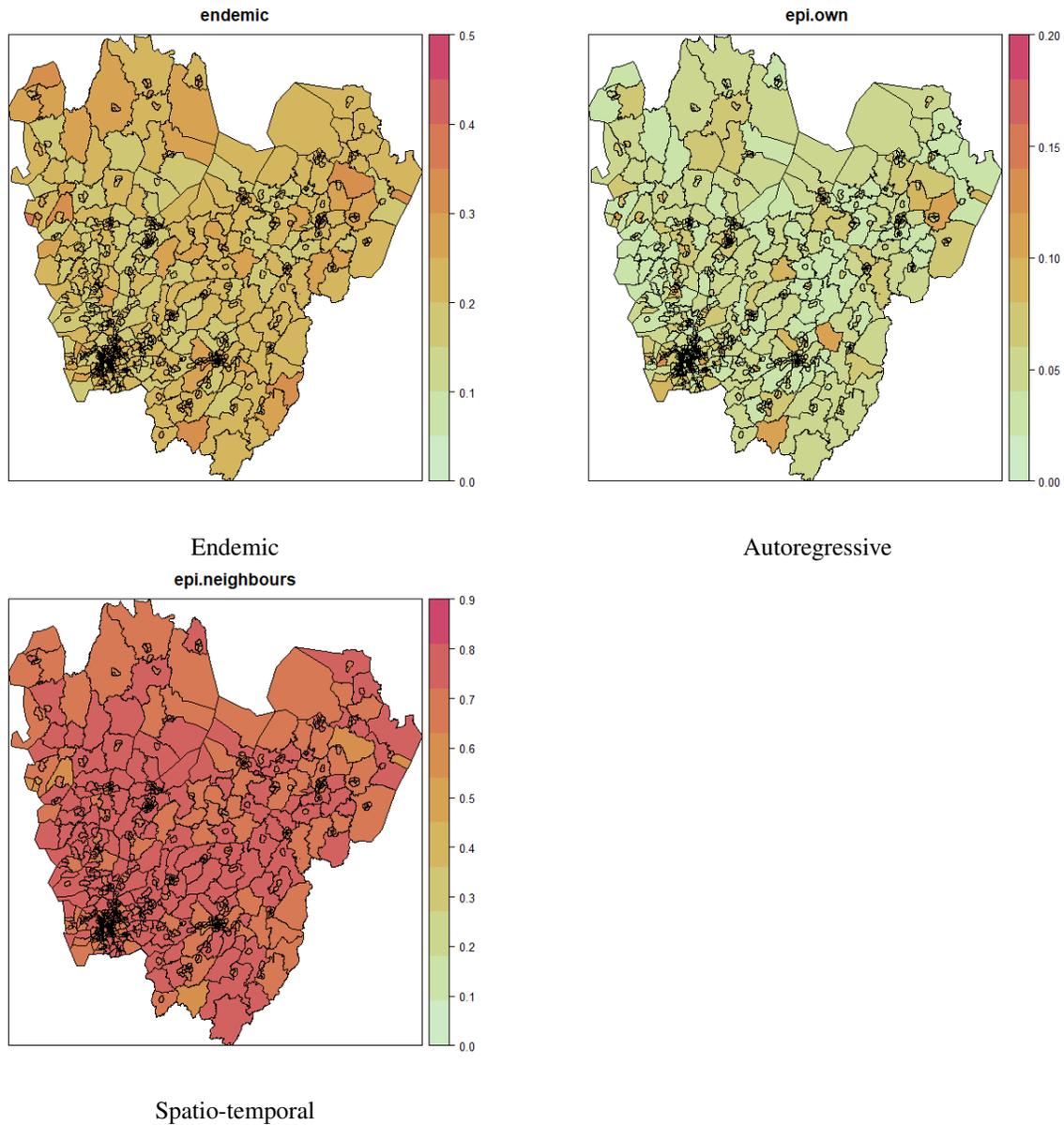
Endemic

Autoregressive

Spatio-temporal

Figure 7: Proportions of the fitted components, averaged over all days for **greedy + ri** model.

represent the multiplicative effect on each component. We see that the deviations for the spatio-temporal component are large for the border DeSOs such as the South-Eastern part of the whole region, but this is explained by the fact that the actual neighbours of the border DeSOs (regions Jönköping, Halland, Örebro) were not included into the analysis. Looking closely at the regions with highest multiplicative effect on the endemic components, we observe that those are the

central regions (svenska: centralort) such as Göteborg, Borås, Skövde, Tibro. The multiplicative effect on the autoregressive part is mostly pronounced in the central regions such as Lysekil, Trollhättan, Falköping, Göteborg, Tibro.
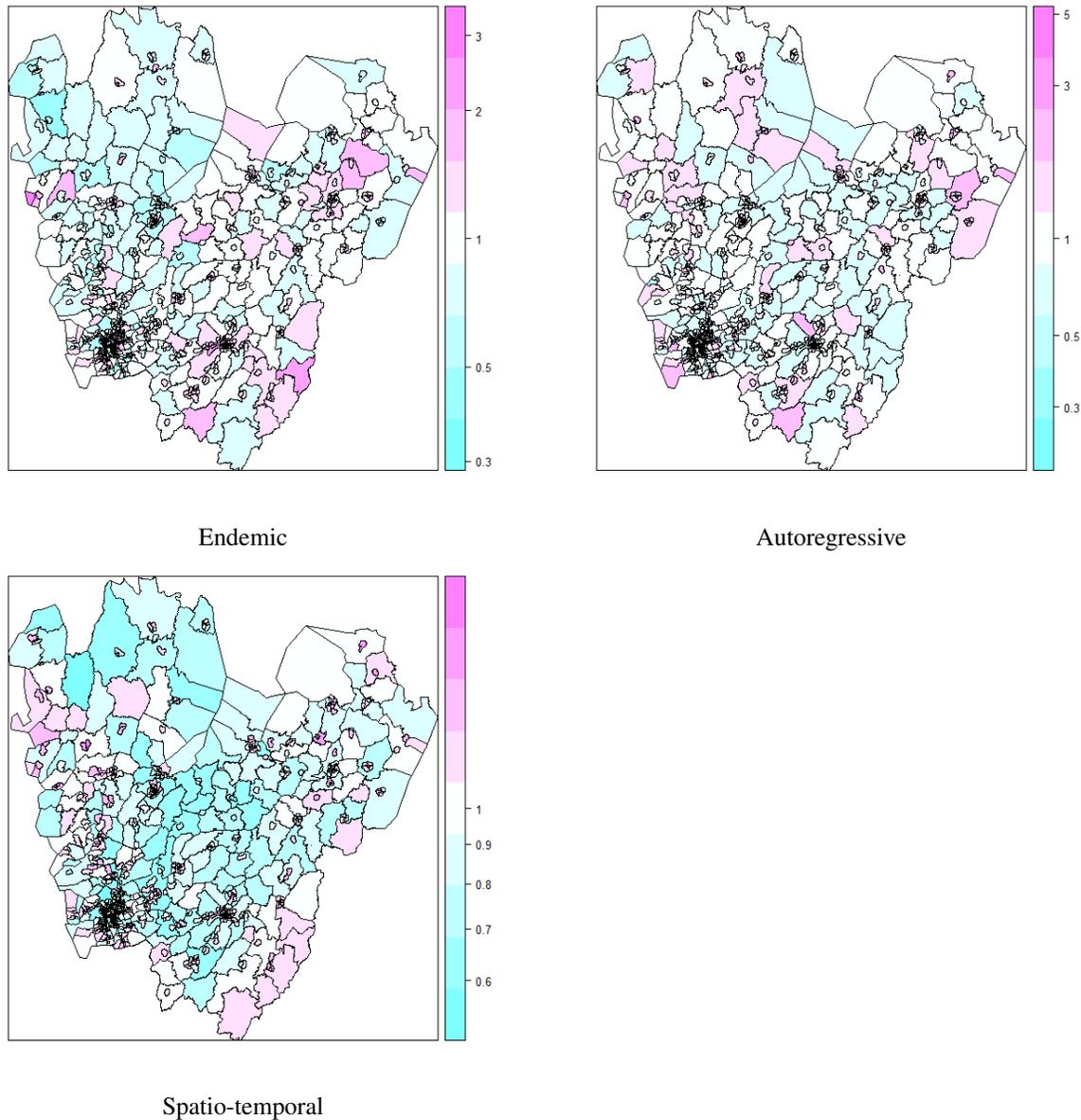

Endemic

Autoregressive

Spatio-temporal

Figure 8: The exp-transformed estimated intercept deviations $\alpha_i - \alpha$, which are the multiplicative effects of district-specific heterogeneity on the components of **greedy + ri**.

4 INLA

4.1 Preface

INLA, which stands for Integrated Nested Laplace Approximations is a technique developed for a particular class of regression models, namely latent Gaussian models (LGMs), which are a subclass of the generalized additive models (GAMs). INLA was proposed in the works of Rue and Martino (2007), Martino and Rue (2009), Rue et al. (2009) and developed further in Martins et al. (2013), Rue et al. (2017) and Chiuchiolo et al. (2022). Classical MCMC (Markov chain Monte Carlo) sampling methods are widely used in Bayesian statistics, in particular in spatio-temporal statistical setups, like the ones initially intended to be employed for the COVID-19 data modelling. Unfortunately, for very large models, these methods can be infeasible due to the associated heavy computational burden. Our observation data contains 537*992 = 532704 points, and here the associated MCMC algorithms took more than two days to obtain a small sample of size 500 (after burn-in period). The chains showed poor mixing and didn't converge, so that a larger sample would be needed for more conclusive results. Since it was infeasible, it was decided to switch to the INLA approach. INLA employes a deterministic algorithm instead of posterior sampling, and its computational speed and high accuracy make it more suitable for large datasets. INLA-package is available on the website https://www.r-inla.org/.

We start this part with an introduction of the Laplace approximation method in Section 4.2 and a brief description of several properties and computational benefits of the Gaussian Markov Random Fields (GMRFs) in Section 4.3. After that we will look into details on how INLA works in Section 4.4. The application of INLA for our problem and the obtained results are presented in the Section 4.5.

4.2 Laplace approximation

Laplace approximation is a method for approximating integrals. Let $\log f(x)$ to be a log density function, and denote the mode of this density by $x^* = \mathrm{argmax}_x \log f(x)$.

We write the Taylor series expansion of order two evaluated at the mode as

$$\log f(x) \approx log f(x^*) + (x - x^*)\frac{\partial \log f(x)}{\partial x}\bigg|_{x=x^*} + \frac{(x - x^*)^2}{2}\frac{\partial^2 \log f(x)}{\partial^2 x}\bigg|_{x=x^*}.$$

The second term of the series expansion turns into 0 at the mode. We can write the integral of the density $f(x)$ as

$$\int f(x)dx = \int \exp(\log f(x))dx$$

$$\approx \int \exp\left(\log f(x^*) + \frac{(x - x^*)^2}{2}\frac{\partial^2 \log f(x)}{\partial^2 x}\bigg|_{x=x^*}\right)dx.$$

We observe that the second quadratic term in the integrand above reminds of the density of a Normal distribution. If we set $\sigma^{*2} = -1/\frac{\partial^2 \log f(x)}{\partial^2 x}\bigg|_{x=x^*}$, then we can rewrite the approximation

$$\int f(x)dx \approx \exp(\log f(x^*)) \int \exp\left(-\frac{(x - x^*)^2}{2\sigma^{2*}}\right)dx.$$

Thus, in order to obtain a Normal approximation of the given density by the Laplace method, we need to find the mode by setting the first derivative to zero, and the variance is found by evaluating the negative inverse of the second derivative at the mode. The general setting of the method for approximating an integral I_n is the following:

$$I_n = \int \exp(nf(x))dx$$

$$\approx \int \exp\left(n\left(f(x^*) + \frac{1}{2}(x - x^*)^2 f''(x^*)\right)\right)dx$$

$$\approx \exp(nf(x^*))\sqrt{\frac{2\pi}{-nf''(x^*)}} = \tilde{I}_n.$$

As $n \to \infty$, the approximation will be exact, if the central limit theorem holds. This approximation has en error satisfying

$$I_n = \tilde{I}_n(1 + \mathcal{O}(n^{-1})).$$

which is not an additive, but a relative error with rate n^{-1} (Rue et el., 2016).

Assume we have some data **y**, and the parameter $\boldsymbol{x} = (x_1, x_2)$, controlling the distribution of **y**. We would like to compute the posterior marginal densities for the parameters. The marginal posterior density for x_1 can be approximated by the Laplace method as

$$\pi(x_1|\boldsymbol{y}) = \frac{\pi(x_1, x_2|\boldsymbol{y})}{\pi(x_2|x_1, \boldsymbol{y})} \approx \frac{\pi(x_1, x_2|\boldsymbol{y})}{\tilde{\pi}(x_2|x_1, \boldsymbol{y})},$$

where $\tilde{\pi}(x_2|x_1, \boldsymbol{y})$ is the Gaussian approximation to the full conditional density of x_2. In the same way we can approximate the posterior marginal for x_2. Of course, we could make the Laplace approximations directly to the $\pi(\boldsymbol{x}|\boldsymbol{y})$ which would be two-dimensional in this case, and then compute the marginal from it, but such approximations often might be very off. Instead, we obtain the approximations to the conditional distributions, which turn out to be more accurate. Such marginals are closer to Gaussian densities than their joint distributions. Before we look into details of how INLA produces the approximations, we should briefly introduce several important concepts about GMRFs.

