Bayesian Spatial-Temporal Statistics for Epidemic Risk Estimation and Modelling

A thesis presented in partial fulfilment of the

requirements for the degree of

Master of Science

In

Statistics

at Massey University,

Manawatū, New Zealand.

You Zhou

2024

Abstract

This thesis focuses on employing Bayesian methods for spatiotemporal modeling of various types of epidemiological or health-related spatial-temporal data. Spatial data types include point pattern data, point reference (geostatistical) data, and area-level (lattice) data. Different types of spatial data have different spatial resolution and virtual assumptions. Therefore, the spatial and spatiotemporal modeling approaches for different data types also differ.

The thesis introduces spatial modeling methods for three types of epidemiological data, including discrete spatial model, linear geostatistical model, generalized linear geostatistical model, Poisson point process, and log-gaussian Cox process. Additionally, we elaborate on extending spatial modeling to spatiotemporal modeling for various data types. This extension is relatively intuitive to implement due to the flexibility of Bayesian methods and hierarchical Bayesian models. We used two epidemiological datasets as examples; one is Campylobacteriosis cases in the Manawatu region of New Zealand during the period Mar 2005-Sep,2016, and another dataset is from the SARS-CoV-2 wastewater surveillance program launched in 2022 in Aotearoa, New Zealand, which is used to monitor and track potential cases of COVID-19. When modeling COVID-19 using wastewater epidemiology data in New Zealand, we employ the INLA-SPDE method. This approach, a Bayesian analysis method for spatial data on intricate grids, represents a new frontier in Bayesian disease mapping techniques and has practical applications.

Regarding Bayesian computational methods, we introduced the traditional Monte Carlo sampling method, the Markov chain Monte Carlo (MCMC) method, and the integrated nested Laplace approximation (INLA), an approximate Bayesian inference method. Compared to commonly used MCMC methods, INLA has a significant advantage in providing precise parameter estimates in less time and is user-friendly through the R-INLA package. R-INLA conducted all the Bayesian-related computations for this thesis. Finally, we discussed the transformation of data types related to the flexible use of spatial epidemiological data and modeling methods. It's crucial to model flexibly according to the problem we aim to solve in epidemiological modeling rather than adhering to specific data formats corresponding to particular models.

Acknowledgements

I want to thank all the people who helped me in the study, especially my supervisor, Associate Professor Jonathan Marshall, for his advice, patience, and help. Thank Massey University's subscription to New Zealand eScience Infrastructure (NeSi) allow for high-performance computing (HPC) clusters. I am also grateful to the Institute of Environmental Science and Research (ESR) for providing data support.

Contents:

Chapter 7: Discussion

Chapter 8: Conclusion

Chapter 1
Introduction

1.1 History of Spatial Statistics and Why Spatial Statistics in Epidemiology Modeling

Moran (1950) first proposed spatial autocorrelation measures to study the phenomena of multiple dimensional spatial random distributions in 1950. Next, in 1951, Krige (1951) introduced the fundamental concepts of spatial statistics. He discussed that people in the mining field needed to know the difference between samples and blocks and how to use regression to enhance prediction. Later refined by Matheron (1963,1967), leading to the formulation of geostatistics and kriging techniques in 1963 and 1967. They are important components of spatial interpolation and spatial statistics. In 1973, “Spatial Autocorrelation” from Cliff & Ord (1973) was published and it first elaborated in detail on the problems of identifying spatial correlation, assessing spatial autocorrelation strength, and the solutions to these issues. This work elevated geographical research to a higher level. Afterward, this book was expanded and reissued as "Spatial Processes: Models & Applications" in 1981 (Cliff & Ord, 1981). In 1993, Cressie (1993) published "Statistics for Spatial Data," a book that systematically introduces statistical analysis methods for spatial data. It contributed to the enhancement of the theoretical framework of spatial statistics. Naturally, when we seek to study spatial patterns at different time points, spatiotemporal statistical models are then expanded. Cressie also published "Statistics for Spatio-Temporal Data" in 2015. This book provides a detailed discussion of statistical analysis methods for spatiotemporal data, including spatiotemporal interpolation, spatial and temporal correlations, and more (Cressie et al, 2015). In recent years, the significant development in spatial analysis systems (such as GIS, Geoda, Winbugs, SaTScan, etc) and the growing availability of modeling packages (such as R-INLA) have also made great progress in health/disease-related

spatial statistics (Beale et al., 2008).

Why Spatial Statistics in Epidemiology Modeling? It is estimated by the Pan American Health Organization, PAHO (1996) that 80% of epidemiological data related to infectious diseases possess spatial attributes and factors such as the distribution of hosts or vectors, temperature, humidity, rainfall, and the incidence of illness in humans or animals are all linked to geographic location. Given this context, we must consider how to utilize spatial statistical methods in handling public health disease surveillance data, fully acknowledging its spatial characteristics to ensure more authentic and reliable analysis results. According to Cressie et al. (2022), spatial statistical methods can be classified according to spatial data types and specificities into methods based on point data analysis, area data analysis, continuous data analysis, spatial regression analysis, spatial interpolation, and so on. Their increasingly widespread application in public health involves primarily sampling disease surveillance points, estimating and interpolating incidence rates, analyses of disease clustering, exploring risk factors, determining spatial characteristics of diseases, and forecasting spatiotemporal patterns of disease (Elliott et al. 2016). This is spatial epidemiology, and it is also known by various other names, including ecological epidemiology, small-area health study, and disease mapping, among others. But regardless of its name, Lawson et al. (2016) think it should exhibit two key characteristics. First, the distribution in space or geographical location holds significant importance as it enables spatial statistics to function as part of geographic information systems. Second, geo-referenced disease data influences our analysis of disease indicators, requiring consideration of the spatial distribution of diseases. In addition to this, naturally, if we can access spatial data across different time points, we must consider its spatiotemporal distribution.

1.2 History of Bayesian method in Spatial Epidemiology Modeling and Why the Bayesian method in spatial epidemiology modeling.

Spatial epidemiology, or disease mapping, has a long history. Some of the earliest spatial epidemiological studies were purely geographical, for example, Snow's map of cholera cases in London in 1854 (Koch et al., 2009). Later, Haviland's (1871, 1888) mapping observed disease and

cancer incidence and mortality rates for the counties of England, displaying the geographic distribution of disease and illustrating potential associations between disease and the environment (Haviland,1871,1888). Since health-related data is typically released based on health administrative regions, disease mapping originally comes from a topic in spatial statistics for lattice data, which calculates the crude rates of a rare disease for (small) geographic areas (MacNab,2004). Earlier statistical research of disease mapping focused on developing empirical Bayes (EB) models and related estimation procedures for smoothing maps of disease rates or relative risks (Lawson, 2018). Since the 1990s, there has been a rising trend in Bayesian statistics and Bayesian hierarchical approaches. Specifically, employing hierarchical Bayesian models to construct complex models, typically conducting posterior estimation and inference of unknown quantities and parameters through Markov Chain Monte Carlo (MCMC) simulations. A shift from EB to a fully Bayesian (FB) approach to disease mapping occurred at that time. It is an analytically natural transition because a fully Bayesian approach facilitates "data-driven" posterior estimation and inference of the prior parameters (MacNab,2022). Bayesian scanning methods to detect disease areas with the most significant excess risk are also down by Kulldorff & Nagarwalla (1995) and Openshaw et al. (1987), respectively. Besag et al. (1991) introduced Bayesian models fitting random effects through MCMC for risk estimation and modeling. Since then, Bayesian statistics has gradually gained widespread and effective application in most areas of disease mapping.

Why the Bayesian method in spatial epidemiology modeling? Various sources of uncertainty exist in spatiotemporal modeling, requiring a flexible structure to connect them. The hierarchical model employs conditional probability to link three distinct structures together, inferring process parameters and hyperparameters upon obtaining observed data. Simultaneously, it addresses various sources of uncertainty (Lawson et al., 2016). In a Bayesian hierarchical inferential framework, the prior distribution—like different classes of CAR structures used for spatial information sharing- could be considered a latent process model for underlying risks leading to the observed disease counts (Blangiardo and Cameletti, 2015). Then, statistical inference utilizes the Bayesian method to compute posterior distributions for all unknown quantities and parameters. In a Bayesian hierarchical framework, exploring thoughtful prior options aids in building models, estimating Bayesian posteriors, and studying the inference of relative risk and other parameters of interest. This method

helps understand uncertainties around risks and unknown parameters in a systematic and principled manner (Lawson, 2018), which is a “data-driven” process (MacNab,2022). Another reason is that, compared to frequentist methods, Bayesian approaches are able to tolerate a certain degree of noise and data incompleteness when modelling (Carlin and Louis, 2010; McElreath, 2016). This is particularly important for health-related/disease data, where missing values are quite common due to various reasons (such as privacy concerns, access restrictions, the need for long-term tracking and observation, etc.) (Pearce and Merletti, 2006). The Bayesian approach can prevent introducing extra bias that may arise from interpolating to obtain complete data.

The Bayesian method and hierarchical models have established their unique advantages in disease mapping and spatial epidemiology. In recent decades, spatial biostatistics or spatial epidemiological statistics have experienced significant advancements. Bayesian disease mapping is a prominent topic within spatial statistics; with ongoing explorations in public policy, public health, and spatial epidemiology, its scope and complexity continue to expand. Now, proposals of the INLA-SPDE approach (Lindgren et al.,2011) in Bayesian spatial data analysis across high-resolution grids represent a new innovative frontier in Bayesian disease mapping and its practical applications (MacNab,2022).

1.3 Thesis Objectives:

This thesis has two main goals to achieve:

1: For the two types of common spatial disease data, namely areal (or lattice) data, and point referenced (or geostatistical) data, we will introduce their corresponding Bayesian spatial-temporal epidemiology risk estimation modeling method, specifically involving the creation of different Gaussian random fields in a hierarchical model. From Besag et al. (1991), different classes of CAR structures are used in the Bayesian statistical model as spatial information-sharing structures would be introduced for areal data study. Besides that, we will also explore the ecological regression model, which is a kind of spatial/spatial-temporal model with covariates. When identifiable factors pose risks,

the research aims to assess how these factors influence the disease incidence or mortality rate (Blangiardo et al., 2015). We introduce the linear geostatistical model (LGM) and the generalized linear geostatistical model (GLGM) for point-referenced data. At the same time, we explore an epidemiological modeling approach based on wastewater as a case study, where virus concentrations in wastewater typically are sampled from various catchment areas. However, these catchment areas are often not adjacent to each other. Instead of using the polygons of catchment areas, employing geometric centroids or population-weighted centroids as substitutes is a good approach for large-scale epidemic risk modeling. In this context, we will utilize the INLA-SPDE approach (Lindgren et al., 2011) and apply potential latent Gaussian fields to conduct Bayesian spatial data analysis with high-dimensional information from many points (areas) distributed over high-resolution grids. Additionally, the INLA method can provide a relatively accurate and fast approximation.

2. The time resolution of disease data is another focus of our attention. If we can obtain data for the same target at different time points, then we can study the development trend of the disease, which is crucial for epidemics. Therefore, our second goal focuses on extending existing spatial disease data exploration and analysis methods to spatiotemporal disease data modeling. In modeling both types of data, benefiting from the flexibility of the Bayesian hierarchical framework, we focus on creating spatiotemporal interaction variables, namely, Gaussian random fields with time components, and incorporate them into the hierarchical model, thus achieving the goal of spatiotemporal modeling.

1.4 Thesis Outline：

Chapter 1:

First, we introduce the development of spatial statistics and its inseparable relationship with epidemic statistics modeling. Then, we will discuss research paths related to Bayesian statistics, spatial epidemiology, or disease mapping. In terms of models, hierarchical models remain the mainstream choice. In Bayesian methods, there has been a transition from empirical Bayes to full Bayes. In computational methods, there has been a shift from Markov Chain Monte Carlo (MCMC) to

Integrated Nested Laplace Approximation (INLA).

Chapter 2: Hierarchical model and Computational Bayesian Inference.

We first introduce hierarchical models, including the latent Gaussian model that a hierarchical structure can represent. In terms of econometric methods, we discusse Monte Carlo (MC), Markov Chain Monte Carlo (MCMC), and Integrated Nested Laplace Approximation (INLA). In MCMC, we will discuss the two most commonly used methods, Metropolis-Hastings (MH) and Gibbs sampling. In INLA, we initially introduce the Laplace approximation, and the Gaussian random fields. Integrated Nested Laplace Approximation (INLA) is designed specifically for LGMs and especially focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF) (Rue et al. 2009). Then, we introduced the properties of Gaussian random fields, the SPDE method for Gaussian random fields, and spatial autocorrelation functions. Finally, we discuss Bayesian model comparison method.

Chapter 3: Spatial Discrete Model for Area-level Epidemiological Data

We first introduce the spatial weight matrix used to define spatial relationships. Then, we discusse spatial sharing structures suitable for lattice data and the models that incorporate them, namely the Intrinsic Conditional Autoregressive (ICAR) Model, The Proper CAR (pCAR) Model, The Besag, York, and Mollié (BYM) Model. We also introduce the spatial model that includes area-level covariates, called the ecological model. Finally, we discuss how to establish a spatiotemporal interaction variable, following the approach outlined by Knorr-Held (2000), to extend spatial epidemiological modelling to spatiotemporal modelling.

Chapter 4: Case study of the campylobacteriosis in Manawatu, New Zealand

A case study about campylobacteriosis in Manawatu, New Zealand from March 2005 to 2016 by using spatial and spatial-temporal discrete model.

Chapter 5: Spatial Point Process Models

First, we briefly introduce the point pattern process here. Since it is the data with the highest spatial resolution and usually has high confidentiality, the development of this type of epidemic modeling has been limited to a certain extent. We introduce the Poisson process model, which is a frequentist inference. A Bayesian method, the log-Gaussian Cox process model, used in spatial epidemiology to analyze the spatial distribution of case events, will also be presented. Next, for the point-referenced (geostatistical) model, we introduced the linear geostatistical models (LGM) and generalized linear geostatistical models (GLGM), respectively. In GLGM, we primarily introduced two types of link functions: binomial, Poisson. In terms of methods, we introduced the principle of SPDE methods to establish non-stationary and non-isotropic Gaussian random fields and how to use INLA-SPDE to construct high-resolution grids for spatial modeling. Finally, we also discuss how to extend spatial models to spatiotemporal models by creating spatiotemporal interaction variables in point-referenced data.

Chapter 6: Case Study About the Aotearoa SARS-CoV-2 Wastewater Surveillance Programme

A case study about modeling COVID-19 using wastewater epidemiology data by INLA-SPDE is also attached, and the data is from the SARS-CoV-2 wastewater surveillance program launched in 2022, which is used to monitor and track potential cases of COVID-19 in Aotearoa New Zealand.

Chapter 7: Discussion.

We discussed the transformation of data types related to the flexible use of spatial epidemiological data and modeling methods. It's crucial to model flexibly according to the problem we aim to solve in epidemiological modeling rather than adhering to specific data formats corresponding to models.

Chapter 8: Conclusion

We outline our conclusions here and discuss possible further work.

1.5 Data

In this thesis, we use two epidemic-related datasets, both of which report cases or case-related indicators in two dimensions (time and space). One is the lattice data about Campylobacteriosis cases in Manawatu, New Zealand from Mar 2005 to Sep 2016 and the other is a point-referenced data about the Aotearoa SARS-CoV-2 wastewater surveillance programme to monitor and track potential cases of COVID-19. They will be discussed in detail in Chapter 4 and Chapter 6 respectively.

Chapter 2 Hierarchical Model and Computational Bayesian Inference.

In this chapter, we will briefly introduce Bayesian inference. Bayesian inference is easy to understand in theory, but computing the Bayesian posterior distribution can be challenging, especially when the model has non-conjugate priors. In Bayesian inference, sampling-based methods are the most commonly used, including Monte Carlo sampling (MC) and Markov Chain Monte Carlo (MCMC), among which we also introduce Metropolis-Hastings (MH) and Gibbs algorithms as representatives of MCMC. We also discuss the Integrated Nested Laplace Approximation (INLA), an approximate Bayesian posterior estimation method. Then, we move to the hierarchical model. Before introducing the Gaussian random field (GRF), we first discuss the special form of the hierarchical model, and the latent Gaussian model (LGM). The Bayesian approach of INLA can provide accurate and fast inference for LGM models (Rue et al., 2009). Latent Gaussian field and Gaussian random field are often embedded into certain levels of hierarchical models to handle spatial correlations or other structural features in the data. At the end of this chapter, we introduce the comparison of the Bayesian models. For Bayesian inference, please see Gelman et al. (2013) and Carlin & Louis (2010). For hierarchical models, please see Hoff (2009) and Congdon (2010). For Bayesian computational inference, please see McElreath (2016), and more details about INLA can be found in Rue et al. (2016).

2.1 Bayesian Inference

We will briefly introduce the basic concepts of Bayesian statistical inference and decision-making. We are starting from the Bayesian formula in probability theory. Assuming there is a complete event group in the sample space D, consisting of a finite number of events $B_1,\ B_2\ \ldots\ B_n$, satisfying the conditions of being pairwise disjoint and their union forming the complete sample space. So, if A is

an event in the sample space, then the law of total probability is given as follows:

$$P(A) = \sum_{i=1}^{n} P(AB_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \tag{2.1}$$

Under the conditions of the law of total probability, where $B_{i,\ i=1\ldots n}$ is any event in the sample space with a probability greater than 0. The methods for calculating conditional probability are as follows:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)} = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)} \tag{2.2}$$

This formula is called Bayes' formula, which is formally just a conditional probability formula. We know that in statistical models, the task is to make inferences about the population using samples, and this information is referred to as sampling information. However, Bayes' formula introduces a new type of information known as prior information, which refers to some information about unknown parameters in statistical inference before sampling. The basic concept of Bayesian statistics is to simultaneously utilize sample information and prior information in the inference process. We can obtain a prior distribution when the prior information is sufficiently rich. Any probability distribution on the parameter space can be referred to as a prior distribution. $\pi(\theta) = P(\theta = \theta_i)$ is used to represent the probability distribution of a random variable θ happened at time $\{\theta = \theta_i\}$. The prior distribution represents the understanding of the possible values of the parameter before getting samplings S. Once the sampling information is obtained, people's understanding of the parameter θ changes, and the adjusted result is the new understanding of θ, known as the posterior distribution $\pi(\theta|s) = P(\theta = \theta_i|S)$. The posterior distribution can also be considered as the adjustment of the prior distribution using prior information.

Given $\boldsymbol{S} = \boldsymbol{s}$, that is, after obtaining sample $\boldsymbol{s}$, the conditional distribution of $\boldsymbol{\theta}$, which is the posterior distribution, has a probability density function:

$$\pi(\theta \mid s) = \frac{\pi(s|\theta)\pi(\theta)}{\int \pi(s|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{\pi(\mathbf{s} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{s})} \tag{2.3}$$

And the posterior is sometimes expressed as:

$$\pi(\boldsymbol{\theta} \mid \mathbf{s}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{s} \mid \boldsymbol{\theta}) \tag{2.4}$$

2.2 Bayesian Computing

There are different types of prior distributions in Bayesian statistics, but an important concept is conjugate prior distributions. If the posterior distribution belongs to the same family as the prior distribution, then the prior and posterior distributions are termed conjugate distributions. If the prior distribution is not a conjugate prior distribution, then the posterior distribution is often no longer standard. More details about the prior could be found at Bolstad et al. (2016). Therefore, the calculation of the posterior distribution often lacks an explicit expression and faces various integration computation issues. In this chapter, we will mainly introduce two Bayesian computation methods. First is the Monte Carlo sampling method, followed by the Markov chain Monte Carlo (MCMC) method.

2.2.1 Monte Carlo (MC) sampling

Monte Carlo (MC) sampling is the most common method. As we desire numerical characteristics of the posterior distribution, such as posterior expectation, posterior quantiles, etc., within an integral expression, to estimate these numerical characteristics of the posterior distribution, we can extract enough samples from the population and then estimate the population's numerical features based on the sample's numerical characteristics. Suppose multiple independent observations can be generated from the posterior distribution when the sample size is sufficiently large. In that case, this estimation is referred to as Monte Carlo approximation, and this method is known as Monte Carlo sampling (Gelman et al. 2013). However, this method is seldom directly applicable because, in most cases, the

posterior distribution is not a standard distribution, making it difficult to sample from. Monte Carlo Importance Sampling is one method, but it has some evident shortcomings.

2.2.2 Markov Chain Monte Carlo (MCMC)

The MCMC method uses Markov Chain Monte Carlo integration, with the basic idea being to construct a Markov chain whose stationary distribution is the posterior distribution of the parameter to be estimated (Hastings, 1970). Monte Carlo integration is performed using samples from this Markov chain when it reaches its stationary distribution. The construction of the Markov chain's transition kernel is crucial in employing the MCMC method, and different methods of constructing this kernel led to various MCMC techniques (Gelman et al., 2013). Currently, two commonly used MCMC methods are the Metropolis-Hastings algorithm and Gibbs sampling.

2.2.2.1 Metropolis-Hastings Sampling Method

Metropolis-Hastings (MH) Sampling algorithm is one of the most commonly used MCMC sampling algorithms. Metropolis first proposed it (Metropolis and et al. 1953) and later extended by Hastings (1970). We use f(.) to represent the target distribution and g(.) to represent the proposal distribution First, to generate a Markov chain, starting from an initial value x_0 and specifying a rule for transitioning from the current value x_t to the next value x_{t+1}, thus generating the Markov chain $x_{t,\ t=0,1,\ldots}$. More specifically, given the current value x_t, a candidate point x' is generated from the proposal distribution $g(.|x_t)$. If the candidate point is accepted, it transitions to the next moment, setting $x_{t+1} = x'$. If the candidate point is not accepted, the Markov chain remains in the state x_t, setting $x_{t+1} = x_t$. Whether the candidate point is accepted is referred to as the next value of the chain, determined by the acceptance probability $\partial(x') = \min(1, \mathrm{C})$, where $C = \frac{f(x')g(x_t|x')}{f(x_t)g(x'|x_t)}$.

The selection of the proposal distribution g(.) also has certain conditions. Apart from needing to ensure that the generated Markov chain satisfies regularity conditions (irreducible, aperiodic, and possessing a stationary distribution), it should also include:

1. Easy sampling from the proposal distribution, often a known and commonly used distribution.
2. The acceptance probability of the proposal distribution should be easily computable.
3. The tails of the proposal distribution should be thicker than those of the target distribution.
4. The frequency of rejecting new candidate points should not be too high, otherwise, efficiency will be low.

The process of generating a Markov chain satisfying regular conditions using the Metropolis-Hastings algorithm in a Bayesian framework is as follows:

1. choose a proposal distribution $g(.|x_t)$.
2. Generate the initial value $\boldsymbol{\theta}_0$ from the proposal distribution. Note that x is replaced by the parameter θ in the Bayesian model, while the target function is replaced by the posterior distribution $\pi(\theta|s)$.
3. Repeat the following steps for time t = 1, 2, 3...:

   A: Generate a candidate value from the proposal distribution.

   B: Generate a random number u from a uniform distribution.

   C: Calculate the accept/reject decision for the candidate θ', if

$$u \leqslant \frac{\pi(\theta' \mid s)g(\theta_t \mid \theta')}{\pi(\theta_t \mid s)g(\theta' \mid \theta_t)}$$

Then accept θ' and set $\theta_{t+1} = \theta'$, otherwise set $\theta_{t+1} = \theta_t$;

4. t=t+1; return to step 1.

The selection of the proposal distribution g provides flexibility to the MH Sampling method. Depending on the different structures of the proposal distribution, the Metropolis-Hastings Sampling method can also lead to independent sampling methods, Metropolis random walk sampling methods, and component-wise Metropolis-Hastings Sampling methods, among others. Each of them has its advantages and disadvantages.

2.2.2.2 Gibbs Sampling

Gibbs sampling was first proposed and applied to Gibbs lattice distributions by Geman et al. (1984), hence its name. It's another specific case of MCMC sampling, and its significance lies in transforming sampling from a multivariate distribution into sampling from a univariate target distribution.

In the Bayesian framework, let the parameters $\theta = (\theta_1, \dots, \theta_k)$, and $\theta_{-j} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots \theta_k)$. The target distribution f(.) is represented as the posterior probability π(θ | x). The algorithm for generating a Markov chain using Gibbs sampling is as follows:

1. Choose an initial value for the parameters.
2. Repeat the following steps for t = 1, 2...J

a) Set $\theta_1, \theta_2, \dots, \theta_k = \theta^{(t-1)}$

b) For each component $j = 1 \dots k$: generate ccandidate point θ_j^{t} from $\pi(\theta_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{x})$ and update $\theta_j = \theta_j^{(t-1)}$

$$\begin{aligned} \theta_1^{(t)} \mid \boldsymbol{x} &\sim \pi\left(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_k^{(t-1)}, \boldsymbol{x}\right), \\ \theta_2^{(t)} \mid \boldsymbol{x} &\sim \pi\left(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_k^{(t-1)}, \boldsymbol{x}\right), \\ &\vdots \\ \theta_k^{(t)} \mid \boldsymbol{x} &\sim \pi\left(\theta_k \mid \theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{k-1}^{(t)}, \boldsymbol{x}\right). \end{aligned} \tag{2.5}$$

c) Set $\boldsymbol{\theta}^{(t)} = \left(\theta_1^{(t)}, \cdots, \theta_k^{(t)}\right)$

d) t=t+1; return to step 1.

Gibbs samplers have their unique advantages, but if the parameters θ_j, θ_{-j} are highly correlated, the Gibbs sampler will move slowly around the parameter space.

2.3 INLA

The INLA method, introduced by Rue et al. (2009), stands out as a deterministic approach for Bayesian inference, distinct from simulation methods like MC and MCMC introduced above. INLA is especially for latent Gaussian models and focused on approximating Bayesian regression model posterior estimates, this approach can offer substantial computational benefits compared to the commonly employed MCMC methods by practitioners. Using the INLA methodology in R-INLA (Thiago et al., 2012) makes the inference process for latent Gaussian models a valuable practical resource, offering a user-friendly interface for defining stochastic partial differential equations (SPDEs) models (Lindgren et al., 2011). We make a brief introduction on the Laplace approximation, and for more about the INLA algorithm and INLA inference, please see Rue et al. (2009).

2.3.1 From Laplace Approximation

Laplace's method purposed by Tierney et al. (1986) is usually employed as an approximation estimation of the posterior distribution. When we want to compute the following the integral of $f(x)$ which is the density function of random variables x, we show it as below:

$$\int f(x)\mathrm{d}x = \exp\int \quad (\log f(x))\,\mathrm{d}x \tag{2.6}$$

We can represent $\log f(x)$ through a Taylor series expansion of the second order evaluated at the $x = x_0$.

$$\log f(x) \approx \log f(x_0) + (x - x_0)\frac{\partial \log f(x)}{\partial x}\bigg|_{x=x_0} + \frac{(x - x_0)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\bigg|_{x=x_0} \tag{2.7}$$

Assume we set the x_0 is equal to the mode of $\text{argmax}_x \log\big(f(x)\big)$, then the first derivative of the equation $\frac{\partial \log f(x)}{\partial x}\Big|_{x=x_0}$=0 and the approximation going to be :

$$\log f(x) \approx \log f(x^*) + \frac{(x-x^*)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*} \tag{2.8}$$

So, the approximation of the integral of the $f(x) = \exp(\log f(x))$ could be rewrite like:

$$\begin{aligned} \int f(x)\mathrm{d}x &\approx \exp \int \left(\log f(x^*) + \frac{(x-x^*)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*}\right)\mathrm{d}x \\ \exp &= (\log f(x^*)) \exp \int \left(\frac{(x-x^*)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*}\right)\mathrm{d}x \end{aligned} \tag{2.9}$$

It looks somewhat similar to the form of a Gaussian distribution, so the integrand could be associated with the pdf of a normal distribution.