4.3 Gaussian Markov Random fields

A Gaussian Markov Random fields (GMRF) is a random vector $\boldsymbol{x} = (x_1, ..., x_n)$ that follows a Gaussian or multivariate Normal distribution with some mean $\boldsymbol{\mu}$ and the precision matrix $\boldsymbol{Q}$, which is an inverse of the covariance matrix; and it has a Markov property: x_i and x_j are conditionally independent, given all other indices except for the i and j. Write it as $x_i \perp x_j | x_{-ij}$. The concept of conditional independence is encoded in the precision matrix of the GMRF: the elements of the precision matrix are zero for conditionally independent elements. Thus, the precision matrices for GMRFs are sparse, which decreases the computational burdens substantially. The covariance matrix of such a field, on the other hand, may be a dense matrix and can be computationally too expensive to work with directly if the data dimension is large. The sparsity of the precision matrix is one of the key attributes of the INLA method.

We state the formal definition of a GMRF and an important theorem from (Rue and Held, 2005). Define a labelled graph $G = (V, E)$, where V are vertices $V = (1, ..., n)$ and the edges E connect the vertices. There is no edge between the vertices x_i and x_j iff they are conditionally independent, given x_{-ij}.

Def. A random vector $\boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ is called a GMRF with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$ wrt a labelled graph $G = (V, E)$, iff its density has the form

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2}\exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})\Big), \qquad (15)$$

where

$$Q_{ij} \neq 0 \iff \{i, j\} \in E \text{ for all } i \neq j.$$

Theorem. Let **x** have a Normal distribution with mean μ and a precision matrix $\boldsymbol{Q} > 0$. Then for for $i \neq j$

$$x_i \perp x_j | x_{-ij} \iff Q_{ij} = 0.$$

The proof of the theorem is in the Appendix section.

This result means that we can read off from the graph which elements are zeros in the precision matrix. For a GMRF the pairwise Markov property, defined above, is equivalent to the local Markov property (conditional independence, given the neighbors) and a global Markov property. The global Markov property is a stronger concept, implying the local and pairwise properties. The global Markov property Rue and Held (2005):

$$\boldsymbol{x}_A \perp \boldsymbol{x}_B | \boldsymbol{x}_C,$$

for all disjoint sets A, B, C, where C separates A and B, and A and B are non-empty (see Figure 9) The simplicity of the computation of the precision matrix **Q** is proven by the following results: Let **x** be a GMRF as in (15), and let **L** be a Cholesky lower triangular matrix of the matrix **Q**. For the vertices $1 \leq i < j \leq n$ define the set

$$F(i, j) = \{i + 1, ..., j - 1, j + 1, ..., n\},$$

then

$$x_i \perp x_j | \boldsymbol{x}_{F(ij)} \iff L_{ji} = 0.$$

Hence, the global Markov property implies that if $F(i, j)$ is a separating set for $i < j$, then $L_{ji} = 0$. So the sparsity of precision matrix **Q** is inherited by the Cholesky triangles **L**. If we have a graph with vertices and edges, which in our
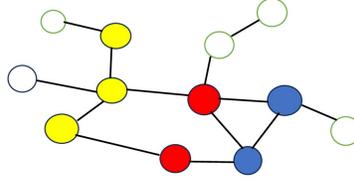
Figure 9: The yellow set A and the blue set B are conditionally independent, given the red set C.

case is a map with the regions and their neighborhood structure, then we know which elements must be zero in $\mathbf{L}$ and we will save time by not computing them. But what is the interpretation of the non-zero elements L_{ji}? Assume we want to sample $x \sim N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ by the following algorithm Rue and Held (2005):

- Compute the Cholesky factorisation $\boldsymbol{Q} = \boldsymbol{LL}^T$

- Sample $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})$

- Solve $\boldsymbol{L}^T \boldsymbol{v} = \boldsymbol{z}$

- Compute $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{v}$

- Return $\boldsymbol{x}$

Since $\boldsymbol{L}^T$ is an upper triangular matrix, the equation $\boldsymbol{L}^T \boldsymbol{v} = \boldsymbol{z}$ is solved by back substitution:

$$
\begin{bmatrix}
L_{11} & L_{12} & \cdots & L_{1n} \\
0 & L_{22} & \cdots & L_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & L_{nn}
\end{bmatrix}
\times
\begin{bmatrix}
v_1 \\
v_2 \\
\cdots \\
v_n
\end{bmatrix}
=
\begin{bmatrix}
z_1 \\
z_2 \\
\cdots \\
z_n
\end{bmatrix}
$$

$$
v_n = \frac{z_n}{L_{nn}},
$$

$$
v_{n-1} = \frac{z_{n-1}}{L_{n-1,n-1}} - \frac{L_{n,n-1}}{L_{n-1,n-1}} v_n,
$$

$$\ldots$$

We observe that this solution by the back substitution method allows us to define a GMRF **x** wrt to a labelled graph G with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{Q}$ backward in time or in indices by specifying the expectation and precision, conditional on the future

$$E(x_i|\boldsymbol{x}_{(i+1):n}) = \mu_i - \frac{1}{L_{ii}} \sum_{j=i+1}^{n} L_{ji}(x_j - \mu_j),$$

$$Prec(x_i|\boldsymbol{x}_{(i+1):n}) = L_{ii}^2,$$

where $i = 1, ..., n$ are the vertices or the nodes in the graph G. In our case the regions of the map represent the vertices or the nodes. To simplify the computations further, it is possible to permute the nodes to make the matrix as sparse as possible. In general, banded matrices are the fastest in computations, because they have elements on diagonals and sub-diagonals, and zero everywhere else. There is no natural ordering for the regions, so we can assign the order of nodes in such a way to make the matrix a band matrix. Also, the Cholesky triangle **L** will inherit the bandwidth of the precision matrix **Q**. It is possible to find the best order of the nodes by listing out all $n!$ possible permutations of the matrix **Q**. This method is obviously infeasible for large n. Instead, we can apply a reordering based on a nested dissection algorithm. Select a small set of nodes such that it divides the whole graph into two completely independent sub-graphs, reorder the nodes and proceed division of the sub-graphs further (see details in (Rue and Held, 2005)).

4.4 INLA settings

4.4.1 Hierarchical LGM model

Latent Gaussian models (LGMs) constitute a particular class of models, where a latent field follows a Normal distribution. Each observation y_i from $\boldsymbol{y} = (y_1, ...y_n)$ is assumed to belong to the exponential family, where the parameter ϕ_i (usually the mean μ_i) is linked to the predictior η_i by the link function $g(\phi_i) = \eta_i$. The linear predictor η_i is called latent, because it is usually unobserved or partially observed. The η_i has an additive structure, allowing to introduce various covariates

of different forms as

$$\eta_i = \beta_0 + \sum_{j=1}^{J} \beta_j z_{ij} + \sum_{k=1}^{K} f_{k,j_k(i)}. \tag{16}$$

Here, β_0 is an intercept, $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_J)$ are covariates with corresponding fixed effects $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)$, $\boldsymbol{f} = (f_1, ..., f_K)$ is a collection of independent model components which can represent spatial, spatio-temporal effects, autoregressive models, splines, random effects and more. A noise term ϵ_i can also be added in the predictor specification (16). We assign a Gaussian prior to all the components of the model, except for the hyperparameters. Then the linear predictor η_i must be Gaussian as well. Denote by $\boldsymbol{x} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{\eta}\}$, the vector of all Gaussian parameters (we use the notation according to the paper of Rue et al. (2009)). Hence, $\mathbf{x}$ is assigned a joint Gaussian prior distribution with a (often) zero mean and a precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_2)$. The hyperparameters control the effects of the model, they are typically variances, precisions, correlations. Write the hyperparameters as one vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ is the hyperparameters controlling the distribution of data $\mathbf{y}$ and $\boldsymbol{\theta}_2$ is the hyperparameters of the latent field $\mathbf{x}$. We assume it follows a joint prior or a product of prior distributions. The hierarchical LGM model is defined as three-staged

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1 \sim \prod_i \pi(y_i|x_i, \boldsymbol{\theta}_1),$$
$$\boldsymbol{x}|\boldsymbol{\theta}_2 \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}_2) = N(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)),$$
$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

The first line reads as y is conditionally independent, given the hyperparameters $\boldsymbol{\theta}_1$ and we assume that each data point y_i is connected to the latent field via only one element x_i. Only some x are observed in the latent field $\mathbf{x}$. The second stage of the above specification is the Normal distribution of the latent field with the mean $\boldsymbol{\mu}(\boldsymbol{\theta}_2)$ and the precision matrix $\boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)$. The third stage is the prior distribution of the hyperparameters $\boldsymbol{\theta}$.

The key factors that make INLA work fast and accurate are the following: the dimension of the $\boldsymbol{\theta}$ must be small, usually 2-6, but not more than 20. If dimension

of **x** is large, say, around $n = 10^3 - 10^5$, then **x** must be a GMRF. We can write the joint posterior density for the latent field components and hyperparameters as :

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) \pi(\boldsymbol{x} | \boldsymbol{\theta_2}) \prod_i \pi(y_i | x_i, \boldsymbol{\theta_1}) \tag{17}$$

$$\propto \pi(\boldsymbol{\theta}) |\boldsymbol{Q}(\boldsymbol{\theta_2})|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta_2})\boldsymbol{x} + \sum_i \log\{\pi(y_i | x_i, \boldsymbol{\theta_1})\}\right) \tag{18}$$

The posterior marginals for each parameter are computed by integrating out other parameters from the corresponding densities

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

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

From the marginals of the parameters we can obtain any summary of interest: means, variances, quantiles. The INLA method constructs the Laplace approximations to the posterior marginals

$$\tilde{\pi}(\theta_j | \boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta} | \boldsymbol{y}) d\boldsymbol{\theta}_{-j}, \tag{19}$$

$$\tilde{\pi}(x_i | \boldsymbol{y}) = \int \tilde{\pi}(x_i | \boldsymbol{\theta}, \boldsymbol{y}) \tilde{\pi}(\boldsymbol{\theta} | \boldsymbol{y}) d\boldsymbol{\theta}. \tag{20}$$

where $\tilde{\pi}(\cdot)$ stands for an approximation. The marginals $\tilde{\pi}(\theta_j | \boldsymbol{y})$ can be computed by integrating out θ_{-j}, as the dimension of θ is relatively small. But instead, a more computationally efficient approach is used here. The marginals $\tilde{\pi}(x_i | \boldsymbol{y})$ are computed by numerical integration by summing over θ with the area weights Δ_k

$$\tilde{\pi}(x_i | \boldsymbol{y}) = \sum_k \tilde{\pi}(x_i | \boldsymbol{\theta_k}, \boldsymbol{y}) \tilde{\pi}(\boldsymbol{\theta_k} | \boldsymbol{y}) \Delta_k. \tag{21}$$

4.4.2 Details of approximation

We will look closer into details of approximations (19) and (20). For simplicity, write $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$. Firstly, we need to approximate the joint posterior of the hyperparameters as follows:

$$\pi(\boldsymbol{\theta} | \boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x} | \boldsymbol{\theta}, \boldsymbol{y})} \approx \frac{\pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \pi(\boldsymbol{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\tilde{\pi}_G(\boldsymbol{x} | \boldsymbol{\theta}, \boldsymbol{y})}\bigg|_{x=x^*(\boldsymbol{\theta})} := \tilde{\pi}(\boldsymbol{\theta} | \boldsymbol{y}), \tag{22}$$

where $x = x^*(\boldsymbol{\theta})$ is the mode of the full conditional for **x**, and the denominator $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is its Laplace approximation. As $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ has a Laplace prior distribution, then $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is close to Gaussian. Therefore, it's Gaussian approximation should be reasonably accurate. It has the following form:

$$\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\Big(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_i \log \pi(y_i|x_i, \boldsymbol{\theta}) \Big) \tag{23}$$

$$\propto |\boldsymbol{P}(\boldsymbol{\theta})^{1/2}| \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T \boldsymbol{P}(\boldsymbol{\theta})(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) \Big), \tag{24}$$

where $\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{Q}(\boldsymbol{\theta} + diag(\boldsymbol{c}(\boldsymbol{\theta})))$, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the location of the mode, $\boldsymbol{c}(\boldsymbol{\theta})$ contains the negative second derivatives of the log-likelihood at the mode (Rue et al., 2016). So the contribution from the likelihood enters the expression of the density only in the diagonal of the precision matrix $\boldsymbol{P}(\boldsymbol{\theta})$, whereas the dominating term is the contribution from the Gaussian prior (Eq. (23)). This approximation is also a GMRF with respect to the same graph. Hence, the Markov property of the precision matrix $\boldsymbol{Q}$ is preserved in the matrix $\boldsymbol{P}$, implying the high computational speed even for non-Gaussian data. If $\boldsymbol{y}$ has Gaussian distribution, the approximation will be exact.

The density $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ departs from Gaussian, and in order to compute it's normalizing constant, Rue et al. (2009) suggest an approach, which is based on numerical exploration of the $\boldsymbol{\theta}$ space to find good points for further integration. The following illustrations of the INLA method are inspired by Rue et al. (2009) and Blangiardo and Cameletti (2015).

1. Locate the mode θ^* of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ by optimizing $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$ wrt θ by, for example, the Newton-Raphson method.