We could set:

$$\sigma^{2*} = -\frac{1}{\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*}} \tag{2.10}$$

And then we get

$$\int f(x)\mathrm{d}x \approx \exp \approx (\log f(x^*)) \exp \int \left(-\frac{(x-x^*)^2}{2\sigma^2}\right)\mathrm{d}x \tag{2.11}$$

The integrand above is the kernel of a Gaussian distribution with mean=x^*, and the variance equal to σ^{2*}. Be more specific, the integral evaluated in the interval $(\alpha,\ \beta)$ is approximated by.

$$\int_\alpha^\beta f(x)\mathrm{d}x \approx f(x^*)\sqrt{2\pi\sigma^{2*}}\big(\Phi(\beta) - \Phi(\alpha)\big) \tag{2.12}$$

Where Φ is the cumulative density function of the Gaussian distribution $(x^*,\ \sigma^{2*})$. It's worth noting that Laplace's method could be extended to multiple case. In upcoming sections, latent gaussian and

how it associated with INLA will be introduced.

2.4 Hierarchical Model and Gaussian Random Field

2.4.1 Introduction to Hierarchical Model

In spatial epidemiological modeling, the data we obtain often possess both spatial and temporal attributes. These data need a versatile modeling framework capable of integrating complex structures into both data and model parameters to address different sources of uncertainty (Lawson, 2013). Hierarchical model is a good choice. It is a statistical model described in multiple levels (hierarchical form), utilizing Bayesian methods to estimate parameters of the posterior distribution. We could combine submodels to form a hierarchical model and integrate these submodels with observed data using the Bayesian theorem to explain all existing uncertainties (Carlin & Louis, 2010). The outcome of this integration is the posterior distribution, also known as updated probability, as it acquires additional evidence regarding the prior distribution. The hierarchical model, employing Bayesian statistics to use prior estimates for posterior inference, offers a flexible model with many applications in the social, political, and health sciences (Banerjee et al., 2015).

The hierarchical modelling framework formalizes a statistical model into three components: a data model, a process model, and a parameter model. Each level of the model is designed to establish a flexible structure for handling the uncertainty associated with the data, the process, and the parameters. This three-stage hierarchical Bayes model has a joint density that is the product of three stages. According to Bayesian theory:

$$\begin{aligned}\Pr(\text{ process, parameters } \mid \text{ data }) \propto & \Pr(\text{ data } \mid \text{ process, parameters }) \\ & \times \Pr(\text{ process } \mid \text{ parameters }) \\ & \times \Pr(\text{ parameters })\end{aligned} \tag{2.13}$$

Written by the phase of observation denoted as $\mathbf{y}$ is defined by data conditioned on all latent parameters ϕ and hyperparameters $\boldsymbol{\theta}_1$.

$$\mathbf{y} \sim \pi(\mathbf{y} \mid \boldsymbol{\phi}, \boldsymbol{\theta}_1) \tag{2.14}$$

The phase of latent process denoted as ϕ s defined by latent parameters ϕ given hyperparameters $\boldsymbol{\theta}_2$.

$$\phi \sim \pi(\phi \mid \boldsymbol{\theta}_2) \tag{2.15}$$

and the hyperparameter stage is defined by hyperparameters $\boldsymbol{\theta} = (\theta 1; \theta 2)$.

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \tag{2.16}$$

Then, the joint density of the model is:

$$\pi(\mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{\phi} \mid \boldsymbol{\theta})\pi(\mathbf{y} \mid \boldsymbol{\phi}, \boldsymbol{\theta}) \tag{2.17}$$

With the Bayes theory, the posterior distribution of parameter and hyperparameter is like:

$$\pi(\boldsymbol{\phi}, \boldsymbol{\theta} \mid \mathbf{y}) = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{\phi} \mid \boldsymbol{\theta})\pi(\mathbf{y} \mid \boldsymbol{\phi}, \boldsymbol{\theta})}{\int_{(\Theta,\Phi)} \pi(\boldsymbol{\theta})\pi(\boldsymbol{\phi} \mid \boldsymbol{\theta})\pi(\mathbf{y} \mid \boldsymbol{\phi}, \boldsymbol{\theta}) d(\boldsymbol{\phi}, \boldsymbol{\theta})} \tag{2.18}$$

And the posteriors of the $\boldsymbol{i_{th}}$ latent parameter and hyperparameter are like:

$$\begin{aligned} \boldsymbol{\pi}(\boldsymbol{\phi_i} \mid \mathbf{y}) &= \int \boldsymbol{\pi}(\boldsymbol{\phi_i} \mid \boldsymbol{\theta}, \boldsymbol{y})\boldsymbol{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} \\ \boldsymbol{\pi}(\boldsymbol{\theta_i} \mid \mathbf{y}) &= \int \boldsymbol{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} \end{aligned} \tag{2.19}$$

Here, we need the Bayesian computation methods we introduced earlier because integrals are usually too complex to solve.

2.4.2 Latent Gaussian Model

Latent Gaussian models, a versatile subset within hierarchical models, have gained widespread popularity across statistical realms and diverse application fields (Rue & Martino, 2008). More details about this could be found at Rue & Martino (2008), and Martins et al. (2013). The class of LGM could be represented by a hierarchical structure containing three hierarchies. We go from the bottom to the top. The bottom stage is formed by the conditionally independent likelihood function:

$$\pi(y \mid \phi, \theta) = \prod_{i=1}^{n} \pi(y_i \mid \eta_i(\phi), \theta) \tag{2.20}$$

Where $y = (y_1, y_2, \dots y_n)$ is the observation which is the response vector.

$\phi = (\phi_1, \phi_2, \dots \phi_n)$ is the latent field parameter, which is the vector of all unknown latent parameters of link function $\eta_i(\phi)$ which serves as a linear predictor.

$\theta = (\theta_1, \theta_2, \dots \theta_n)$ is the hyperparameter vector.

$\eta_i(\phi)$ is the i-th linear predictor that connects the data to the latent field.

The middle stage is formed by the latent Gaussian field. where we attribute a Gaussian distribution with mean $\mu(\theta)$ and sparse precision matrix $\mathrm{Q}(\theta)$ to the latent field ϕ is conditional on the hyperparameters θ, that is:

$$\phi \mid \theta \sim N\big(\mu(\theta), Q^{-1}(\theta)\big) \tag{2.21}$$

The top stage is the hyperparameter model (or hyperpriors), which is the prior distribution for the hyperparameters.

$$\theta \sim \pi(\theta) \tag{2.22}$$

Let the response variable $y = (y_1, y_2, \dots y_n)$ belong to the exponential family (like Gaussian, Poisson, Binomial distribution and etc), where the mean u_i is linked to a structured additive predictor η_i through some link function g(.) such that g (u_i)= η_i.This predictor accounts for the different covariates additively and a structured additive regression model is given by:

$$\eta_i = \alpha + \sum_{j=1}^{n_l} f_{(L)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i \quad (2.23)$$

Where:

The α term represents the overall mean.

$f_{(L)}$ are unknown functions of the covariates.

β_k represent linear effects of the known covariates z_i

ϵ_i are the error terms.

ϕ is the vector of all unknown parameters, $\theta = (\alpha, f_{(L)}, \beta_k, \epsilon_i)$.

This serves as the overall definition, and if a model aligns with this layered framework, it can be classified as a Latent Gaussian model (LGM). At times, recognizing whether a specific model conforms to this hierarchical structure isn't straightforward. The terms can manifest in various ways: smooth, nonlinear impacts, time trends, seasonal effects, random intercepts, slopes, and temporal or spatial random influences (Martins et al., 2013). This adaptability renders the latent Gaussian model category highly versatile and capable of embracing an extensive array of models—from generalized to dynamic linear models such as spatial and spatiotemporal models (Rue et al., 2009).

2.4.2.1 Computation of LGM

Using the MCMC method to compute complex LGM model posteriori can still be problematic. This is partly due to the potential strong correlation among latent field components and the interdependency between hyperparameters and latent parameters, especially when dealing with many latent parameters (Martins et al., 2013). Integrated Nested Laplace Approximation (INLA) is specifically designed for LGMs and especially focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF) (Rue et al., 2009). It operates as a deterministic algorithm, employing Laplace approximations and a set of carefully chosen points to estimate posterior distributions. Compared with different MCMC algorithms, the INLA approximation technique for the latent Gaussian dramatically reduces the computational requirements. More details can be found in

(Rue et al.,2017).

2.4.3 Gaussian Random Field

As described in the previous section, the latent Gaussian model is used to describe the relationship between observed data and Gaussian random field. In the latent Gaussian model, observed data are assumed to be generated by an unknown Gaussian random field, which is then observed through an observation model. This observation model is typically a linear or nonlinear transformation that maps the Gaussian random field to the observation space (Haining, 2003b). Therefore, the latent Gaussian model allows for the consideration of uncertainty in modelling observed data and can describe spatial correlation structures through parameterization of the Gaussian random field (Adler, 2009). The Gaussian random field is an essential concept in spatial models, and it is also important for geostatistics (Cressie, 1993; Diggle & Ribeiro,2006). A Gaussian Random Field (GRF) is a type of random process characterized by the property that the joint distribution of any finite set of points on its domain follows a multivariate Gaussian distribution; for instance, geostatistical data or point-referenced data is the realization of a spatial process guided by a spatial point index. This means that in a Gaussian Random Field, the joint distribution of the corresponding random variables can be described by a multivariate Gaussian distribution for any set of locations. Specifically, a random variable is associated with each point for a Gaussian Random Field defined over a specific domain (Diggle & Giorgi, 2019). A mean function and a covariance function describe the joint distribution of these random variables. The mean function represents the average level of the random field at each point, while the covariance function represents the correlation between different points (Cressie, 1993). Usually, $S(\mathrm{x})$ is used to denote a Gaussian Random Field on a specific domain, where x represents a location on the domain. The mean function is typically denoted by $u(\mathrm{x})$, and the covariance function is usually donated by C (x, x'), representing the covariance of random variables at positions x and x' (Cressie, 1993).

Next, we introduce two essential properties of Gaussian random fields: isotropy and stationarity. In traditional geostatistics, the spatial process, which is defined by a Gaussian field, is often

established under the assumption of isotropy and stationarity (Cressie & Moores, 2022). Stationary means all statistical properties are invariant under translation, which means properties do not change at any point of the study area, and they are only a function of the relative position of two locations. That is to say, the covariance function only depends on two relative locations: $C(S_i, S_j) = C(S_j, S_i)$, then the GRF is stationary. Isotropy means that the process is invariant under rotation, which means the statistical properties all depend on the Euclidean distance $||\mathbf{s}_i\text{-}\mathbf{s}_j||$. The direction of the spatial point does not make an effect. These spatial models are stationary isotropic models. For whom, moving or rotating the map does not make a change. Specifically, the continuously indexed spatial process is a Gaussian random field when there is at least one spatial location S_i $(i \geq 1)$ (Haining, 2003b). The vector of all finite collections $y(S_1, S_2, \dots, S_n)$ are jointly Gaussian distributed with the mean function $\mu = \mu(S_{i\,(i\geq 1)})$ and spatial structured variance-covariance matrix $\mathbf{\Sigma} = \left(C(\mathbf{s}_i, \mathbf{s}_j)\right)$. Here we consider using the simplest Gaussian random field as example, which is the stationary and isotropic gaussian random field. So, the covariance between two locations is $COV\left(y(S_i), y(S_j)\right) = C(S_i, S_j)$, meaning the dependence between gaussian random field S (x_i) and S (x_j) is only determined by the locations. The assumption of $S(x)$ being stationary and isotropic has the inference like:

$$\text{Cov}\{S(x), S(x')\} = \sigma^2 \rho(u; \phi) \tag{2.24}$$

Where σ^2 is the variance of the Gaussian random field S. $\rho(u; \phi)$ is the spatial correlation function.

2.4.3.1 Gaussian Random Field Defined by Stochastic Partial Differential Equation (SPDE)

As our need to model complex phenomena increase, the stationary and isotropic assumptions are frequently impractical due to the effect of regional factors on the correlation pattern in reality. When using areas such as coastal regions for aquatic species modelling, a stationary isotropic model could

not be aware of physical barriers like coastlines. Similarly, when studying atmospheric phenomena, assuming the same spatial correlation across the entire area might not hold due to influences from topographical factors like altitude, rivers, and lakes altering the spatial dependence structure (Lindgren, 2011).

We are not content with the limitations of stationary and isotropic spatial models. So, the non-isotropic model was proposed to overcome the isotropic plane assumption first. Many geographic models rely on the idea of Euclidean distance, which forms the fundamental basis of the isotropic plane assumption. Modern Geographic Information Systems possess the capability to challenge the assumption of an isotropic plane, so non-isotropic models introduce directional influences that affect the spatial process (Haining, 2003). The SPDE method (Lindgren, 2011) implements a non-isotropic model by allowing a non-isotropic Laplacian and including a directional derivative term. The second thing to overcome is to build a non-stationary model. Motivated by the necessity to capture intricate dependence patterns in environmental occurrences, developing non-stationary spatial modelling frameworks has been a vigorously researched domain for over twenty years. Many people have proposed different approaches, but they can be broadly divided into two categories. The first class of methods is through warping the spatial domain into a space where the process is stationary. For example, (Paul et al., 1989) use biorthogonal grids from morphometrics to depict the thin-plate spline mappings. These mappings represent anisotropy and non-stationarity of the sample covariance matrix. The second class of methods focuses on how to induce the properties of covariance functions according to the dependence on local covariate information chosen, and SPDE also belongs to this class.

2.4.3.2 Family of Spatial Correlation Functions

According to Tobler's first law of geography (Tobler, 1993), "Everything is related to everything else, but near things are more related than distant things." This law was proposed by geographer Waldo Tobler, and it emphasizes the idea that things closer together in geographical space tend to have stronger relationships or interactions than those farther apart. So, in terms of spatial correlation

functions to possess, we may want to except it has the following two properties (Diggle, 2013): 1. When the distance between x and x' increases, the correlation between $C(x)$ and $C(x')$ is not expected to increase. 2. Secondly, given the variability in units used to measure distance, the correlation function must encompass a scaling parameter. In general, the spatial covariance function needs to satisfy some basic properties, such as positive definiteness and symmetry $(C(x,x') = C(x',x)$. This is a very important property that ensures the consistency of the probability distribution of the random field (Haining, 2003). There are several correlation functions commonly used. We define $\rho(u)$ is spatial correlation function and u is the distance between x and x'.

1.

The Matern function (Matérn, 2013) has become the most popular family of geostatistical correlation functions. The (isotropic) Matern -correlation has the inconvenient formula.

$$\rho(u;\kappa) = \frac{1}{\Gamma(\kappa)2^{\kappa-1}}(\sqrt{8\kappa}|u|)^{\kappa}K_{\kappa}(\sqrt{8\kappa}|u|) \tag{2.25}$$

$\Gamma(.)$ is a gamma function and is a modified Bessel function of the second kind of order κ. The parameter controls the differentiability of the correlation function and GRF, with small κ producing surfaces with sharp spikes and dips and large κ given more gently rounded surfaces.

2. Gaussian correlation function

$$\rho(u;\kappa) = \exp\left(-\frac{u^2}{\kappa^2}\right) \tag{2.26}$$

The Gaussian model is also a commonly used spatial correlation function.

3. Exponential correlation function

$$\rho(u;\kappa) = \exp\left(-\frac{u}{\kappa}\right) \tag{2.27}$$

The scaling parameter κ is the single parameter which is also called range parameter. When there is a fixed unit of distance, a larger value of ϕ indicates that spatial correlation exists over a longer range.

2.4.3.3 Bayesian Model Comparison

In Bayesian modeling, we often encounter situations where several possible models are available, and we want to find the optimal one. These models may involve only minor changes during the modeling process, such as different variations in the prior of a random variable. Alternatively, they may exhibit more noticeable differences, such as incorporating different fixed effects (covariates) and random variables. In addition to validating results through cross-validation tests, model-fitting comparisons are frequently conducted by evaluating specific indices. Within the Bayesian framework, we compare different models using the Deviance Information Criterion (DIC) (Spiegelhalter, 2002), which is expressed as follows:

$$DIC = \overline{D} + pD \tag{2.28}$$

DIC, the deviance information criterion which is a hierarchical modeling generalization of the Akaike information criterion (AIC) (Akaike, 1974). It is especially useful in Bayesian model selection, and generally, the smaller the value of the deviation information criterion DIC, the better the model. When evaluating two models, a general guideline is that if the difference between two DIC values greater than 5, the model with the smaller DIC value is preferred. A DIC difference of less than 5 implies that the two models cannot be discerned significantly based on DIC. (Lunn et al., 2012). $\overline{D}$ is the posterior mean of the deviance and pD is the effective number of parameters. They are used to measure the goodness of fit and the complexity of the model respectively. In terms of model complexity, based on the Occam's razor principle, under the same goodness of fit, we typically opt for the simplest model.

Chapter 3
Spatial Discrete Model for Area-level Epidemiological Data

3.1 Introduction of the Spatial Data

Before we introduce spatial models, let's briefly review the types and characteristics of spatial data, as they determine the design and application approach of spatial models, as well as the appropriate methods and techniques we need to adopt when dealing with spatial data.

3.1.1. Spatial Data Type

Spatial data supports decisions in many fields, including the environment, public health, ecology, agriculture, urban planning, economy, and society. These data come from various sources and are provided in multiple formats. According to Cressis (1993), spatial data usually could be seen as the result of an observed stochastic process occurring in space.

$$\{Z(s): s \in D \subset \mathbb{R}^d\}$$

Where $Z(s)$ denotes the attribution which we observe at generic locations s and where D is a random set of $\mathbb{R}^2$. According to the structure D, there are three main types of spatial data: areal (or lattice) data, geostatistical data, and point patterns can be distinguished (Cressis, 1993).

For areal or lattice data:

In area or lattice data, the domain D comprises a set number of areal units where variables are

observed. Aggregating the occurrences of a particular variable over a region is frequently illustrated using this type of data. For example, in spatial epidemiology, individuals with a specific illness are often aggregated within administrative areas or say, counts data are spatially aggregated from case event data. By taking into account the neighbourhood's configuration structure and other factors known to affect disease risk, these area or lattice data can be analysed to understand spatial patterns and identify disease risk factors (Moraga, 2023). In spatial epidemiology, this is the most easily obtained type of data, and the associated modelling methods have seen the richest development (Lawson et al. 2016). Because both point-referenced and point-pattern data involve specific geographic locations and the individuals on maps may be identified, they both require a higher level of confidentiality (Lawson, 2013).

For point referenced data (Geostatistical data):

In point referenced data, or geostatistical data, D the subset of $\mathbb{R}^2$ is used to define a continuous spatial field. We use spatial index $s = \{s1, s2 \dots sn\}$ to define the locations of every record in this spatial field and $\{Z(s1), Z(s2) \dots, Z(sn)\}$ are observed data at known spatial locations s. Geostatistics in epidemiology is usually confined to modelling environmental exposure at a small scale. For example, using spatial interpolation to predict relative risks at other locations based on the relative risks from existing locations.

For point pattern data:

The D is a set of points in $\mathbb{R}^2$ where case or event happened. The Domine D is random, representing the location of the occurrence an event themselves are random. The index set $s = \{s1, s2 \dots sn\}$ means the locations of the random case event data, and $Z(s) = \{Z(s1), Z(s2) \dots, Z(sn)\}$ usually equal to one which means the occurrence of random case or event. Such data is usually described for events that occur at random. Point patterns emerge when the variable being studied corresponds with the occurrence locations of events. This type of data used in spatial epidemiology usually is not for risk modelling or specific features of a particular disease or outcome but focuses on the data analysis of observed spatial point patterns of events, like the understanding spatial patterns of disease transmission or assessing the impact of sources of pollution on the health status of communities (Lawson et al. 2016).

3.1.2 Spatial Data Property

Then, we introduce two characteristics of spatial data because it poses challenges to traditional statistical modelling methods and determines the spatial structure in the model. They are spatial dependence and spatial heterogeneity.

a). Spatial Dependence

For spatial or spatiotemporal data, the spatial observation values must have spatial dependence, which means that the data in a particular area usually contains some information about the data values of the same variable in other nearby areas. Characteristics or events in a specific area do not occur entirely randomly but are affected by neighbouring areas or environments (Haining and Li, 2022). This means that spatially close locations may show similar characteristics or trends. In contrast, locations far apart may have different characteristics or trends when there is a positive spatial dependence.

b). Spatial Heterogeneity

Sometimes, heterogeneity is denoted as the second fundamental characteristic of spatial data, following spatial dependency. Spatial heterogeneity refers to the nature of differences or different characteristics between different regions or locations in space. In different spatial regions, there may be various characteristic structures or attributes (Haining and Li, 2022). These differences lead to uniqueness between areas.

3.2 Introduction of the Spatial Discrete Model

In this chapter, we focus on the discrete spatial model for areal (lattice) data. Due to the relative accessibility of regional epidemic data, such models have seen widespread development and application in epidemiological modeling. Information sharing is the core idea of spatial modeling, and the dependence of spatial data makes it possible. The information sharing, we consider in spatial

modeling is based on our prior assumptions about the spatial dependence structure of region-specific parameters, and the hierarchical model provides a powerful modeling and inference foundation for discrete spatial modeling, seamlessly combining various parameters and random effects in discrete spatial models. So, we need to incorporate this spatial dependence structure for estimation purposes in Bayesian model to realize these information sharing processes. We will first discuss the conditional autoregressive (CAR) structure introduced by Besag (1974), followed by the intrinsic conditional autoregressive (ICAR). Before we talk about the famous Besag–York–Mollié (BYM) model (Besag, 1991), which is the additive combination of ICAR structure and an exchangeable structure, we also introduce the Proper CAR (pCAR) structure. Afterward, we will introduce the ecological regression model, which incorporates the fixed effect into the hierarchical model, as well as the adaptive conditional autoregressive model, which is a CAR structure with a flexible spatial weight matrix. Finally, we elaborate the principle on how to construct spatiotemporal non-separable model and the spatiotemporal interaction variables (Knorr-Hold, 2000).

3.3 Spatial Weight Matrix

First, we need to introduce spatial weight matrices, which are matrices used to describe the spatial relationships between different geographic units. They are an indispensable component of discrete spatial models.

3.3.1 Specifying Spatial Relationships by The Weight Matrix.

The W matrix is an essential component in all forms of spatial analysis. Here, we will introduce how to use the spatial weight matrix W to define the relationships among polygons in space. In spatial data modeling, the spatial weights matrix, denoted as W, holds significance as it provides information about the spatial connections among the referenced areas where the observed data are located geographically. These spatial connections among areas represent a vital aspect of the data, and generating the W matrix enables us to translate these connections into numerical representations (Cressie, 1993). The spatial related data, which is the weight matrix W, is often based on our

assumptions about how the areas are connected to each other. This includes methods for defining adjacency between different regions and the strength of connections among these regions. These methods can be based on the geographical layout within administrative boundaries (where most commonly used polygons are based on different administrative divisions), or they can rely on attribute variables between the regions of interest to define the strength of connections among areas (Haining and Li, 2022). For more details about the spatial weight matrix, please see Fischer and Getis (2010) Haining (2003) and Haining and Li (2022). R package 'spdep' (Bivand, 2022) provides a variety of methods for generating spatial weight matrix.

3.3.2 How to Define the Spatial Matrix W.

In modelling a spatial dataset with N areas, a spatial weights matrix, denoted as W, is an N×N matrix that outlines our chosen method for representing the spatial connections between these areas. ω_{ij} to denote the element of W on the ith row and jth column (i=1…, N and j =1…, N). It is used to describe the spatial relationship between the area represented by the ith row and the area represented by the jth column.

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

We will introduce several commonly used methods for defining elements within the W matrix. Although different ways of defining the W matrix exist, it should be noted that the most frequently employed criteria are the contiguity and geographical distance.

3.3.3 Contiguity-Based Spatial Weights

Contiguity refers to the situation where two spatial units share a boundary of non-zero length. There are two methods under the definition of contiguity. Practically, we can further distinguish between a

rook and a queen criterion of contiguity, akin to how the respective chess pieces move on a chessboard (Cliff and Haggett, 1988). The rook criterion establishes neighboring units by having a shared boundary between them. The queen criterion is broader, considering units as neighbors if they share either a boundary or a vertex. Consequently, the number of neighbors following the queen criterion will always be equal to or greater than that under the rook criterion. In these two methods, $\boldsymbol{W}$ is a symmetric matrix, which means $W_{ij} = W_{ji}$.

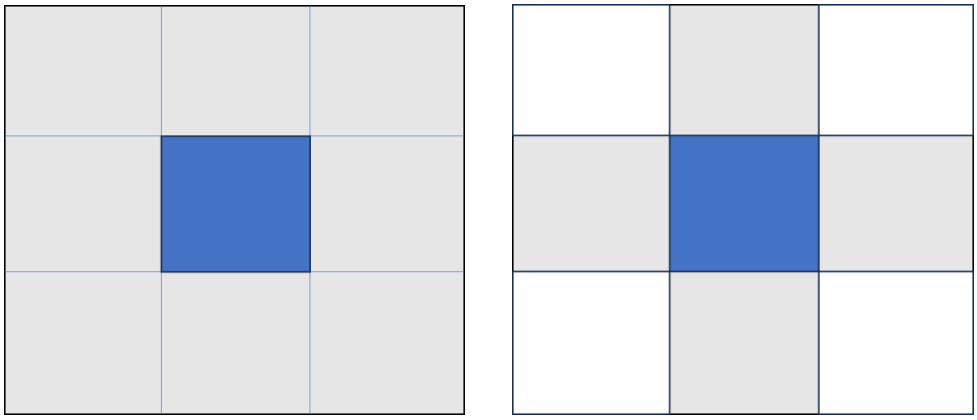


Figure 3.1 Schematic representation of spatial contiguity definitions. The blue squares are the target areas, and the light gray ones are the defined neighbors (A) queen criterion of contiguity. B) rook criterion of contiguity.

For the first method, which is called "rook's move." Only common borders of the polygons are considered to define the neighbour relation (common vertices are ignored). In this method, if areas *i* and *j* share a common border, then $w_{ij} \neq 0$. In other words, these two areas are defined as neighbors. Typically, $w_{ij} = 1$. If two areas do not share a common boundary, then $w_{ij} = 0$ and they are not neighbors. The second method, which is called "queen's move." The difference between the rook and queen criterion to determine neighbors is that the latter also includes common vertices. In this method, if they share either a common border *or* a common vertex (two areas touch at just one point). Otherwise, they are considered as non-neighbors ($w_{ij} = 0$). This would make the greatest difference for regular grids (square polygons), where the rook criterion will result in four neighbors (except for edge cases) and the queen criterion will yield eight. In small areas like census tracts (which are almost

irregular grids), the distinction between the significance of a "queen's move" versus a "rook's move" is much less. This is due to their small size and often irregular shapes. Either of methods of defining Contiguity is reasonable. However, to handle potential inaccuracies in the polygon file, such as rounding errors, practical advice suggests using the queen criterion (Li and Haining, 2020). In addition to this binary definition, there is another called proportion of border based spatial weights, usually used for irregular space areas (polygons). This approach uses common boundary proportions. It involves refining the spatial contiguity weights matrix by determining each weight (representing the strength of contiguity) based on the length of the shared border as a fraction of the total border of area i. That is:

$$W_{ij} = \left(\frac{l_{ij}}{l_i}\right)^a$$

Where l_{ij} is the length of the shared border between i and j and l_i is the length of the border of spatial unit i.