2. At the mode θ^* compute the negative Hessian **H**. Set the covariance matrix for $\boldsymbol{\theta}$ as $\boldsymbol{\Sigma} = \boldsymbol{H}^{-1}$. Do the eigendecomposition of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$. Define a new standardized variable **z** with mutually orthogonal components

$$\boldsymbol{\theta}(\boldsymbol{z}) = \boldsymbol{\theta}^* + \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{z}$$

3. Explore the density of $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$ to detect the regions, where the bulk of probability mass is located. For visualisation purposes, let $\boldsymbol{\theta}$ space be two-dimensional. In Figure 10 the new coordinate axes z_1 and z_2 and the contour

plot of the $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$ are shown, together with the points with substantial probability mass. Starting from the mode, where $\mathbf{z} = 0$, we move along the axis with some step length, δ_z. Then the other axis is explored in the same way, yielding the black points in the Figure 10(a). The grey points are the intermediate combinations of the black points. We collect the points $\boldsymbol{\theta}_k$ together with the area weights Δ_k, which we will need for the numerical integration (equation (21)). This method is called the grid strategy (Figure 10 (a)). When the dimension of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ grows, and more points are built, the computational costs grow exponentially. The other approach was developed to overcome that difficulty, a central composite design strategy (CCD) to be applied when the dimension of the hyperparameters is larger than 2. Instead of building a regular grid, the CCD method selects the points (using the mode and the Hessian) and evaluates the curvature. These few but carefully selected points allow to explore the space of θ with less computational burden (Figure 10(b)).

4. The marginal densities $\tilde{\pi}(\theta_j|\boldsymbol{y})$ are obtained by constructing an interpolant, evaluated at the collected points $\boldsymbol{\theta}_k$. The density is to be normaliazed numerically.
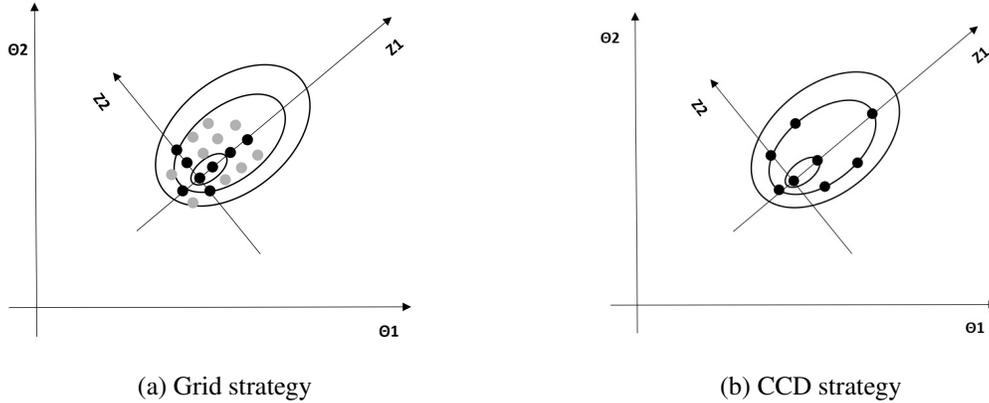

(a) Grid strategy   (b) CCD strategy

Figure 10: Illustration of the exploration of $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$. The contour plots with the new axes z_1 and z_2 are shown. The marked points are those with a significant probability mass.

Secondly, we need to approximate the densities $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ which later will be used in the numerical integration to obtain the marginals for the components of the latent field. There are several ways to compute these marginals: the Gaussian,

the Laplace and the simplified Laplace. The simplest way, the Gaussian method, is to compute the marginals from $\pi_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ (Eq. 23). This method is fast but can give errors in the location and the absence of skewness. The second method, the Laplace approximation uses

$$
\pi(x_i|\boldsymbol{y}, \boldsymbol{\theta}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \approx \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\tilde{\pi}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})}\bigg|_{\boldsymbol{x}_{-i}=\boldsymbol{x}_{-i}^*(\boldsymbol{\theta}, x_i)} := \tilde{\pi}(x_i|\boldsymbol{y}, \boldsymbol{\theta}).
$$
(25)

Here, the denominator is a Laplace approximation with a mode $\boldsymbol{x}_{-i}^*(\boldsymbol{\theta}_{x_i})$. This method gives good results, but is computationally more expensive, as it has to recompute the approximations for each point of θ and x_i, which can be too expensive for large $\boldsymbol{x}$. The next cheaper method, which is a default method as well, is a simplified Laplace strategy. The numerator and the denominator of the expression (25) are expanded as Taylor series around $x_i = \mu_i(\theta)$. The third term of the Taylor expansion corrects for the skewness of the distribution. In the recent paper Chiuchiolo et al. (2022), the authors propose an Extended Simplified Laplace method with the fourth order Taylor expansion.

4.4.3 Example

We will look at a simple graphical example of how INLA works, taken from the book of Blangiardo and Cameletti (2015). We keep the notations in accordance with Rue et al. (2009) as before. Assume $\boldsymbol{y} = (y_1, ..., y_n)$ are independent and have marginal Normal distributions with $y_i \sim N(\mu, \sigma^2)$. Set independent priors for μ and $\theta = 1/\sigma^2$:

$$
\mu \sim Normal(\mu_0, \sigma_0^2),
$$
$$
\theta \sim Gamma(a, b).
$$

Assume $x_i = \mu$, so that each data point y_i is connected to x_i, which is Gaussian. In this particular case, the full conditional distribution $\pi(x|\theta, \boldsymbol{y})$ has the Gaussian form $N(x_n, \sigma_n^2)$, where

$$
x_n = \frac{\theta \sum y_i + \frac{\mu_0}{\sigma_0^2}}{n\theta + \frac{1}{\sigma_0^2}} \quad \text{and} \quad \sigma_n^2 = \frac{1}{n\theta + \frac{1}{\sigma_0^2}}.
$$

To find the posterior for the hyperparameter θ we use eq. (22) and get

$$\pi(\theta|\boldsymbol{y}) \propto \left.\frac{\pi(\boldsymbol{y}|x,\theta)\pi(x)\pi(\theta)}{\pi(x|\theta,\boldsymbol{y})}\right|_{x=x_n} = \left.\frac{1}{1/\sqrt{2\pi\sigma_n^2}}\pi(\boldsymbol{y}|x,\theta)\pi(x)\pi(\theta)\right|_{x=x_n},$$

where the equality followes from the evaluation of the density $\pi(x|\theta,\boldsymbol{y})$ at the mode x_n which resulted in $1/\sqrt{2\pi\sigma_n^2}$. In general, INLA computes Laplace Normal approximation for the denominator as shown in the equation (23). But in this simple example we know that the denominator is Gaussian, because y_i has a Normal distribution. We still need to compute the normalizing constant. As performed by INLA, the space of θ is explored by computing the points θ_k and evaluating the density at each of them, yielding the weights Δ_k. Starting from the mode, we evaluate the curvature at the mode, and the exploration proceeds further until the main bulk of density masses are found, as shown on Figure 11. Thus we find the black points θ_k, evaluate the shape and obtain the marginal distribution for θ by an interpolation (Figure 11). The normalizing constant will be computed from the interpolated marginal. At the second step we evaluate the posterior density $\pi(x|\boldsymbol{y})$ using numerical integration as in eq. (21). For each point θ_k and each point in the set of points of **x** we compute the full conditional distribution $\pi(x|\theta,\boldsymbol{y})$. These full conditionals are shown in Figure 12 (a). Then we weigh this distributions with the corresponding posteriors of θ. In Figure 12(b) the weighted posterior distributions are shown $(\tilde{\pi}(x_i|\boldsymbol{\theta_k},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta_k}|\boldsymbol{y})\Delta_k)$ and above them is the solid line of the marginal posterior density $\pi(x|\boldsymbol{y})$ which is obtained by vertically summing up the weighted posteriors for each point of x.

4.5 Modelling

The modelling approaches, attempted in this thesis, are stemming from the book Blangiardo and Cameletti (2015). In Section 4.5.1 the general model setup is described. In Section 4.5.2 we present the criteria for model selection. The Sections 4.5.3 and 4.5.4 formulate two model specifications for the time trend, followed by the results presented in 4.5.5.
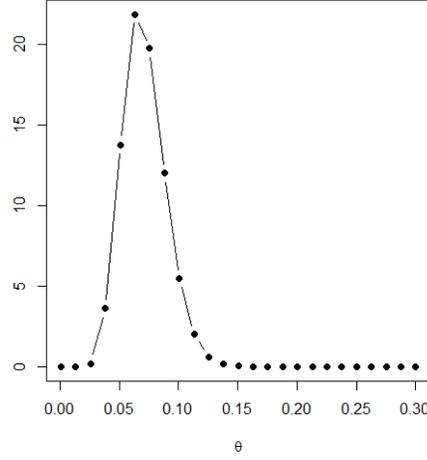
Figure 11: The evaluation of the posterior distribution of θ by finding points with the main bulk of probability masses and then computing an interpolant.

4.5.1 General approach

We assume that the data is Poisson distributed with the average number of cases λ_{it} for each area i and time point t, so that

$$y_{it} \sim \text{Poisson}(\lambda_{it}), \tag{26}$$
$$\lambda_{it} = E_{it}\rho_{it}, \quad \log(\rho_{it}) = \eta_{it}, \tag{27}$$
$$\eta_{it} = b_0 + u_i + v_i + \text{Temporal}_t. \tag{28}$$

The mean value λ_{it} is equal to the expected number of cases in region i at time t, E_{it}, multiplied with a relative risk ρ_{it} for this area; see (27).The expected number of cases is modelled as an offset. The offset is unknown, so we need to calculate it by using the reference rate

$$r = \frac{\sum_{it} y_{it}}{\sum_i n_i},$$
$$E_{it} = \frac{r * n_i}{T},$$

where n_i is the total population in district i, T is the total number of time points, which is 537 days (from the 1st of January 2020 till 21st of June 2021).
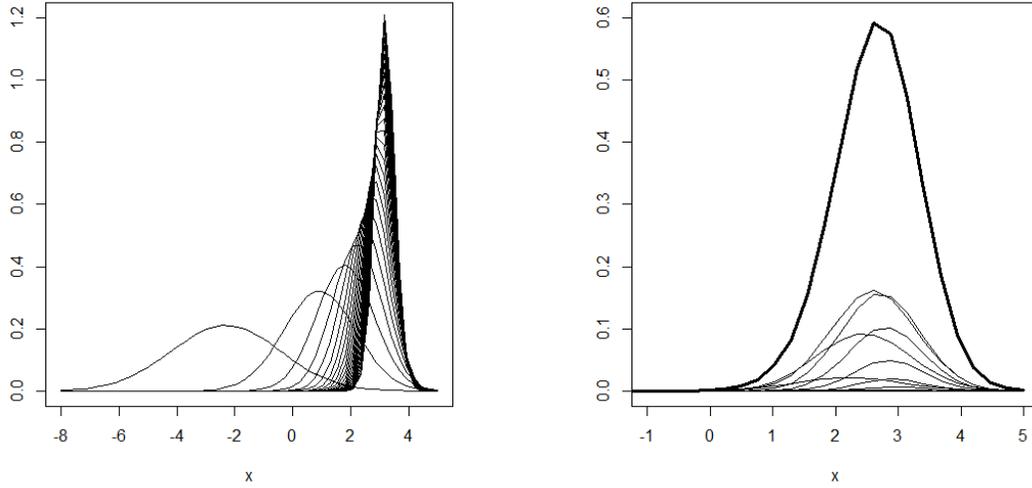
(a) Full conditional distributions for x, given y and hyperparameter θ for each point θ_k .

(b) Weighted joint posterior distributions of x, the solid line is the posterior distribution of x.

Figure 12: Computing the posterior distribution for x by numerical integration.

The rate ρ_{it}, which is the relative risk, is log-linked to the linear predictor η_{it}, the additive structure of which allows to include the fixed effects, random effects, and different models for time trends. If the rate or relative risk ρ_{it} is larger than one, then the risk in area i at time t is larger than the average in the region. Basically, in disease mapping analysis, the areas with an increased relative risk is of main interest for identification.

The predictor (28) includes the overall intercept b_0, the spatially structured area-specific effects u_i, the unstructured area-specific effects v_i and a temporal trend. The structured area-specific effect has a conditional intrinsic autoregressive specification (Besag, 1974)

$$u_i|\boldsymbol{u}_{-i} \sim \text{Normal}\left(\mu_i + \frac{1}{N_i}\sum_{j=1}^{n} a_{ij}(u_j - \mu_j), \frac{\sigma_u^2}{N_i}\right), \qquad (29)$$

where a_{ij} is an indicator function for the neighborhood for i: if j is a neighbour of i, $a_{ij} = 1$, and $a_{ij} = 0$, if they are not. Each unit has a number of neighbours N_i. If the effects are assumed to be zero mean, then the expectation for the effect u_i is an averaged value over the neighboring regions' effects. It is a natural as-

sumption about the smoothness of the area-specific structured effect: we suppose that neighboring areas exhibit a similar picture in terms of the number of cases. From the variance specification in equation (29) one sees that the more neighbors an item i has, the smaller its corresponding variance is. So we expect to be more certain about the structured area-specific effect for the inland sub-regions than for the isolated regions. Combined with the unstructured effect v_i, it is called a Besag-York-Mollié model (Besag et al., 1991). The unstructured area-specific effect v_i, or noise, is assumed to follow a zero mean Normal distribution

$$v_i \sim \text{Normal}\left(0, \sigma_v^2\right),$$

and it accounts for overdispersion due to some heterogeneity across the regions. A priori, we don't have any information on the possible variation of the effects in our models, so we will assign weakly informative priors to our precision parameters of spatial effects (this is default in INLA)

$$\log(\tau_u) \sim \text{logGamma}(1, 0.0005),$$
$$\log(\tau_v) \sim \text{logGamma}(1, 0.0005),$$

where $\tau_u = 1/\sigma_u^2$, and $\tau_v = 1/\sigma_v^2$ are precisions of the effects. The distribution logGamma on the log precision is equivalent to a Gamma distribution on the precision. Each fixed effect has a non-informative normal distribution

$$\beta_i \sim \text{Normal}(0, 0.001),$$

which is also default in INLA.