However, in certain specific geographical contexts, defining a spatial weight matrix based on contiguity is worth discussing:

1. If the spatial system comprises blocks of areas that are non-contiguous such as an island. For this kind of spatial unit i, the i_{th} row of the W matrix under spatial contiguity would be entirely zeros. This is not permitted when using spatial weight matrices for modelling, meaning gaps are not allowed. This is also why lattice data is often suitable for small areas studies. A common approach involves connecting islands to the closest spatial units on the mainland (Haining, 2003). Alternatively, population-weighted centroids for each area could be used to calculate the distance between isolated areas and others. By setting a threshold distance, any distance of the area below that threshold can be considered a neighbor of the isolated area.

Figure 3.2 Schematic representation of the first situation.

For example, Kapiti Island is a small island off the west coast of New Zealand. When conducting spatial modelling, it may be considered adjacent to certain regions by calculating the geometric centroid or population weighted centroid.

2. When one area entirely surrounds another area, the surrounded area has only one neighbour which may be unreasonable (Anselin, 1995). Anselin (1995) pointed out that the boundary geographic units may have limited connections with external geographic units, necessitating a more careful construction and analysis of spatial weight matrices.

One approach is to consider a higher-order contiguity, such as setting the k in the k-order neighbour of the spatial contiguity for this surrounded area to 2; In other words, the neighboring status of this area extends to the lattices directly connected to its first-order neighbors.

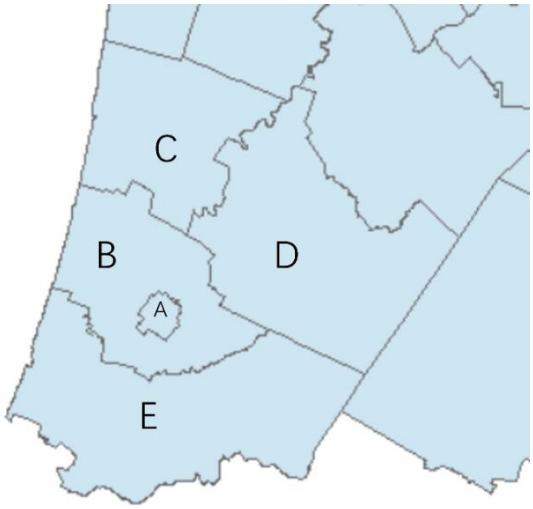


Figure 3.3 Schematic representation of the second situation.

For example, region A is completely surrounded by region B. 1st-order neighbour of area B are area C, D, and E. So, in the in the 2-order neighbour of the spatial contiguity for this surrounded area A, its neighbours are B, C, D, and E.

3.3.4 Specifying a Spatial Weight Matrix by Attached Values.

Specifying attached values as a weight matrix typically refers to defining the elements of the weight matrix based on certain attached or assigned values, which is a flexible and intuitive method. There are various approaches commonly used, including: Distance-Based Weights, K-Nearest Neighbor Weights, Threshold Weights, Attribute Similarity Weights, etc. (Fischer and Getis, 2010; Haining, 2003)

1. Geographical distances based spatial weight matrix.

This method employs a collection of points to represent polygons, referred to as lattice data. Then, it establishes spatial relationships of areas in space based on the distances between their corresponding points. Any collection of spatial units in polygonal form could be represented using a set of points. We can take the geometric centroid or population-weighted centroid of the polygon, where the population-weighted centroid is more suitable for cases where the population within the polygon is non-uniformly distributed. So, let d_{ij} be the Euclidean distance. Listed below are two different functions to defend the weight between those two points:

A). Inverse distance.

$$w_{ij} = d_{ij}^{-\gamma} (\gamma > 0) \tag{3.1}$$

B). Negative exponential.

$$w_{ij} = \exp(-\lambda \cdot d_{ij}) \, \lambda > 0 \tag{3.2}$$

2. Attribute values based spatial weight matrix.

Contiguity can be defined in terms of how similar two areas are based on some attribute.

$$w_{ij} = \left|x_i - x_j\right|^{-1} \tag{3.3}$$

x_i and x_j are attribute values of area i and j respectively.

3. Network-based spatial weight matrix.

Long-distance interaction has always been a challenge in spatial statistics (Riley et al. 2015). In spatial statistics modelling based on lattice data, using a spatial weight matrix established on attribute values of interactions is one of the solutions. Information of interactions, like the movement of goods, people, or communications, can determine specific values. The core notion is that tangible movements could indicate the degree of connection or interaction between diverse locations (Bavaud, 1998). This could assist in clarifying differences in spatial diversity. This method considers the flow paths of people or goods, using factors such as the size of the flow and distance to determine the spatial connectivity strength between units. Here we introduce one network-based spatial weight matrix proposed by Bavaud (1998). It shows two different ways to define the weight matrix:

Export - based weights: $w_{ij} = \frac{f_{i\to j}}{f_{i\to .}}$

Import - based weight: $w_{ij} = \frac{f_{j\to i}}{f_{\to i.}}$

$f_{i\to j}$ measures the export from area *i* to area *j*.

$f_{j\to i}$ measures the import from area *j* to area *i*.

$f_{i\to .}$ is the total export from *i* to all other areas in the study region.

$f_{\to i.}$ measures the total import from all other areas to *i*.

The W matrix serves a crucial function by transforming spatial connections among N areas or points into tangible information. By utilizing models, this type of data supplements traditional geographic referenced datasets (Bavaud, 1998). This augmentation enables us to enhance estimations at a localized level, explore how characteristics of one area influence others (such as spatial spillovers), and derive a more precise understanding of how a known factor impacts an outcome (Haining and Li,

2022). It is also a good structure used for long distance interaction in spatial epidemiology modelling.

3.4 Spatial Dependence Structure

3.4.1 CAR structure

CARs are commonly motivated for borrowing information and modelling smoothly varying disease risks as effects of omitted covariates (MacNab,2022). Given a set of observations collected across N distinct areas within a study area. Let S_i represent the parameters specific to each area, labeled from S_1 to S_N, highlighting they are spatially structured.

In this normal condition distribution, each S_i follows a conditional normal distribution with mean equal to the sum of the weighted values of its neighbours and has unknown variance.

$$S_i \mid S_j, j \neq i, \sim \mathrm{N}\left(\sum_{j=1}^{n} w_{ij} S_j, \sigma^2\right) \tag{3.4}$$

The joint distribution by the local specification of the conditional distributions of the individual random variables defines a Gaussian Markov random field (GMRF). Besag (1974) showed that the corresponding joint specification of S_i is a multivariate variable followed by a normal distribution with 0 mean.

$$\mathbf{S} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\Sigma} = \boldsymbol{\sigma^2 D^{-1}(I_N - \rho W)^{-1}} \tag{3.5}$$

The variance-covariance matrix Σ, which is a function of W, determines the properties of local information sharing.

Where:

$\boldsymbol{S_j}$: area-specifc parameters excluding $\boldsymbol{S_i}$ itself.

W: the N x N spatial weight matrix. We use the simplest spatial weigh matrix is the binary adjacency spatial weight matrix as example, and the all diagonal elements {i,i } are zero. If area i and j are adjacent, the off-diagonal elements {i,j} are 1 otherwise 0.

D: $N \times N$ diagonal matrix (all off-diagonal elements equal to zero) and where {i,i} are the number of neighbors of area i.

ρ : spatial correlation parameter which controls the amount of spatial autocorrelation; ρ=0 implies spatial independence and ρ=1 implies complete spatial correlation.

$\boldsymbol{I}_N$: N*N identity matrix

3.4.2 ICAR structure

An Intrinsic Conditional Auto-Regressive (ICAR) model (Besag, 1991) is a CAR model when spatial correlation parameter α=1, so it assumes complete spatial correlation between areas.

Let's still take the example of defining adjacent structures with a spatial weight matrix **W** with binary (0 or 1) entries. The corresponding conditional distribution specification based on binary weight matrix of ICAR model is:

$$S_i \mid S_j, j \neq i, v, W \sim N\left(\frac{\sum_{j \in \Delta i} S_j}{m_i}, \frac{{\sigma^2}_i}{m_i}\right) \tag{3.6}$$

m_i : is the number of neighbors for region m_i .

Δi: set of neighbors of area i as defined in $\boldsymbol{W}$.

${\sigma^2}_i$: unknown variance parameter.

The ICAR model requires the introduction of an important constraint, which limits the sum of these region-specific parameters to zero (sum to zero constraint). ICAR model is also called the improper CAR model because the joint distribution of ICAR is improper (Guinness and Baak, 2021). Simply, under the ICAR model, the mean of S is not defined, so it is an improper distribution. Without constraints, the model may face multiple equivalent solutions, which can make parameter estimates

ambiguous and unstable (Besag, 1991). This is why the ICAR model is only used as a prior for a set of area-specific parameters but not as a likelihood function to model the data directly.

Specifically, In the ICAR model, the joint distribution is:

$$\mathrm{S} \sim \mathrm{N}(0, \Sigma)$$

$$\Sigma = \sigma^2 D^{-1}(I_N - W)^{-1}$$

So,

$$\mathrm{S} \sim \mathrm{N}(0, \sigma^2 D^{-1}(I_N - W)^{-1}) \tag{3.7}$$

The pairwise difference formulation of the joint distribution is:

$$p(\mathrm{S} \mid \sigma^2) \propto \exp\left\{-\frac{\sigma^2}{2}\sum_{i\sim j}\left(\mathrm{S}_i - \mathrm{S}_j\right)^2\right\} \tag{3.8}$$

From the pairwise difference formulation, we could see that the joint distribution is non-identifiable (Besag, 1991). Adding any constant to all the elements of leaves the joint distribution unchanged. When the sum of all region-specific parameters is constrained to zero, the model establishes a reference point or baseline. This means that the model is no longer affected by one overall translation but focuses more on the relative impact differences between individual regions (Keefe and Franck, 2018). So, the sum to zero constrain is necessary to ensure that the estimation of spatial effects is valid and interpretable in the analysis of spatial data.

3.4.3 Proper CAR (pCAR) structure

The ICAR (intrinsic conditional autoregressive) model is improper since the precision matrix (Σ^{-1}) is singular. It could be made proper by setting the spatial correlation parameter ρ not equal to 1 of the spatial weights (Besag,1991). The pCAR conditional distribution specification based on general weight matrix is:

$$S_i \mid S_j, j \neq i, v, W \sim N\left(\rho \frac{\sum_{j \in \Delta i} S_j}{m_i}, \frac{\sigma^2{}_i}{m_i}\right) \tag{3.9}$$

The definitions of parameters are the same as previous. Thus, the pCAR variate joint distribution is:

$$S \sim \mathrm{N}(\mathbf{0}, \sigma^2(\boldsymbol{D}_w - \rho \boldsymbol{W})^{-1}) \tag{3.10}$$

The parameter ρ is used to manage the degree of spatial correlation. When ρ>0, it indicates a positive spatial autocorrelation; conversely, when ρ<0, it indicates a negative spatial autocorrelation. and the closer ρ to 1 indicates the stronger spatial autocorrelation.

3.4.4 Besag–York–Mollié (BYM) model

The Besag–York–Mollié (BYM) (Besag et al. 1991) Model is one of the most popular regional data spatial models used in epidemiology, which takes into account the data in a positive spatially correlated. The model combines two model components: a spatial random effect that soothes the data according to the spatial (neighborhood) structure and an unstructured exchangeable component that models uncorrelated variability. The model can be applied to the response variables are both continuous and discrete. For example:

For continuous variable

$$y_i \sim N(\theta_i, \sigma_y^2)$$
$$\theta_i = \alpha + S_i + U_i$$

For discrete variable

$$y_i \sim Possion(\lambda_i)$$
$$\lambda_i = n_i \theta_i$$
$$\log(\theta_i) = \alpha + S_i + U_i \tag{3.11}$$

Where i means area i, the excepted number of cases λ_i which is the mean of Poisson is defined by

the rate θ_i and the population size n_i.

Compared with locally similar under various CAR specifications, both globally and locally similar are considered in BYM model. As we said before, while this ICAR model is non-generating in that it cannot be used as a model for the data, it can be used as a prior for part of a hierarchical model, which is the role it plays in the BYM model. Where S_i is a parameter from the spatially structured (ICAR) component S = (S1, … S_N) and U_i is a parameter from the spatially unstructured component U = (U1, … U_N). These two parameters are all area specific. The spatially unstructured component U in Bayesian hhierarchical model is usually set as an exchangeable structure which follows a zero-mean normal with unknown variance given a vague prior.

3.5 Ecological Regression: Disease mapping with areal-level covariates

3.5.1 Ecological Regression

When pertinent risk factors like exposures, confounders, or other relevant variables are available, they can typically be included as fixed effects in the model alongside random effects. This adjustment means the model would consist of fixed effects combined with random effects, which are specified as ecological regression models. One of the most important purposes of spatial regression is to measure disease risk variations attributed to factors that we are interested in and unexplained attributed to the risk factor (Clayton et al., 1993). It's important to highlight that when incorporating covariates (fixed effects) into the model, the interpretation of spatiotemporal effects (random effects like space, time, spatiotemporal interactions, etc.) will be residuals (Blangiardo and Cameletti, 2015). The discrete ecological regression models can be specified as:

$$\begin{aligned} y_i &\sim Possion(\lambda_i) \\ \lambda_i &= n_i\theta_i \\ log\ (\theta_i) &= \alpha + \beta_i m_i + S_i + U_i \end{aligned} \tag{3.12}$$

Where i means area i, the excepted number of cases λ_i which is the mean of Poisson is defined by the rate θ_i and the population size n_i, and $\beta_i m_i$ are covariates and covariate parameters.

3.5.2 Challenges in Ecological Regression.

One of the most obvious challenges in ecological regression is spatial confounding. In ecological regression modelling, it's common to incorporate covariates, and when these covariates vary across areas or exhibit spatial correlation, spatial confounding is highly likely to occur. Spatial confounding, that is, collinearity between fixed effects and random effects in a spatial generalized linear mixed model, can adversely affect estimates of the fixed effects (Zimmerman & Hoef,2021). There are two main ways to solve this problem: One is de-confound the fixed and random effects. Another is to formulate transformed GMRF or MGMRF, which is called TGMRF or TMGMRF, so that the fixed and random effects can be modeled independently (MacNab,2022).

3.6 Adaptive CARs

The adaptive CAR model is a flexible extension of the traditional CAR structure. It is a CAR model with adaptive weights matrix. The proposal aims to define spatial correlation by creating a stochastic spatial weights matrix W during the fitting process (Griffith,1996). The GMRFs commonly used in disease mapping are undirected graphical models commonly defined for a lattice system of nodes and edges (Rue & Held, 2005). In disease mapping, they hypothesize interactions between neighboring risks, usually resulting in a smooth change of relative risks in the posterior. However, this model has a major flaw; they often lack the adaptability to account for variations in risk both within and between neighborhoods, which are likely to occur in certain areas of the map (MacNab,2022). Therefore, the adaptive CAR model comes into being.

Until now, adaptive conditional autoregressive models have generally been categorized into two main parts. One is to hierarchically formulate adaptive CAR(GMRF) with an unknown adjacency matrix to be modeled and estimated from data. Some people proposed an adaptive CAR model in

which the non-zero elements of W were modeled as Bernoulli random variates or all elements of an unknown adjacency matrix W as a Bernoulli random variate. (Rushworth et al., 2017; MacNab,2022). Formulating adaptively parameterized CARs over a given W is the second approach. In this method, adaptive Conditional Autoregressive models can be employed to represent spatial dependencies, interactions, and heterogeneities that vary locally. These adaptive CARs might capture micro-level phenomena, such as locally varying risk dependencies and influences, and heterogeneities that typically manifest at macro-level phenomena, such as spatial risk heterogeneities and discontinuity (MacNab,2022). These CAR(GMRF) structures often contain two parameters, which are spatial and scale parameters, and the CAR(GMRF) structures typically involve either one of the two or both. According to these two parameters of CAR (GMRF), MacNab (2022) generally classify them into three main groups, they are:

1. The first category consists of CARs only with adaptive spatial parameters; they were typically motivated to model locally varying spatial dependencies.

2. The second category involves CARs with adaptive scale or precision parameters acting as spatial weights. These weights are utilized to model the interactions between pairs of risks with varying weights, addressing local differences in risk heterogeneity.

3. The third category is CARs, which incorporate both adaptive spatial and scale parameters.

3.7 Spatio-Temporal Models

The spatial data modelling of hierarchical models has been discussed. In this section, we will turn to the exploratory modelling method of spatiotemporal data. Compared with spatial data, spatiotemporal data has an additional time dimension, which is not only an increase in the amount of data but also a more complex data structure in space and time. Dependence and heterogeneity are two significant characteristics of spatial data. Now, we also need to deal with these two characteristics in the temporal or space-time interaction dimensions (Cressie, 2015). This means that the observations at a particular

place and time may differ from the observations at other points in time, space, or space-time. In addition, we also need to capture autocorrelated dependence patterns in time, space, and spatiotemporal interactions. Combining these two properties creates the need for modelling to understand complex patterns in spatiotemporal data. Relying on hierarchical modelling, this seems quite clear. The hierarchical model framework allows us to view complex structures in a modular way (Lawson, 2013). It seems that the simplest and the most straightforward way is to incorporate the spatial and the temporal main effect independently, along with the spatial-temporal interaction (Lawson, 2018). Just as we introduced in the previous section, class of CAR structures are used for spatial information sharing in hierarchical modelling. Now we need to explore different structures for information sharing in temporal and spatiotemporal dimension. But before we talk about them specifically, we first have to decide how to add them in to a hierarchical model by a modular way.

3.7.1 Additively or Multiplicatively

Before delving deeper into understanding the Bayesian spatiotemporal model, we need to consider a question: we should combine the space and time components additively or multiplicatively? This topic has been discussed by many people, please see Hastie & Friedman (2009), Gelman et al. (2013), Haining &Li (2020) and Cressie & Wikle (2015) for detail.

Consider expressing the above formula as a matrix model：

For addition structure:

$$y_{it} = \alpha + S_i + T_t$$

Written in matrix form:

$$\begin{pmatrix} y_{11} & \cdots & y_{1T} \\ y_{21} & \cdots & y_{2T} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NT} \end{pmatrix} = \alpha \mathbf{1}_{N\times T} + \begin{pmatrix} S_1 + T_1 & \ldots & S_1 + T_T \\ S_2 + T_1 & \ldots & S_2 + T_T \\ \vdots & \ddots & \vdots \\ S_N + T_1 & \ldots & S_N + T_T \end{pmatrix} \tag{3.13}$$

For multiplicative structure:

$$y_{it} = \alpha + S_i * T_t$$

Written in matrix form:

$$\begin{pmatrix} y_{11} & \cdots & y_{1T} \\ y_{21} & \cdots & y_{2T} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NT} \end{pmatrix} = \alpha \mathbf{1}_{N\times T} + \begin{pmatrix} S_1 * T_1 & \cdots & S_1 * T_T \\ S_2 * T_1 & \cdots & S_2 * T_T \\ \vdots & \ddots & \vdots \\ S_N * T_1 & \cdots & S_N * T_T \end{pmatrix} \quad (3.14)$$

In an additive model, overall spatial effect and overall spatial effect are combined by addition, so the trend of response variable y could be specific and differ by level. But we can observe, from top to bottom, the spatial pattern represented by the spatial main effects, $S = (S_1, \ldots S_N)$, stays consistent across different time points. That is to say, the spatial effects S_i are parallel to the trend of each area rather than intersect. Similarly, the temporal trends of each area mirror the overall temporal pattern outlined by the general (overall) temporal component, T= ($T_1 \ldots$, T_t). In a space-time separable model, the temporal progression of the response variable y for each area is achieved by either raising or lowering the entire collective trend. Each regional trend runs in parallel with one another, aligning to the same pattern as the overall trend T relatively.

However, the situation will become somewhat different when using the multiplication structure. Think about the scenario where you're modelling a collection of continuous outcome data by employing a normal distribution as the probability model.

$$y_{it} \sim N(\mu_{it}, \sigma^2)$$
$$\mu_{it} = \alpha + (S_i \times T_t) \quad (3.15)$$

The implication on the time trends in small areas would be complex, as it is influenced by the product of S and T. The time trend in each area could either compress (for 0<S<1) or elongate (for S > 1) the general overall pattern of change along the response variable positioned on the y-axis. And when S = 1, that specific area precisely follows the overall trend. However, when S< 0, the resulting trend in that area diverges from the overall pattern in the opposite direction. So, using an additive structure will result in a decrease in the model's interpretability. Another reason is that, when modelling count

data, it is easier to interpret the two components with the additive structure either on the log scale (for Poisson likelihood) or the logit scale (for binomial likelihood). They are interpreted differently in two scales.

Additive refers to the process of combining both components on the logarithmic or logit scale by addition (Hastie & Friedman, 2009). Combining the two components on the logarithmic scale through addition is the same as merging the two components multiplicatively on the exponential scale. Like $\theta_{it} = \exp(\alpha)$ ×exp (S_i) ×exp (T_t).

When Poisson likelihood is used

$$\begin{aligned} y_{it} &\sim \text{Poisson}(\lambda_{it}) \\ \lambda_{it} &= n_i\theta_{it} \\ \log(\theta_{it}) &= \alpha + S_i + T_t \end{aligned} \tag{3.16}$$

The interpretations of exp (S_i) and exp (T_t) could be considered as the rate ratio or relative risk. If exp (S_i)>1 is greater than 1, it implies a higher rate (or risk) in area i when compared to the overall average.

When Binomial likelihood is used

$$\begin{aligned} y_{it} &\sim \text{Binomial}\ (n_{it}, \theta_{it}) \\ \text{logit}(\theta_{it}) &= \alpha + S_i + T_t \end{aligned} \tag{3.17}$$

The interpretation of exp (S_i) could be considered as the odds ratios. For instance, exp (S_i)>1 indicates that the likelihood of an event occurring in area i is higher compared to the likelihood for the entire study region. In fact, the multiplicative formulation of (Si *Tt) could be seen as an interaction structure between space and time, but it is not for spatial and temporal main effects. According to this, it is not recommended to involve only a spatial-temporal interaction structure but without spatial and temporal main effects, respectively (Haining &Li, 2020). That is to say, when you want to include

both spatial and temporal interaction effects in the model, you should also include the main effects simultaneously. Taking the Poisson likelihood as an example, it should be:

$$\log(\theta_{it}) = \alpha + S_i + T_t + (S_i * T_t)$$

However

$$\log(\theta_{it}) = \alpha + (S_i * T_t)$$

This is not recommended.

In summary, the additive model will serve as a consistent form of hierarchical Bayesian spatiotemporal models to combine the space and time structure (Gelman et al., 2013). This would be consistently utilized in this thesis. At the same time, this is also the most mainstream approach. But here's what we should note: the constraint of this model is that it is based on the strict assumption of separability of time and space. The spatial pattern at all time points is the same, and all spatial observations in different regions share the same time trend pattern (Haining and Li, 2020). Of course, such an assumption is strict and unreasonable, which brings us to the spatio-temporal interaction variables we will introduce next.

3.7.2 The Space-Time Separable Modelling Framework

There are two frameworks in the hierarchical model used for spatiotemporal modelling. They could be divided into separable and inseparable models according to whether they contain a spatiotemporal interaction structure (Knorr-Held, 2000). In the separable model, there are two sets of terms: one is for capturing spatial structure, and another is for temporal structure. Naturally, there is an additional spatial-temporal interaction in the non-separable model.

For Space-Time Separable Modelling, it is like

$$y_{it} \sim \text{Poisson}(\lambda_{it})$$

$$\lambda_{it} = n_{it}\rho_{it}$$

$$\log(\rho_{it}) = \eta_{it}$$

$$\eta_{it} = \alpha + S_i + T_i \tag{3.18}$$

$$S_i = U_i + V_i$$

$$T_i = \gamma_t + \phi_t$$

Where:

α is the intercept.

$S_i = S_1, S_2, \ldots S_N$ is for spatial component. U_i and V_i are spatially structured and unstructured component respectively.

$T_t = T_1, T_2, \ldots T_T$ is for temporal component. γ_t and ϕ_t are temporally structured effect and unstructured effect respectively.

For temporally structured effect, we can set it followed by random walk or autoregressive with order 1 or 2 respectively.

For random walk:

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

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

For autoregressive:

$\gamma_t \mid \gamma_{t-1} \sim \text{Normal}(\boldsymbol{c} + \boldsymbol{m_1}\gamma_{t-1}, \sigma^2)$

$\gamma_t \mid \gamma_{t-1}, \gamma_{t-2} \sim \text{Normal}(\boldsymbol{c} + \boldsymbol{m_1}\gamma_{t-1} + \boldsymbol{m_2}\gamma_{t-2}, \sigma^2)$

The spatially and temporally unstructured random effect V_i and ϕ_t are often given by a zero mean gaussian with exchangeable (vague) prior $\phi_t \sim N(0, a^2)$. The advantage of using vague priors is that the data dominates the inference process and form a data-driven estimation (Carlin and Louis, 2000; MacNab, 2022), allowing the data to have a stronger influence on the estimation of the variance parameter.

3.7.3 The Space-Time Inseparable Modelling Framework --Space–time interactions

In hierarchical models, it's easy to add a structure $\boldsymbol{\delta}$ to describe the spatiotemporal interaction term, transforming the model into a spatiotemporally inseparable model. This can eliminate the previously stringent assumptions, and the model could be used in a flexible way. It is like:

$$\eta_{it} = \alpha + S_i + \gamma_t + \phi_t + \delta_{ij} \tag{3.19}$$

Where:

$\boldsymbol{\delta_{ij}}$ spatiotemporal interaction term of area i and time point j.

The definition of other variables does not change.

The spatial-temporal interaction which is used to explain the different temporal trend for different areas or the different spatial trend for different time point. Knorr-Held (2000) propose a way on establishing the space-time interaction. There are four different definition of space-time dependence structures matrix for modelling the N X T (N areas and T time points) parameters in the space-time interaction component, δ = ($\boldsymbol{\delta_{11}}$,…,, $\boldsymbol{\delta_{1T}}$, $\boldsymbol{\delta_{N1}}$,.. $\boldsymbol{\delta_{NT}}$). The interaction matrix $\boldsymbol{M_\delta}$ is the spatiotemporal structure matrix which could be factorized by Kronecker product into the structure matrix of the main effects that interact with one another, which identifying the type of temporal and/or spatial dependence between the elements of δ. δ here follows a Gaussian distribution with a precision matrix given by $\tau_\delta \boldsymbol{M_\delta}$,where τ_δ is an unknown scalar (Knorr-Held, 2000).

Type I:

In Type I, all parameters $\boldsymbol{\delta}_{\text{ij}}$ of dependence structures are similar wherever from any spatial-temporal point. The interaction parameter δ is consist of two unstructured effects: S_i spatially unstructured and ϕ_t temporally unstructured effects. So, the structured matrix is:

$\boldsymbol{M_\delta} = \boldsymbol{M_V} \otimes \boldsymbol{M_\phi}$.

Type I space-time interaction is the simplest of the four, it indicates that different configurations of

spatiotemporal parameters don't influence the determination of spatiotemporal dependency structures. In other words, regardless of whether spatiotemporal units (i, t) are close in time to (i, t+1) or (i, t-1), or are adjacent in space to (j, t), they do not participate in determining the spatiotemporal dependency structure. In this model, space-time interaction in type 1 assumes that after factoring in the main spatial and temporal effects, the residual variations in space-time don't display smooth changes across either space or time. Alternatively, if this spatiotemporal effect is used to capture residuals, the residual does not have any spatiotemporal structure.