4.5.2 Model selection criteria

The INLA package allows us to compute the leave-one-out cross-validation indices for model selection without the refitting the model each time. If we split the data into two sets $\boldsymbol{y} = (y_i, \boldsymbol{y}_{-i})$, the model will be fit with the set $\boldsymbol{y}_{-i}$, and one observation y_i is left for validation. Let $y_{i,-i}^{rep}$ denote a new predicted value for y_i based on data $\boldsymbol{y}_{-i}$. The following model selection criteria are used:

The conditional predictive ordinate for leave one out cross-validation

$$\text{CPO}_i = p(y_{i,-i}^{rep}|\boldsymbol{y}_{-i}) = \int p(y_{i,-i}^{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y}_{-i})d\boldsymbol{\theta}$$

The probability integral transform (PIT) for observation i

$$\text{PIT}_i = p(y_{i,-i}^{rep} \leq y_i|\boldsymbol{y}_{-i}).$$

We already used PIT as an informal calibration test. Now we will use PIT for the goodness-of-fit check. The CPO values are the probabilities of the materialized values y_i under the given model. The sum of logarithms of the computed CPO values can also be used as a criterion for the model fit. Small CPO values indicate that these observations are unlikely under the given model. To obtain PIT and CPO we simply need to add `control.compute=list(cpo=TRUE)` in the INLA command. The values of P_{x-1} were computed as the PIT value minus the CPO value. The code for computing and plotting PIT histograms was taken from Andrea Riebler (2014)

The deviance of a model is defined as:

$$D(\boldsymbol{\theta}) = -2\log(p(\boldsymbol{y}|\boldsymbol{\theta})).$$

The deviance information criteria, DIC, introduced by Spiegelhalter et al. (2002), is a measure of model deviance combined with a term quantifying the complexity of the model:

$$DIC = \bar{D} + p_D,$$

where p_D is the effective number of parameters:

$$p_D = E_{\boldsymbol{\theta}|\boldsymbol{y}}(D(\boldsymbol{\theta})) - D(E_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta})) = \bar{D} - D(\bar{\boldsymbol{\theta}})$$

We also want to compare the model results with the endemic-epidemic framework by computing the MSE. In the Bayesian framework we will need to generate posterior samples with the function `inla.posterior.sample`, compute MSE for each of them and take the mean of MSEs.

4.5.3 Parametric trend

The parametric trend model divides the temporal effect into the global linear time trend β and differential area-specific trends δ_i. The linear predictor is

$$\eta_{it} = b_0 + u_i + v_i + (\beta + \delta_i)t. \tag{30}$$

Here, β is assigned a non-informative Normal prior with zero mean and precision of 0.001, as for other fixed effect covariates. The specifications for the differential

trends are

$$\delta_i \sim \text{Normal}\left(0, \frac{1}{\tau_\delta}\right),$$

$$\sum_i \delta_i = 0,$$

The differential time trend allows us to identify the time trend for each area separately: if δ_i is negative, then the time trend is less steep than the main trend, and if δ_i is positive, the area-specific trend is more steep. This is specified in INLA by putting sum-to-zero constraints, as well as for the area-specific spatial effect u_i. The DIC of this model is equal to 764682.14, and the CPO = -382569.2. We add the following explanatory covariates : Older, Young, Income, Employed, Foreigners, Overcrowded, M2 (number of square meters per person), Care workers and Education. These covariates were found important for predicting the number of cases in Söderberg et al. (2022). The DIC and CPO were only slightly improved : DIC = 764657.21 and CPO = -382556.8. The PIT histogram values for the two parametric trend models are shown on the left on the first and the second rows of Figure 13

4.5.4 Nonparametric trend

The temporal trend is modelled as a combination of a random walk γ_t and an unstructured temporal effect ϕ_t, with a linear predictor of the form

$$\eta_{it} = b_0 + u_i + v_i + \gamma_t + \phi_t. \tag{31}$$

The random walk can be of order 1 or 2:

$$\gamma_t|\gamma_{t-1} \sim \text{Normal}(\gamma_{t-1}, \sigma^2),$$
$$\gamma_t|\gamma_{t-1}, \gamma_{t-2} \sim \text{Normal}(2\gamma_{t-1} + \gamma_{t-2}, \sigma^2).$$

The prior for the unstructured temporal effect is set to a Normal distribution,

$$\phi_t \sim \text{Normal}(0, \frac{1}{\tau_\phi}).$$

The default priors for the precisions of the time effects are logGamma with parameters 0 and 0,00005. For the model with random walk of order 1 and no covariates we have DIC = 648152.62 and CPO = -324209. The model with added

covariates has DIC = 648119.55 and CPO = -324186.7. For the similar model with random walk of order 2 and no covariates DIC = 648154.41 and CPO = -324207.9, whereas with the added covariates we have DIC = 648110.77 and CPO = -324182.4. Clearly, the random walk models (formulation (31)) are supported better by the data than the parametric trend models (formulation (30)). Also, the PIT histograms of the random walk models, both with and without covariates, have distributions which are closer to uniform than the models with a parametric trend.
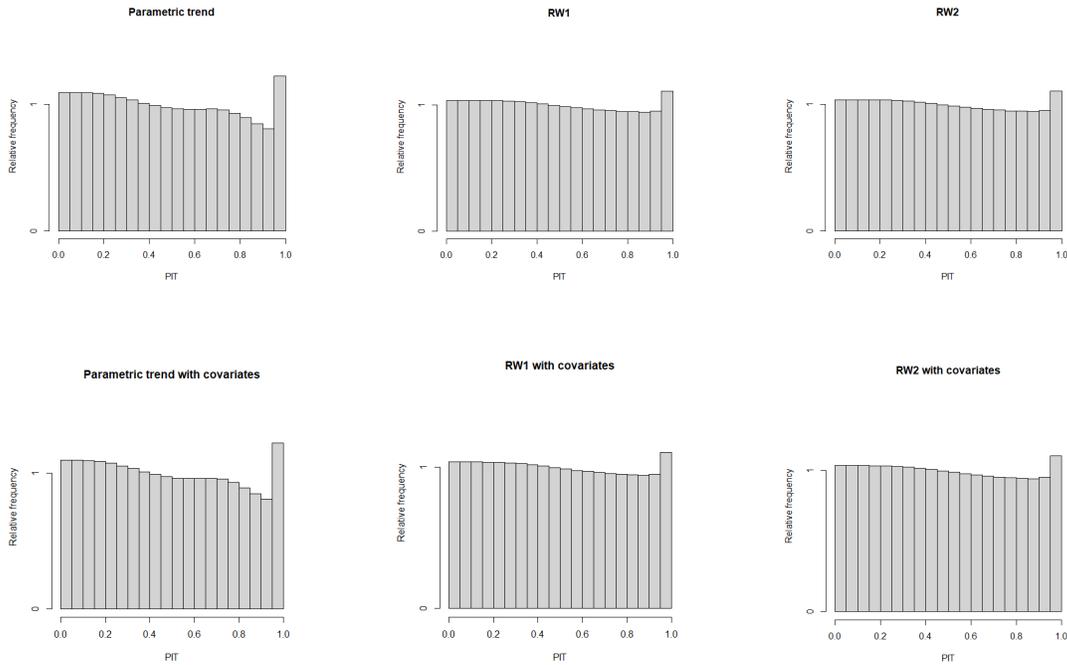
Figure 13: PIT histograms of competing models.

4.5.5 Results

The DIC and CPO criteria are summarized in the table below for the competing models.

Both DIC and CPO point at **rw2 with covs** as best model, although the indices are not far from the **rw1 with covs** model. The covariates do improve both models' fits, but not significantly. If we simulate 10 samples from the posterior distributions and compute the MSE for each them and then take the mean of MSEs, we will get $M\bar{S}E = 0.7906$ for model with **rw1 with covs** and $M\bar{S}E = 0.7918$ for

model	DIC	CPO
param.trend	764682.14	-382569.2
param trend with covs	764657.21	-382556.8
rw1	648152.62	-324209
rw1 with covs	648119.55	-324186.7
rw2	648154.41	-324207.9
rw2 with covs	648110.77	-324182.4

rw2 with covs model. The R code for random walk models is specified as:

```
formula <- observations ~ 1 + log(Older) + log(Young)
 + log(Ink) + log(Empl) + log(Utl) + log(Trang)
 + log(M2) + log(Care.w) + log(Edu)
 + f(area, model = "bym",
 graph = "graph.txt", scale.model = TRUE)
+ f(times, model = "rw1") + f(times.1, model = "iid")
```

Since the model selection criteria for the models **rw1** and **rw2** are very close in values, we would like to compare the posterior credible intervals for the fixed parameters of both models.

covariate	mean	0.025quant	0.975quant	covariate	mean	0.025quant	0.975quant
Intercept	-6.4182	-7.4058	-5.4304	Intercept	-6.4149	-7.4034	-5.4261
log Older	-0.0119	-0.0462	0.0223	log Older	-0.0120	-0.0462	0.0223
log Young	-0.0262	-0.0631	0.0106	log Young	-0.0262	-0.0631	0.0107
log Income	0.5103	0.4194	0.6013	log Income	0.5100	0.4190	0.6009
log Employed	-0.1118	-0.1798	-0.0437	log Employed	-0.1117	-0.1798	-0.0436
log Foreigners	0.1057	0.0747	0.1367	log Foreigners	0.1056	0.0746	0.1366
log Overcrowded	0.0090	-0.0029	0.0209	log Overcrowded	0.0091	-0.0029	0.0210
log M2	-0.2913	-0.4358	-0.1468	log M2	-0.2909	-0.4355	-0.1464
log Care workers	0.0077	-0.0104	0.0258	log Care workers	0.0078	-0.0103	0.0259
log Education	0.0777	0.0338	0.1217	log Education	0.0777	0.0337	0.1217

Table 3: Posterior summary on fixed effects, left: **rw1**, right : **rw2**

In Table 3 we observe that the difference between the intervals is only in the 3rd or 4th number after the comma, which we can consider as negligible for the general interpretation. The intervals that do not contain zero are for those of the intercept and covariates: **Income**, **Employed**, **Foreigners**, **M2**, **Education**. The positive correlation between the number of cases and high education level, as well as between number of cases and the income was a bit surprising. Possibly, this can be explained by the association between the number of cases of viral diseases and metropolitan areas in general: people with higher education level and larger
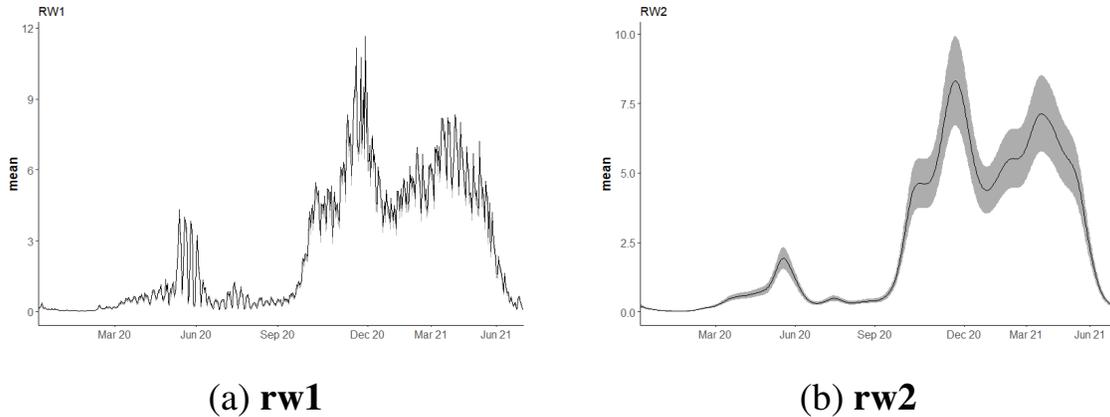
(a) **rw1**

(b) **rw2**

Figure 14: Posterior mean for temporal trend with its credible intervals.

income tend to live in big cities with a higher population density. The negative association of the outcome and the square meters per living person was expected: the less is the living space, the more cases we should expect on average. The number of foreigners, many of whom have to live in overcrowded conditions, also shows a positive association with the number of cases:

The posterior mean with the credible intervals for both models are plotted in Figure 14, where the estimate of **rw2** is much more smooth. Comparing with Figure 1 with the time series of daily counts, one notices that the random walk time model captures very well the temporal dynamics of the cases.

We are interested in the relative risk of the cases for each area, so we map the posterior mean estimates $u_i + v_i$ in Figure 15. The posterior means are exponentiated, so that we see the multiplicative effect of the DeSO region on the number of cases. If the values are larger than 1, then the area shows an excessive relative risk, compared to the average across the whole region. We observe that Gothenburg, its suburbs, as well as North-West and the sub-region bordering to the region Jönköping in the East, exhibit on average higher relative risk of COVID-19 cases. Also, we see that the DeSOs, which denote the dense centre of the local municipality, are those with an excessive relative risk of 1.2 and more (brown colors in Figure 15). The zoomed in map of Gothenburg and its exponentiated posterior mean for spatial effects is shown in Figure 17.