Type II:

The dependence structure δ under type II, the temporal parameters in each area i, $\boldsymbol{\delta}_{i,1:T}=\boldsymbol{\delta}_{i1,\ldots}\boldsymbol{\delta}_{iT}$, are assumed to be temporally structured. But the time trends in different regions (even adjacent regions) are independent. The interaction parameter δ is consist of temporally unstructured effects and a spatially unstructured effect. So, the structured matrix $\boldsymbol{M}_{\delta}= \boldsymbol{M}_{\gamma} \otimes \boldsymbol{M}_{\mathrm{V}}$.

Spatial-temporal interaction in type 2 assumes that spatial effects are localised, i.e., each area is not associated with any other area. When information regarding these localized risk factors is unavailable, the space-time interaction parameters could serve as substitute measures for these effects. This is because The Type II dependence structure is well-suited for capturing unique and specific time trends in local areas that stem from highly localized risk factors without spatial spread (Haining and Li, 2020).

Type III

The dependence structure δ under type III, the area-specific parameters in each time point t, $\boldsymbol{\delta}_{1:I,t}=\boldsymbol{\delta}_{1t,\ldots}\boldsymbol{\delta}_{It}$, are assumed to be spatially structured. But the spatial patterns at different time points (even adjacent time points) like $\boldsymbol{\delta}_{it}$ and $\boldsymbol{\delta}_{it+1}/\boldsymbol{\delta}_{it-1}$ are independent. The interaction parameter δ is consist of an unstructured temporal effect ϕ_t and a spatially structured main effect U_i. So, the structured matrix $\boldsymbol{M}_{\delta}=\boldsymbol{M}_{U} \otimes \boldsymbol{M}_{\phi}$

Consequently, within this model, after considering both main temporal and spatial effects, each time point exhibits a distinct spatial pattern. In other words, the spatiotemporal interaction used to capture the residual in the model, representing unobserved risk factors, still demonstrates a spatial

structure. However, there is no predetermined temporal pattern. For instance, at a specific time point, 't,' departures from the overall spatial main effect often occur within nearby geographical regions, and this spatial structure does not persist into the subsequent time point. Therefore, models including type III are generally suitable for a broader time span, as modelling demands less precision in time (only involving global temporal effects (Haining and Li, 2020)).

Type IV：

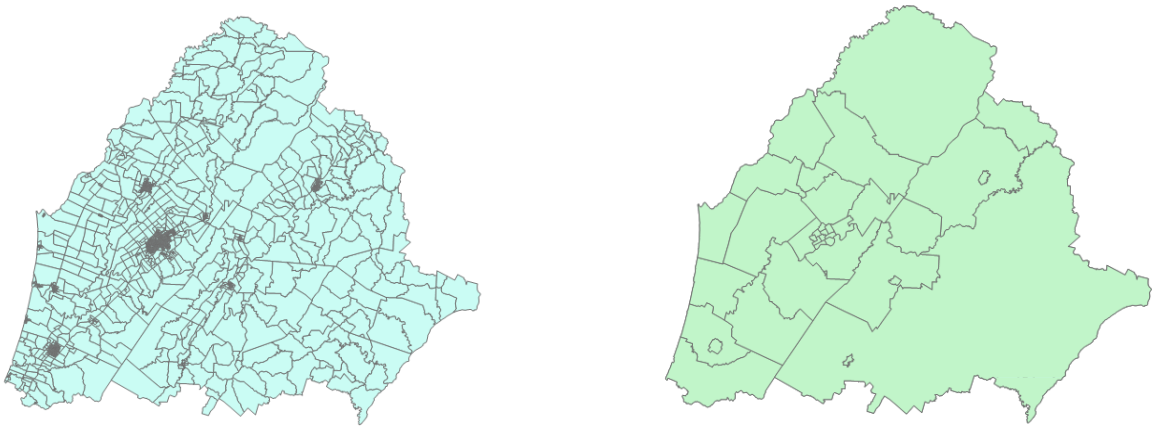
The dependence structure for $\boldsymbol{\delta}$ is no longer just over time (as in II) or just over space (as in III) but is fully dependent on space and time. The interaction parameter δ is consist of two structured effects: S_i spatially and ϕ_t temporally structured effects respectively. So, the structured matrix $\boldsymbol{M}_{\delta}= \boldsymbol{M}_U \otimes \boldsymbol{M}_{\gamma}$

Under this structure, each spatial-temporal interaction $\boldsymbol{\delta}_{it}$ is not only rely on its spatial neighbour $\boldsymbol{\delta}_{jt}$ (area j is assumed to be adjacent to area i.) also rely on its temporal neighbour $\boldsymbol{\delta}_{it+1}/\boldsymbol{\delta}_{it-1}$. This can basically be understood as ttemporal dependency structure for each area is not independent from all the other areas anymore but depends on the temporal pattern of the neighboring areas as well. Spatial dependency structure at each time point is not independent from all periods anymore but depends on the spatial pattern at the adjacent time points.

Chapter 4
Case study of the Campylobacteriosis in Manawatu, New Zealand

4.1 Introduction

Campylobacteriosis stands out as the most frequently reported notifiable infectious disease in New Zealand, happened in numerous other nations (SPENCER et al, 2012). This study focuses on Campylobacteriosis cases using more than ten years of recent data from March 2005 to 2016 across 1834 mesh blocks within 36 subregions in the Manawatu region of New Zealand. The original epidemiological data have high precision regarding geographic location but typically come with high confidentiality. Therefore, when they are disclosed, they are aggregated to a larger geographic level. Here, the Campylobacteriosis data are recorded at the mesh block level, and we aggregate mesh blocks into regions (larger geographic units), totaling 36 regions.



A) B)

Figure 4.1 A) Records by the mesh blocks. B) Records by the regions

Then, spatial/spatio-temporal statistical models are used for smoothing. We will use the INLA method for Bayesian computing of the original BYM (the Besag, York, Mollié) model, the spatial-temporal model with ICAR structure, and the ecological model. The rough expected number of cases is the number of cases that would occur in a community if the disease rate in a more extensive reference population (usually the state or country) happened in that community. So, n_i, the population size of people at risk of each region is calculated by:

$$n_i = m_i \left(\frac{\sum_{i=1}^{36} O_i}{\sum_{i=1}^{36} m_i} \right) \tag{4.1}$$

Where m_i is the number of people of region i and O_i is the number of observed cases. This is a coarse one, and ideally, is to use the age-sex-standardised group to calculate the expected number. The comparison of the risk of the regions is done by relative risk (RR). The relative risk (RR) is calculated by using observed cases divided by the expected cases.

Relative Risk (RR)=Observed / Expected cases

In addition, we also consider some covariates in the model, which are typically believed to be related to the response variable (here, it is the relative risk of Campylobacteriosis). Since campylobacter can colonize farm animals (cattle and sheep), it is widely present in most warm-blooded animals (Hoque et al., 2021). In this study, we added the per capita farm numbers, sheep density, and Social Deprivation Index (SDI) values for 36 regions as covariates. This study serves as a study of the quantifying risk of environmental exposures.

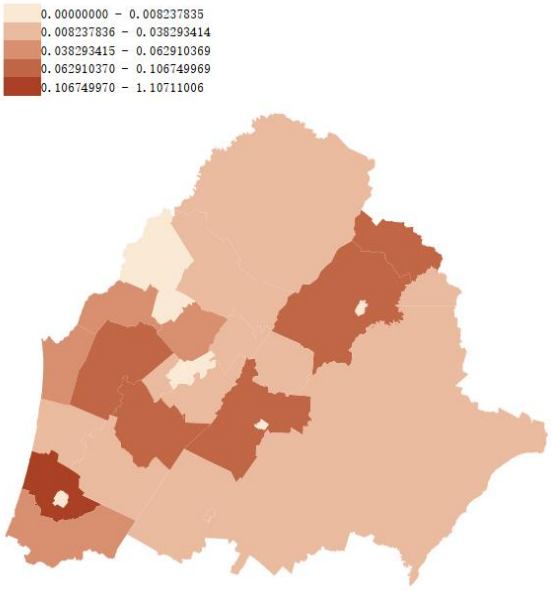


Figure 4.2 Dairy Cattle Density

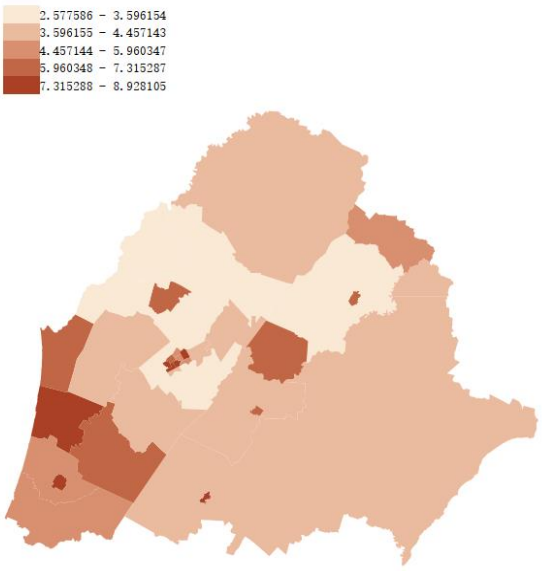


Figure 4.3 Social Deprivation Index (SDI)

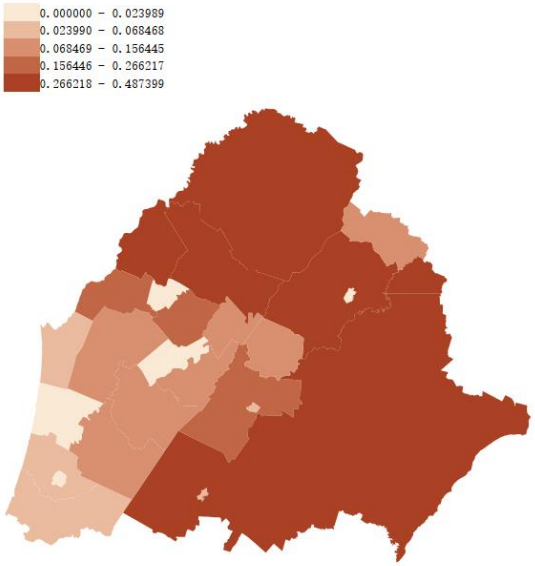


Figure 4.4 Sheep Density

The three covariates all exhibit a clear spatial pattern. In terms of per capita farm quantity, the number of farms per capita in urban areas is generally low, while it is relatively high in other areas. The Social Deprivation Index (SDI) values in urban areas are relatively high and relatively low in places farther from urban areas. Some distinct features can also be observed in the per capita sheep ownership. The per capita sheep quantity in urban areas is relatively low. At the same time, it can be observed that in the southwest direction of this region, the per capita sheep quantity is also relatively low. It can be seen that these three covariates all exhibit distinct urban/rural characteristics. So, we suspect that most of the other effects are probably just an urban/rural effect. We will also consider the urban/rural indicator.

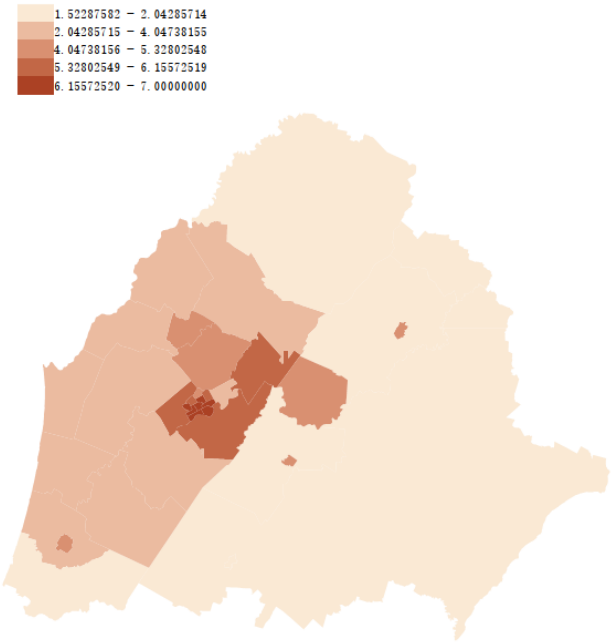


Figure 4.5 Urban/Rural Indicator

4.2 Spatial Modelling

4.2.1 BYM Modelling

First, let's consider the most basic BYM model. Bayesian models like the Besag-York-Mollie (BYM) model for small-area spatiotemporal modelling can effectively investigate regions of high risk for Campylobacteriosis and evaluate the joint role of sociodemographics.

$$y_i \sim Possion(\lambda_i)$$
$$\lambda_i = n_i\theta_i$$
$$\log(\theta_i) = \alpha + u_i + v_i \quad (4.3)$$

Where α quantifies the average incidence rate of all study areas. u is the spatially structured residual, modelled using the ICAR specification and v is the unstructured residual modelled using exchangeability among the 32 boroughs. The excepted number of cases λ_i which is the mean of Poisson is defined by the rate θ_i and the population size n_i.

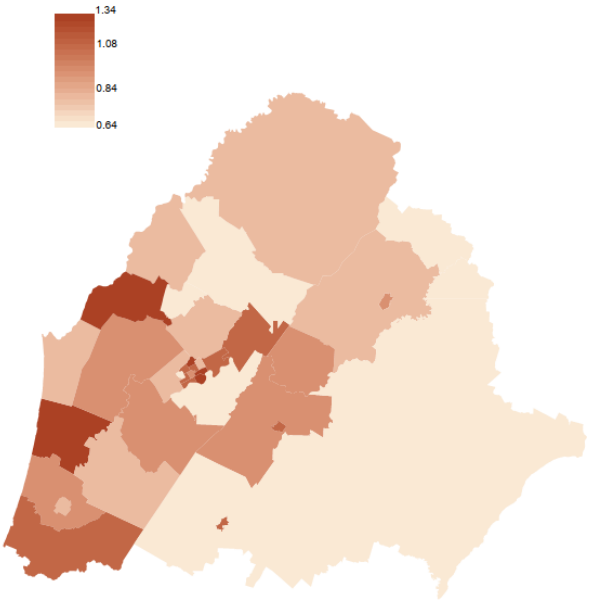


Figure 4.6 Posterior mean of the relative risks (RR=exp $(u_i + v_i)$) of each subregion compared with the whole Manawatu.

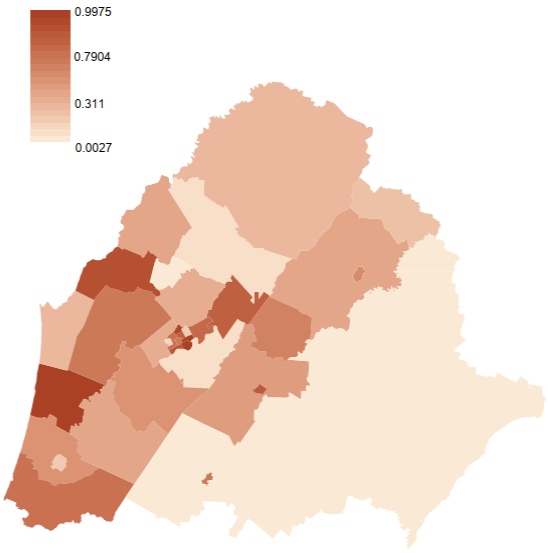


Figure 4.7 Posterior probability for the relative risk shown in figure 3.5 larger than 1:

$$p(RR_i > 1 \mid \boldsymbol{y})$$

4.2.2 Ecological Model

Then, we consider incorporating explanatory variables into the model, transforming it into ecological regression. All covariates have been standardized before being added to the model.

The ecological regression models can be specified as

$$\begin{aligned} y_i &\sim Possion(\lambda_i) \\ \lambda_i &= n_i\theta_i \\ \log(\theta_i) &= \alpha + \beta_i m_i + u_i + v_i \end{aligned} \tag{4.4}$$

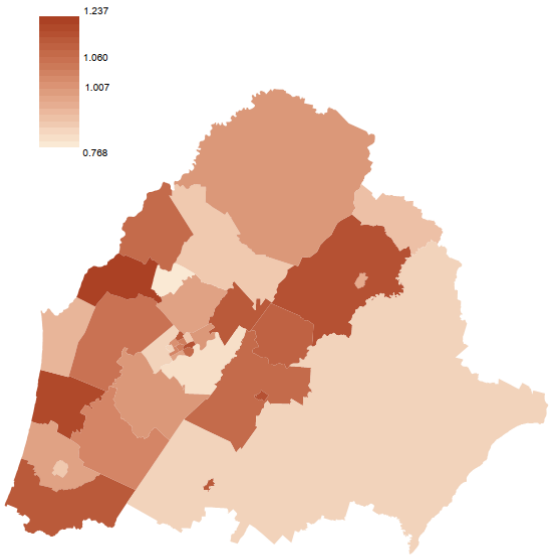


Figure 4.8 Posterior mean for the areas specific relative risk (RR=exp ($S_i + U_i$)) compared with the whole Manawatu in ecological regression.

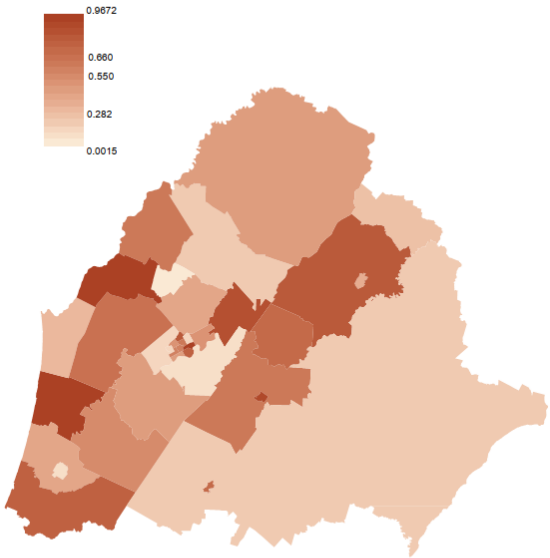


Figure 4.9 Posterior probability for the relative risk shown in figure 3.5 larger than 1 in ecological regression: $p(RR_i > 1 \mid \boldsymbol{y})$

Table 4.1 Summary statistics of the ecological regression model, there are posterior mean, posterior standard deviation (SD) and the 95% confidence interval for the explanatory variables.

	Mean	Sd	0.025	0.5	0.975
Intercept	-0.0641	0.0456	-0.1541	-0.0641	0.0257
SHEEP	-0.2682	0.0783	-0.4236	-0.2678	-0.1151
DAIRY	-0.0158	0.0418	-0.0985	-0.0158	0.0667
SDI	-0.0707	0.0517	-0.1727	-0.0707	0.0317
Urban/Rural	-0.0493	0.0647	-0.1774	-0.0491	0.0778

The fixed effects estimated by INLA are shown in the table above. We can see that covariate SHEEP is significant at a 95% confidence level because the sign doesn't change. All covariates were negatively correlated with the risk of infection with Campylobacter, where the number of sheep in the region is strongly associated with the risk of infection with Campylobacter. The posterior mean of the corresponding regression coefficient is -0.2682. At the standardized scale, the number of sheep is related to a reduction of 23.52% in the risk of infection with Campylobacter. DAIRY, which is the number of dairies, SDI (Socioeconomic Deprivation Indexes), and Urban-Rural indicators, have a weak association with the risk of infection with Campylobacter.

4.2.3 Model Comparison

Recall the Bayesian model comparison introduced in 2.4.2. We compare the following models according to the indicators shown in the table 3.1.

Ecological model 1: $\beta_i m_i$= Intercept+ Sheep Density + Dairy Cattle Density + SDI

Ecological model 2: $\beta_i m_i$= Intercept+ Sheep Density + Dairy Cattle Density + SDI+ Urban/Rural

Table 4.2 A Comparison of the original and ecological model using DIC, and WAIC

Model	$\bar{D}$	pD	DIC	$WAIC$	$pWAIC$
BYM model	234.93	22.11	257.04	258.97	18.27
Ecological model 1	237.96	17.08	255.04	257.71	15.82
Ecological model 2	231.9	20.91	252.81	253.79	16.62

DIC is the deviance information criterion, a hierarchical modeling generalization of the Akaike information criterion (AIC). It is especially useful in Bayesian model selection, and generally, the smaller the value of the deviation information criterion DIC, the better the model. WAIC is short for Widely Applicable (or Watanabe-Akaike) Information Criterion (Watanabe and Opper 2010), as DIC, the smaller the WAIC value, the better the model. pD and $pWAIC$ values are the effective number of parameters which are used to measure the model complexity. Compared to the original BYM model, smaller DIC and WAIC values and smaller pD and pWAIC values are all observed in the ecological model. Considering both the goodness of fit and the level of model complexity suggests that the observed data more strongly supports the ecological model. This reaffirms the necessity of including covariates (explanatory variables). However, the differences in the DIC and WAIC values between the two models are less than 5, and the difference in the goodness of fit between them is also insignificant.

When we incorporate another covariate, the Urban/Rural indicator, into the model, which is ecological model 2, both DIC and WAIC values decrease significantly. Compared to the original model, the difference is close to 5, indicating that the goodness of fit is significant. So, incorporating Urban/Rural is essential. Since the number of covariates added in this model is small, they possess relatively pronounced spatial characteristics, especially the SDI and Dairy covariates, which exhibit apparent and correlated spatial features. We need to be cautious of the decline in model performance due to spatial confounding. Therefore, we attempt to do a variable selection on the covariates in the

model. These are potential models:

For single covariate

Model 1: SDI

Model 2: Dairy Cattle Density

Model 3: Sheep Density

Model 4: Urban/Rural Indicator

Due to the significant influence of Urban/Rural, it seems explain most of the spatial patterns. Therefore, we consider the following models with two covariates:

Model 5: SDI+ Urban/Rural Indicator

Model 6: Dairy Cattle Density + Urban/Rural Indicator

Model 7: Sheep Density + Urban/Rural Indicator

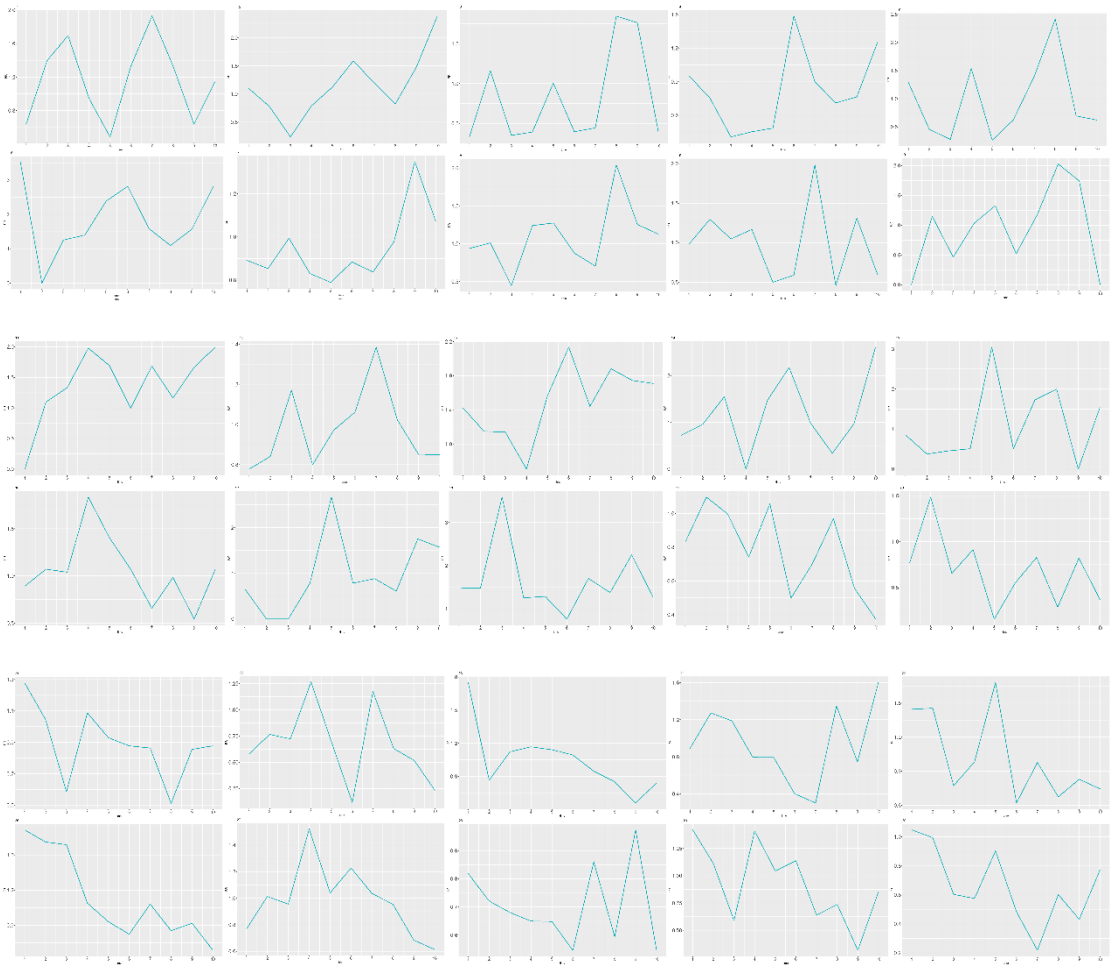
Table 4.3 A Comparison of the model 1-7 using DIC, and WAIC

Model	$\bar{D}$	pD	DIC	$WAIC$	$pWAIC$
Model 1	233.72	23	256.72	258.04	18.36
Model 2	234.6	22.80	257.40	258.98	18.47
Model 3	232.45	19.48	251.93	252.79	15.98
Model 4	233.8	21.98	255.78	257.54	17.99
Model 5	233.18	22.78	255.96	257.47	18.29
Model 6	233.37	22.73	256.10	257.50	18.20
Model 7	231.65	20.17	251.82	252.98	16.26

We know that the full model (ecological model 2 (with four covariates)) has a significant performance improvement compared to the original model. Among the models mentioned above, Model 3 and

Model 7 have the lowest DIC values. Compared to the full model, the DIC values of these two models all show a little decrease, indicating better model performance. This suggests that confounding in the covariates does affect model performance. Especially SDI, Urban/Rural indicator, and dairy cattle density. When our goal is to estimate the effects of covariates, or when the covariates vary over time, meaning they are not fixed in the study, we should be more cautious when adding covariates to the model.

4.3 Spatial-Temporal Modelling



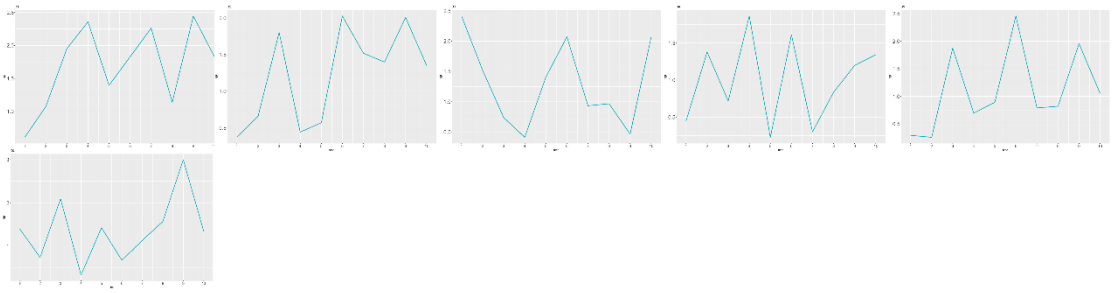
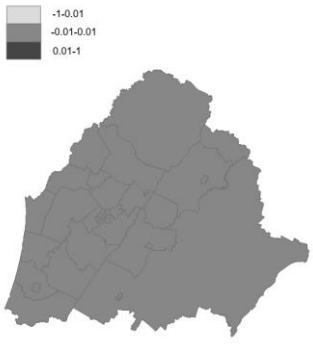


Figure 4.10 Time series plots of the