We can also map the probabilities the extensive risks: Figure 16 for the whole region and Figure 18 for Gothenburg. Here, the most densely populated DeSOs,
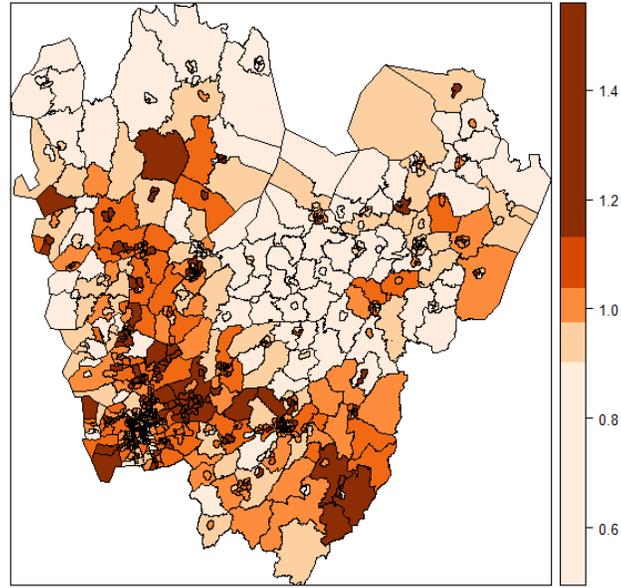
Figure 15: The posterior mean estimates for the area-specific effects $u_i + v_i$. Areas with excessive relative risks have a multiplicative effect value larger than 1.
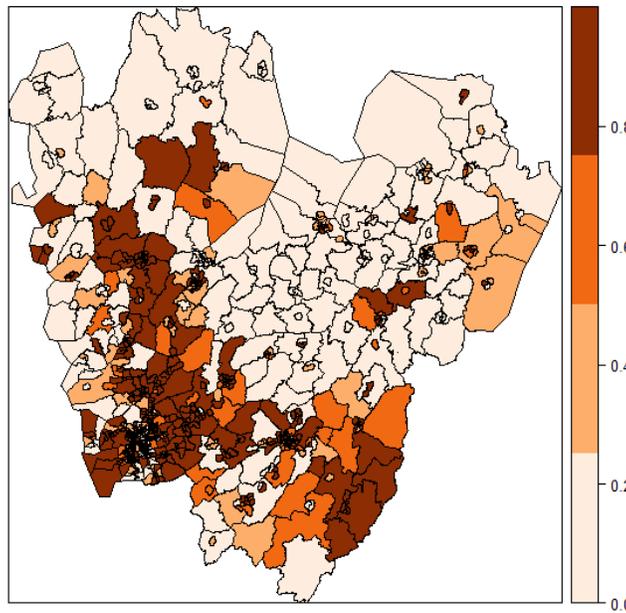

Figure 16: The posterior probabilities that the relative risk is excessive.

Gothenburg and its neighbourhoods, Trollhättan, Borås exceed the relative risk with probability 0.75 and higher. One notes, that although the west of Gothenburg doesn't have an increased relative risk, still the probability to show the higher risk is estimated to be 0.8 and higher. (Figures 17 and 18).
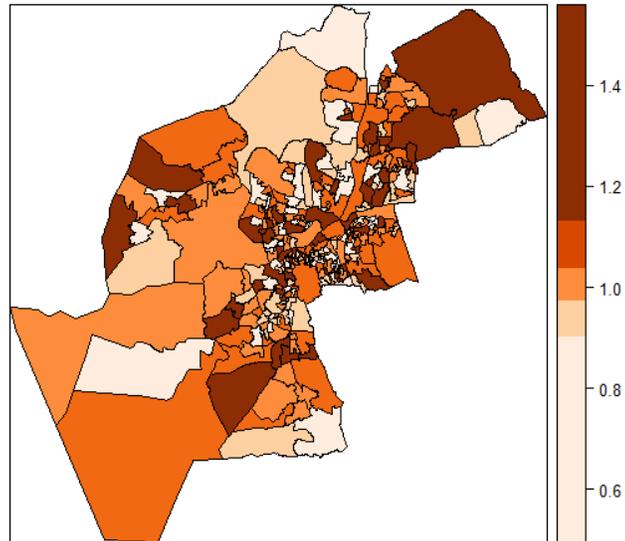
Figure 17: The posterior means for the area-specific effects $u_i + v_i$ in Gothenburg.
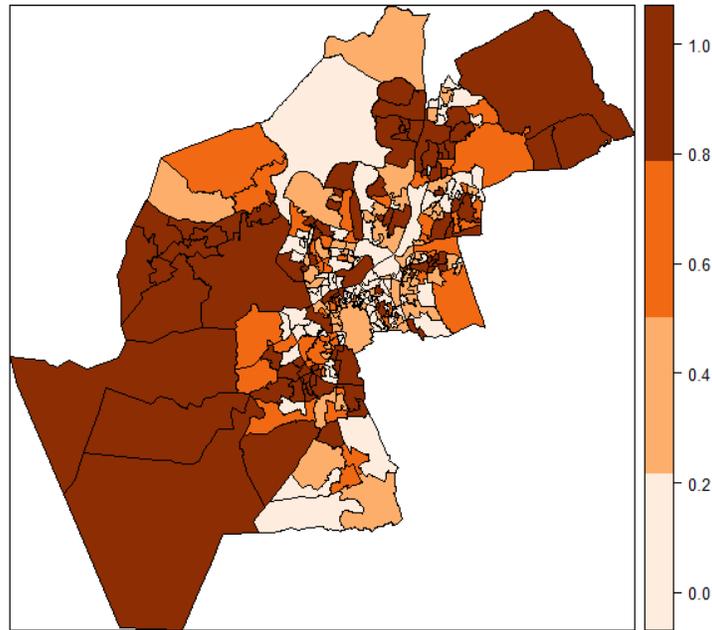
Figure 18: The posterior probabilities that the relative risk is excessive in Gothenburg.

Additionally, we could have mapped the posterior relative risk estimates for each day, but then it would have been 537 figures. Instead, we map the results for the first calendar day of each month to visualize the dynamics of relative risk evolution, so that we get 19 plots in Figure 19. The last figure is for the last day of the study period. The relative risk began to drastically increase in November 2020, being the highest around the period of December 2020, January, March and April 2021, subsided in May and returned to the estimate of 1 in summer 2021.

In the Appendix, we plot the posterior marginal distributions for the fixed parameters in Figure 30, for precisions for hyperparameters in Figure 31, and the variances (the transformed precisions) in Figure 32.
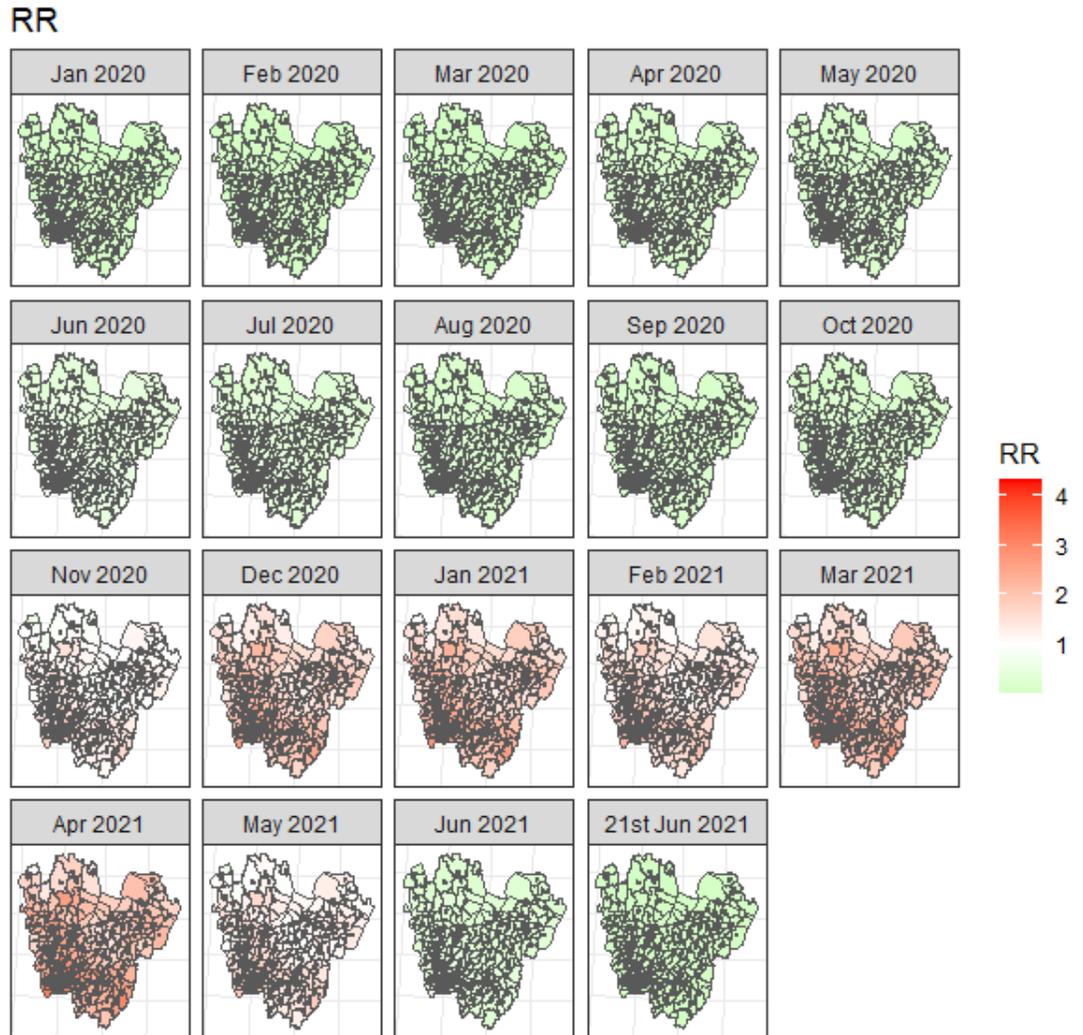
Figure 19: Posterior relative risk estimates of the 1st calendar day of each month.

5 Discussion

5.1 Comparison of two approaches

To compare the goodness of fit of both approaches we use the mean squared errors (MSE). For the endemic-epidemic approach the MSEs are

model	MSE
seasonal + ri	0.4347
rich + ri	0.4346
greedy + ri	0.4347

For the INLA Bayesian approach it is given by

model	mean of MSE
rw1 with covs	0.7906
rw2 with covs	0.7918

Here, we simulated 10 samples from the posterior distribution using the command `inla.posterior.sample`, and then computed the average MSE across these samples. The Endemic-Epidemic model approach performed better in terms of goodness of fit.

For comparison, we present in the table below the mean estimates with the credible bounds and confidence intervals of the significant covariates, found in each of the models: **rw1 with covs**, **greedy + ri**, **rich + ri**. All three models found a positive association between Income and the number of cases, as well as Education, although Education in the autoregressive compartment of the **rich+ri** model shows a negative association. Each model has a positive estimate for the covariate Foreigners: around 0.10 on the logarithmic scale, which means that a 1% increase in the number of foreigners would be associated with a 0.1% increase on average in the number of cases. A larger coefficient for Foreigners found by the **greedy + ri** model in the endemic compartment. The covariate M2 has a positive estimated parameter in the autoregressive components, which is surprising, but the same mean negative estimate of -0.29 in the random walk models and in the **rich + ri** model in the spatio-temporal component. We saw, that the average estimation of the neighbourhood component's contribution to the fitted means was around 81.4% (Section 3.8), so the covariates in the **ne** compartment play major role in the disease spread.

In general, we saw that the inclusion of socio-demographic covariates didn't improve the fit significantly in the both approaches. The flexibility of the model specifications allowed to achieve very good fits even without the given covariates. The possible explanation to that is the nature of COVID-19, a highly viral disease, the transmission of which is not driven by socio-demography, but other factors. More likely, seasonal variations like temperature, humidity play a more important role in the virus spread.

	mean	2.5%	97.5%
rw1 with covs			
log Income	0.5103	0.4194	0.6013
log Employed	-0.1118	-0.1798	-0.0437
log Foreigners	0.1057	0.0747	0.1367
log M2	-0.2913	-0.4358	-0.1468
log Education	0.0777	0.0338	0.1217
greedy + ri			
ar.log M2	1.1890	0.8915	1.48658
ne.log Foreigners	0.1329	0.0949	0.1709
ne.log Income	0.3127	0.2206	0.4047
end.log Foreigners	0.3168	0.2558	0.3778
end.log Education	0.3135	0.2437	0.3832
rich + ri			
ar.log M2	1.2367	0.9422	1.5312
ar.log Education	-0.3046	-0.3993	-0.2099
ne.log M2	-0.2902	-0.4986	-0.0819
ne.log Income	0.3857	0.2608	0.5105
ne.log Foreigners	0.1024	0.0586	0.1462
end.log Education	0.3366	0.2677	0.4056

Table 4: Comparison of the fixed effects, presenting in the selected models

5.2 Limitations and further work

The large dataset comprising approximately half a million points made the computational time for the INLA model around 40 minutes, whereas the `surveillance` package took around 3 hours for the model with random effects in each component. It is possible to introduce the interaction terms in INLA models: there are four types of interaction specifications between space and time, which is implied to be natural inclusion in the spatio-temporal disease modelling. Besides, the interaction terms between the chosen covariates would be needed to further investigate the impact of the driving factors on the disease spread. Also, one can include linear combinations between the intercept and the random effects in order to obtain the posterior distributions of the area-specific intercepts. Unfortunately, again due to large computational intensity, these model extensions were not feasible.

The other limitation were the border effects in the estimation of the neighbourhood influence. Since we took only Västra Götaland region, we did not use the neighboring regions (Swedish: län) in our analysis. As a result of this, the estimate of component contribution might be slightly misleading for the border municipal-

ities. Future work could be to perform a similar analysis on whole Sweden.

The inclusion of all covariates in the model led to misleading and confounding results. It is quite hard to interpret and quantify the impact of each covariate on the number of cases. A possible solution would be to include only one covariate at the time and compare which model makes better prediction.

6 Appendix

Theory of GMRF

This section is based on the material from the Rue and Held (2005) book.

The two random variables x and y are independent, iff $\pi(x, y) = \pi(x)\pi(y)$, and we write it as $x \perp y$. The two variables are conditionally independent iff $\pi(x, y|z) = \pi(x|z)\pi(y|z)$, which we write as $x \perp y|z$.

Theorem 1.
$$x \perp y|z \quad \Longleftrightarrow \quad \pi(x, y, z) = f(x, z)f(y, z) \tag{32}$$
for some functions f and g, and for all z with $\pi(z) > 0$.

Theorem 2. *Let $\boldsymbol{x}$ has a Normal distribution with mean μ and a precision matrix $\boldsymbol{Q} > 0$. Then for for $i \neq j$,*

$$x_i \perp x_j | \boldsymbol{x}_{-ij} \quad \Longleftrightarrow \quad Q_{ij} = 0.$$

Proof. We partition $\boldsymbol{x}$ as $(x_i, x_j, \boldsymbol{x}_{-ij})$. Fix $i \neq j$ and assume $\boldsymbol{\mu} = \boldsymbol{0}$. The density of **x** has the form $\pi(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})\right)$, (eq.15). Then

$$\pi(x_i, x_j, \boldsymbol{x}_{-ij}) \propto \exp\left(-\frac{1}{2}\sum_{k,l} x_k Q_{kl} x_l\right)$$

$$\propto \exp\left(-\frac{1}{2}\underbrace{x_i x_j (Q_{ij} + Q_{ji})}_{\text{term 1}} - \frac{1}{2}\underbrace{\sum_{\{k,l\} \neq \{i,j\}} x_k Q_{kl} x_l}_{\text{term 2}}\right).$$

We see that term 2 doesn't involve $x_j x_j$. Term 1 involves $x_i x_j$ iff $Q_{ij} \neq 0$. Comparing with (32) in Theorem 1 we get

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

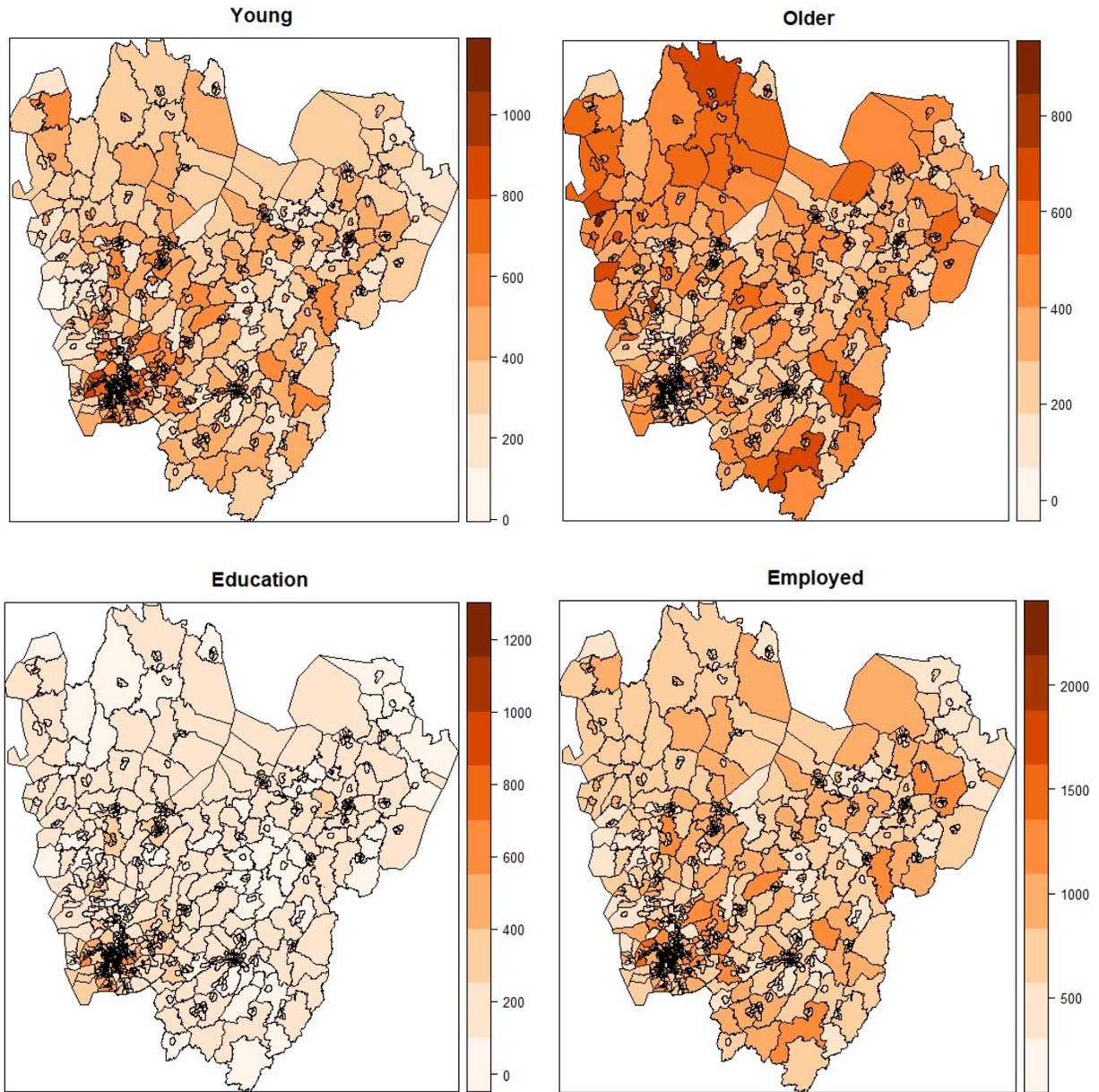for some functions f and g, iff $Q_{ij} \neq 0$. $\qquad\square$

Figure 20: Socio-demographic covariates: Number of people of age 19 and younger, Number of people of age 65 and older, Number of people having 3 or more years of post-secondary education, Number of gainfully employed workers.
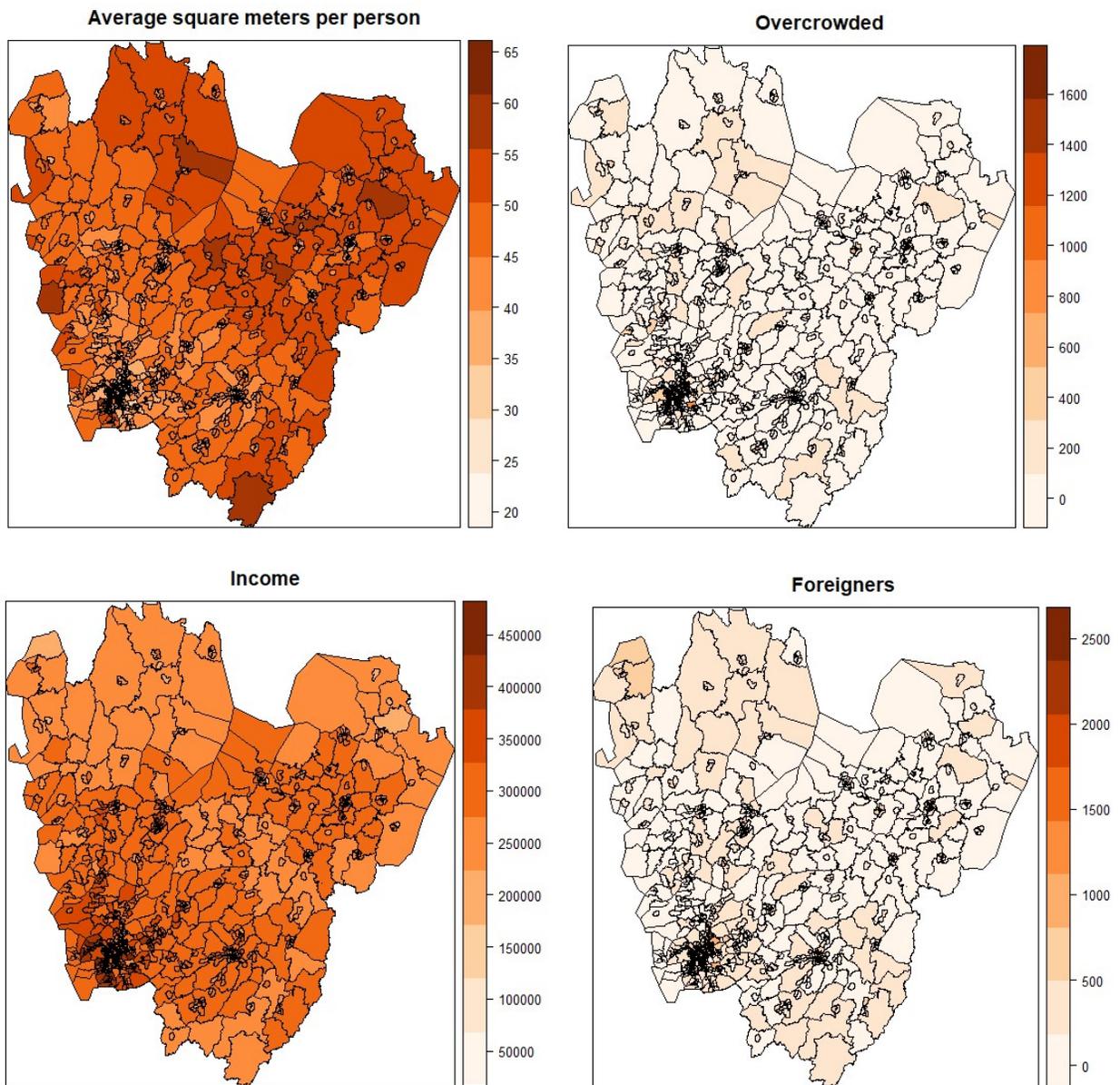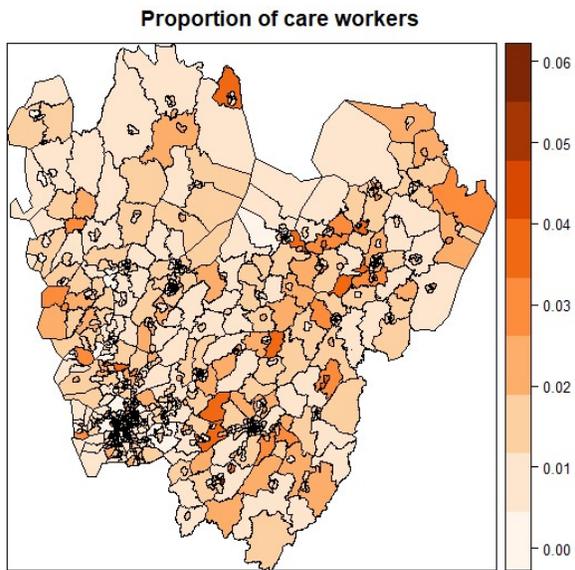
Figure 21: Socio-demographic covariates: Average square meters per person, Number of people living in overcrowded conditions, Average income, Number of people of non-Swedish background.

Proportion of care workers

Endemic

Figure 22: Socio-demographic covariates: The proportion of hospital staff.

	Estimate	2.5 %	97.5 %
ar.sin(2 * pi * t/365)	0.24	0.17	0.31
ar.cos(2 * pi * t/365)	0.32	0.24	0.39
ar.sin(4 * pi * t/365)	-0.23	-0.29	-0.16
ar.cos(4 * pi * t/365)	-0.17	-0.23	-0.11
ar.log(M2)	1.19	0.89	1.49
ar.log(Young)	0.60	0.47	0.73
ar.ri(iid)	-10.55	-12.07	-9.03
ne.log(pop)	0.76	0.68	0.84
ne.sin(2 * pi * t/365)	0.32	0.27	0.36
ne.cos(2 * pi * t/365)	0.32	0.29	0.34
ne.sin(4 * pi * t/365)	-0.14	-0.16	-0.12
ne.cos(4 * pi * t/365)	0.00	-0.02	0.03
ne.log(Trang)	0.05	0.03	0.06
ne.log(Utl)	0.13	0.09	0.17
ne.log(Older)	0.07	0.03	0.12
ne.log(Ink)	0.31	0.22	0.40
ne.ri(iid)	0.48	-1.02	1.99
end.t	-0.00	-0.00	-0.00
end.sin(2 * pi * t/365)	-1.99	-2.18	-1.80
end.cos(2 * pi * t/365)	-0.90	-1.01	-0.79
end.sin(4 * pi * t/365)	-1.56	-1.67	-1.45
end.cos(4 * pi * t/365)	0.54	0.44	0.63
end.log(Edu)	0.31	0.24	0.38
end.log(Utl)	0.32	0.26	0.38
end.ri(iid)	2.20	1.80	2.60
neweights.d	1.80	1.76	1.83
overdisp	0.28	0.27	0.29

Table 5: Estimates with confidence intervals, model **greedy + ri**

	V1	2.5 %	97.5 %
ar.sin(2 * pi * t/365)	0.24	0.17	0.31
ar.cos(2 * pi * t/365)	0.31	0.23	0.38
ar.sin(4 * pi * t/365)	-0.24	-0.31	-0.17
ar.cos(4 * pi * t/365)	-0.17	-0.23	-0.11
ar.log(M2)	1.24	0.94	1.53
ar.log(Young)	0.81	0.65	0.96
ar.log(Edu)	-0.30	-0.40	-0.21
ar.ri(iid)	-10.37	-11.95	-8.79
ne.log(pop)	0.74	0.63	0.84
ne.sin(2 * pi * t/365)	0.31	0.26	0.35
ne.cos(2 * pi * t/365)	0.32	0.29	0.34
ne.sin(4 * pi * t/365)	-0.14	-0.16	-0.12
ne.cos(4 * pi * t/365)	-0.00	-0.03	0.02
ne.log(Trang)	0.04	0.02	0.06
ne.log(M2)	-0.29	-0.50	-0.08
ne.log(Older)	0.13	0.06	0.19
ne.log(Ink)	0.39	0.26	0.51
ne.log(Edu)	-0.00	-0.05	0.04
ne.log(Utl)	0.10	0.06	0.15
ne.ri(iid)	0.14	-1.37	1.66
end.t	-0.00	-0.00	-0.00
end.sin(2 * pi * t/365)	-1.95	-2.14	-1.76
end.cos(2 * pi * t/365)	-0.93	-1.03	-0.82
end.sin(4 * pi * t/365)	-1.54	-1.65	-1.43
end.cos(4 * pi * t/365)	0.51	0.42	0.60
end.log(Kids)	-0.30	-0.40	-0.21
end.log(Edu)	0.34	0.27	0.41

Table 6: Estimates with confidence intervals, model **rich + ri**
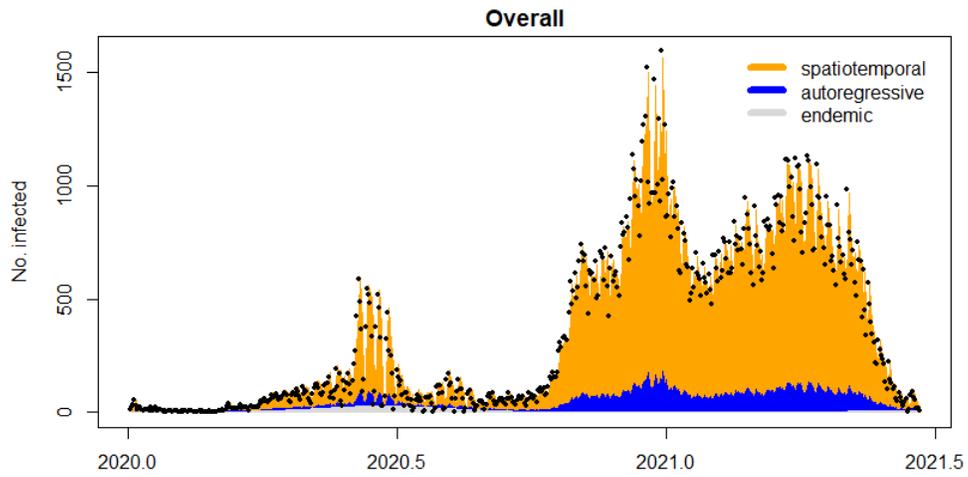
Figure 23: The fitted mean components for the **basic** model, aggregated over all regions.
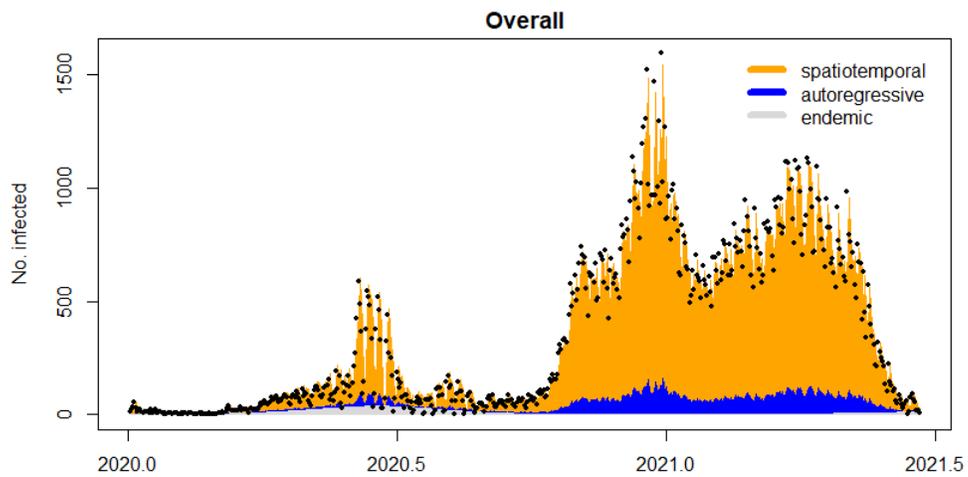

Figure 24: The fitted mean components for the **basic + ri** model, aggregated over all regions.
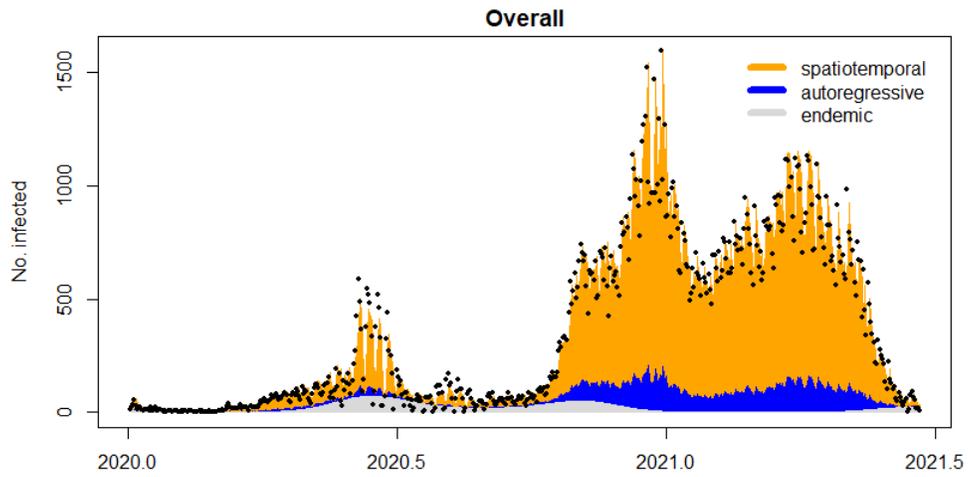
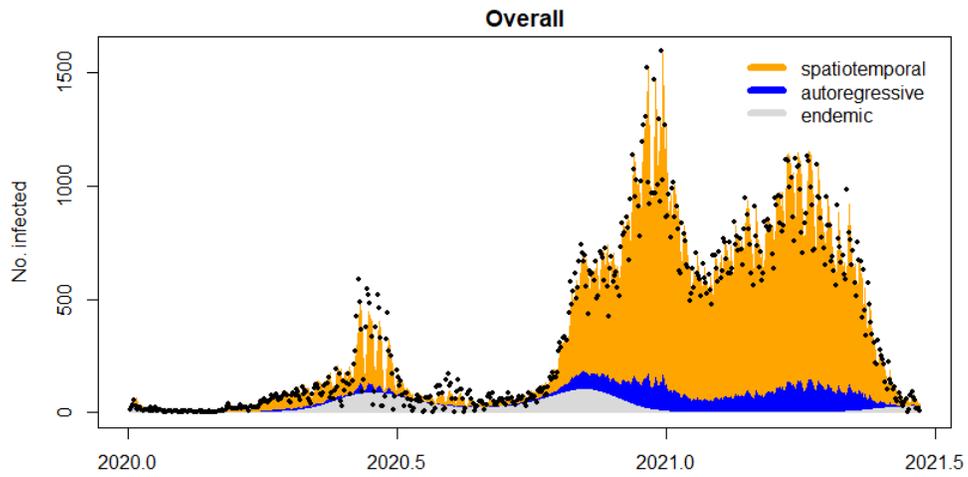Figure 25: The fitted mean components for the **seasonal** model, aggregated over all regions.

Figure 26: The fitted mean components for the **seasonal + ri** model, aggregated over all regions.
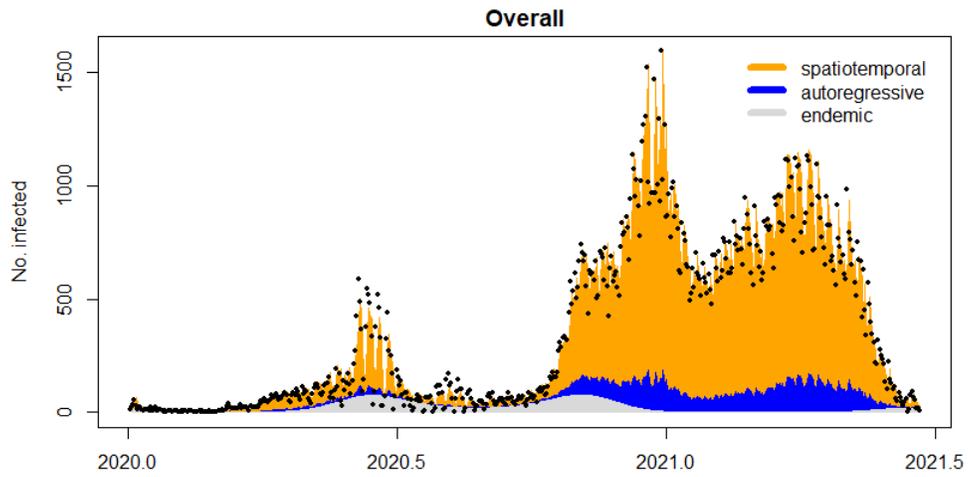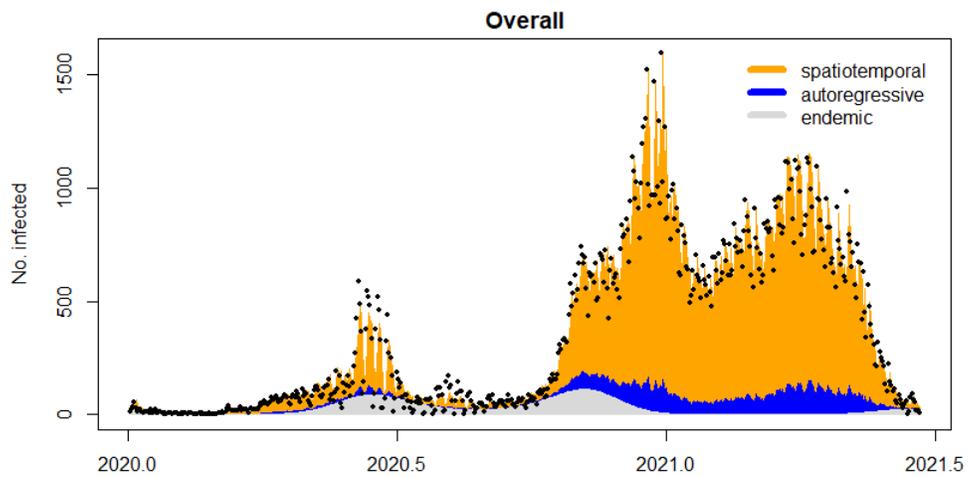
Figure 27: The fitted mean components for the **rich** model, aggregated over all regions.

Figure 28: The fitted mean components for the **rich + ri** model, aggregated over all regions.
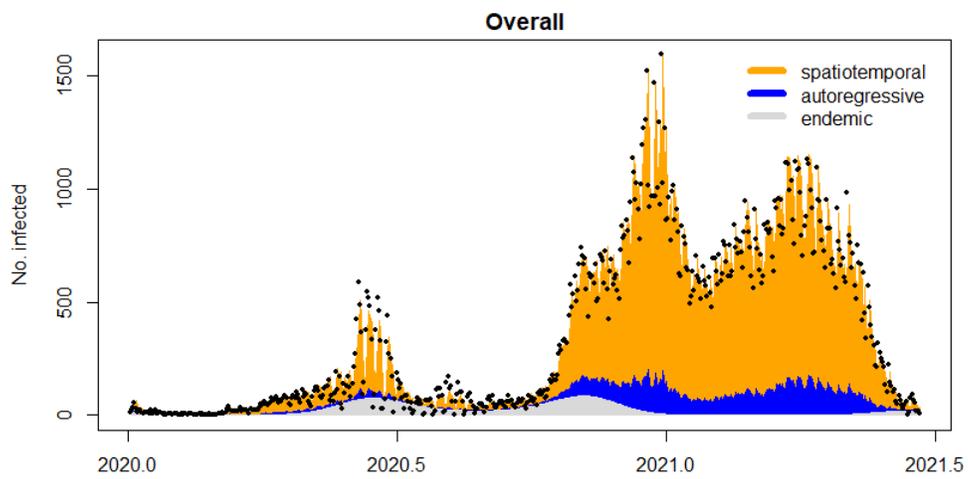
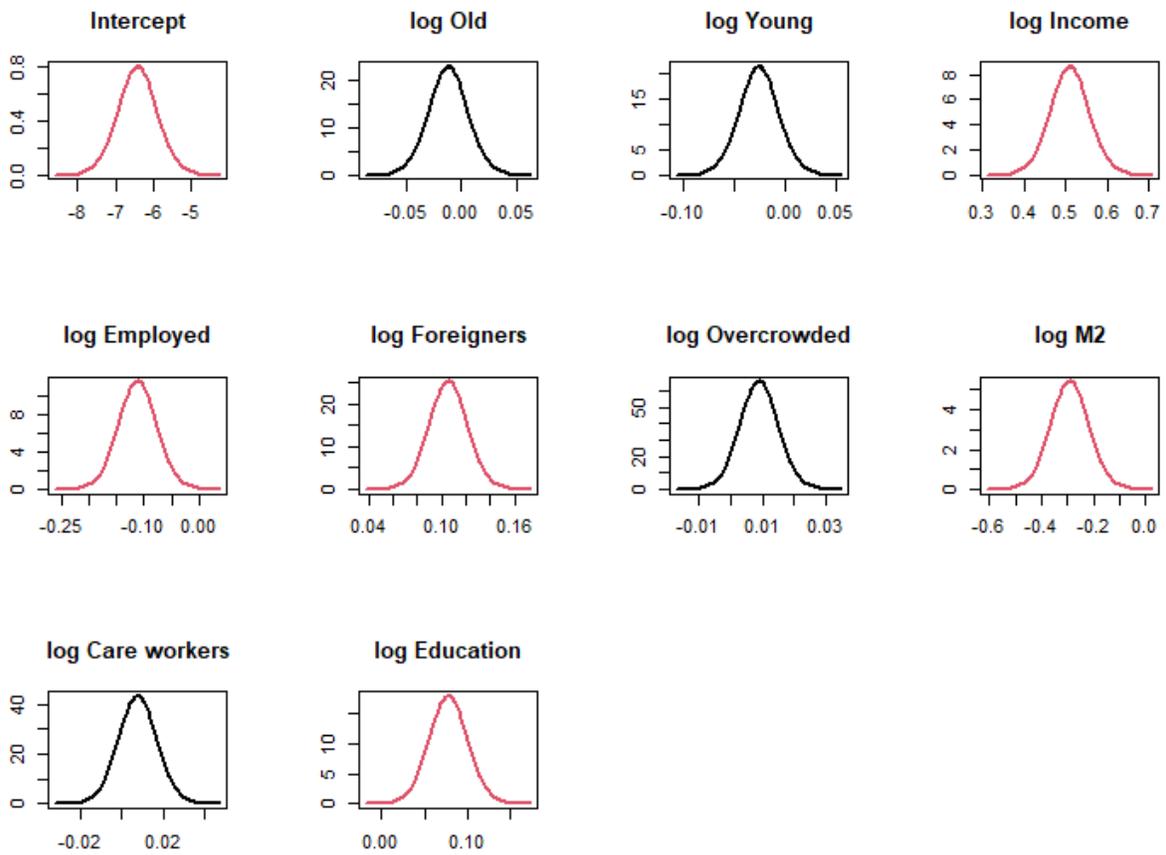Figure 29: The fitted mean components for the **greedy** model, aggregated over all regions.

Figure 30: Posterior marginal distributions for the fixed parameters of the **rw1 with covs** model. The distributions with credible intervals not containing 0 are shown in red.
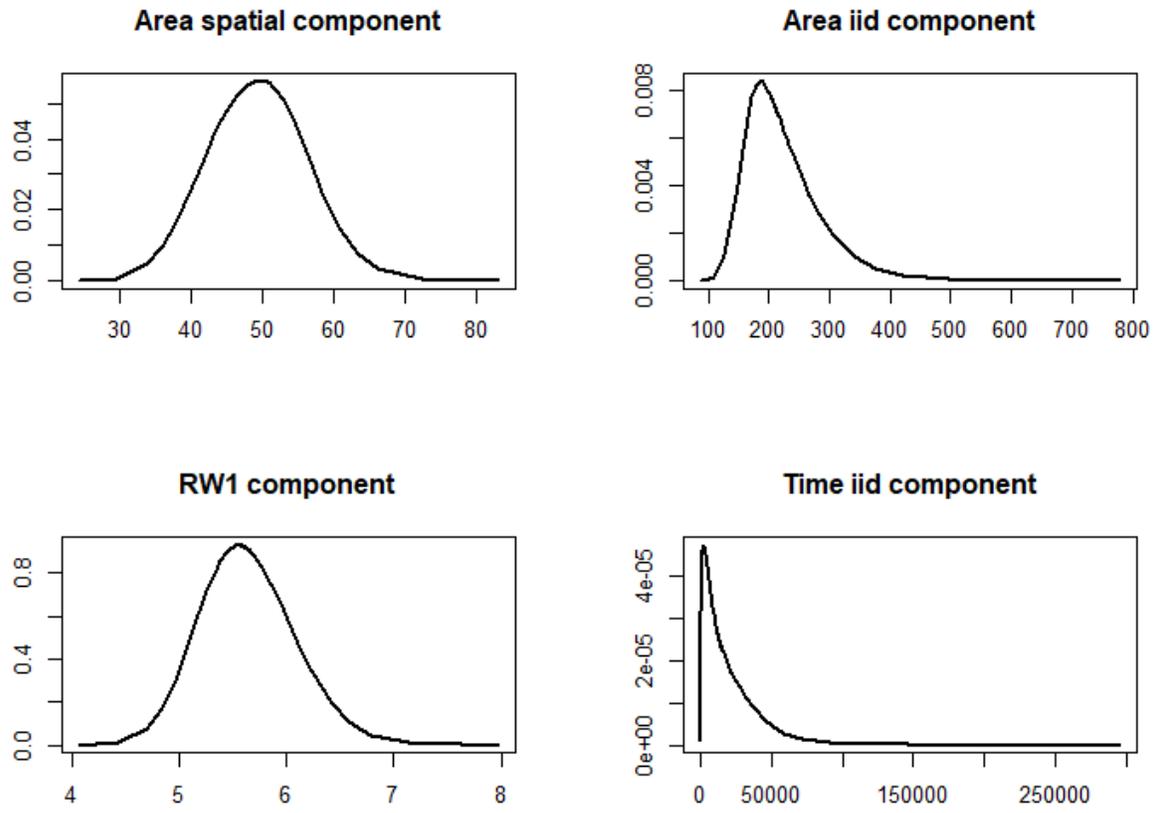
Figure 31: Posterior marginal distributions of precisions of the **rw1 with covs** model. The first row shows the precisions for the area components. The second row shows the precisions for the time model,
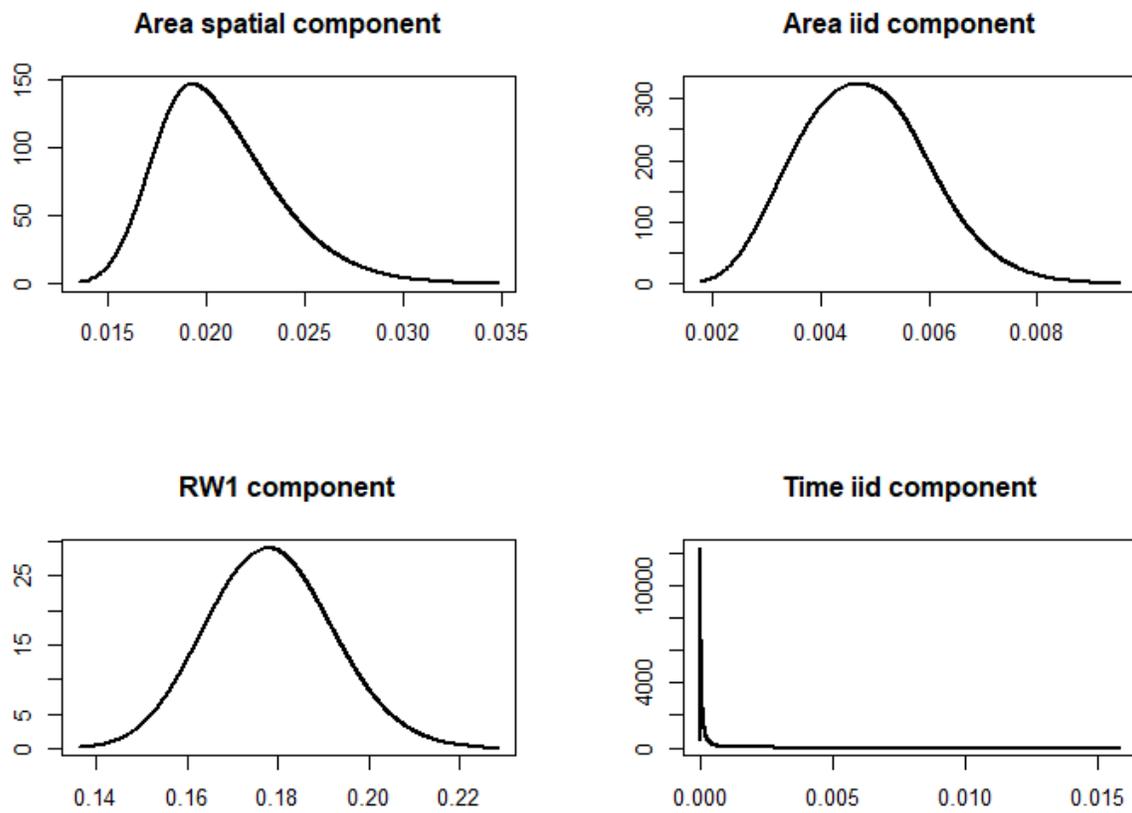
Figure 32: Posterior marginal distributions for the variances (transfromed precisions) of the **rw1 with covs** model.

References

L. Held, M. Höhle, and M. Hofmann, "A statistical framework for the analysis of multivariate infectious disease surveillance counts," *Statistical modelling*, vol. 5, no. 3, pp. 187–199, 2005.

H. Rue and S. Martino, "Approximate bayesian inference for hierarchical gaussian markov random field models," *Journal of statistical planning and inference*, vol. 137, no. 10, pp. 3177–3192, 2007.

M. Söderberg, O. Cronie, M. Adiels, and A. Rosengren, "The influence of overcrowding and socioeconomy on the spatio-temporal spread of covid-19–a swedish register study," 2022.

J. A. Andersen, B. Rowland, S. M. Ratcliff, H. C. Felix, and P. A. McElfish, "Relationship between sociodemographic factors, perceived covid-19 risk, and engagement with health protective behaviors," *Southern Medical Journal*, vol. 115, no. 5, p. 340, 2022.

A. López-Gay, J. Spijker, H. V. Cole, A. G. Marques, M. Triguero-Mas, I. Anguelovski, M. Marí-Dell'Olmo, J. A. Módenes, D. Álamo-Junquera, F. López-Gallego *et al.*, "Sociodemographic determinants of intraurban variations in covid-19 incidence: the case of barcelona," *J Epidemiol Community Health*, vol. 76, no. 1, pp. 1–7, 2022.

R. Markovič, M. Šterk, M. Marhl, M. Perc, and M. Gosak, "Socio-demographic and health factors drive the epidemic progression and should guide vaccination strategies for best covid-19 containment," *Results in Physics*, vol. 26, p. 104433, 2021.

M. V. Ibañez, M. Martínez-Garcia, and A. Simó, "A review of spatiotemporal models for count data in r packages. a case study of covid-19 data," *Mathematics*, vol. 9, no. 13, p. 1538, 2021.

M. Paul, L. Held, and A. M. Toschke, "Multivariate modelling of infectious disease surveillance data," *Statistics in medicine*, vol. 27, no. 29, pp. 6250–6267, 2008.

M. Paul and L. Held, "Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts," *Statistics in medicine*, vol. 30, no. 10, pp. 1118–1136, 2011.

L. Held and M. Paul, "Modeling seasonality in space-time infectious disease surveillance data," *Biometrical Journal*, vol. 54, no. 6, pp. 824–843, 2012.

S. Meyer and L. Held, "Power-law models for infectious disease spread," 2014.

H. Abbey, "An examination of the reed-frost theory of epidemics," *Human biology*, vol. 24, no. 3, p. 201, 1952.

J. Wakefield, T. Q. Dong, and V. N. Minin, "Spatio-temporal analysis of surveillance data," in *Handbook of Infectious Disease Data Analysis*. Chapman and Hall/CRC, 2019, pp. 455–475.

S. Meyer, L. Held, and M. Höhle, "Spatio-temporal analysis of epidemic phenomena using the r package surveillance," *arXiv preprint arXiv:1411.0416*, 2014.

D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

A. P. Dawid, "Present position and potential developments: Some personal views statistical theory the prequential approach," *Journal of the Royal Statistical Society: Series A (General)*, vol. 147, no. 2, pp. 278–290, 1984.

C. Czado, T. Gneiting, and L. Held, "Predictive model assessment for count data," *Biometrics*, vol. 65, no. 4, pp. 1254–1261, 2009.

A. P. Dawid and P. Sebastiani, "Coherent dispersion criteria for optimal experimental design," *Annals of Statistics*, pp. 65–81, 1999.

S. Martino and H. Rue, "Implementing approximate bayesian inference using integrated nested laplace approximation: A manual for the inla program," *Department of Mathematical Sciences, NTNU, Norway*, 2009.

H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *Journal of*

the royal statistical society: Series b (statistical methodology), vol. 71, no. 2, pp. 319–392, 2009.

T. G. Martins, D. Simpson, F. Lindgren, and H. Rue, "Bayesian computing with inla: new features," *Computational Statistics & Data Analysis*, vol. 67, pp. 68–83, 2013.

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

C. Chiuchiolo, J. van Niekerk, and H. Rue, "An extended simplified laplace strategy for approximate bayesian inference of latent gaussian models using r-inla," *arXiv preprint arXiv:2203.14304*, 2022.

H. Rue and L. Held, *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

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

J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 192–225, 1974.

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

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the royal statistical society: Series b (statistical methodology)*, vol. 64, no. 4, pp. 583–639, 2002.

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY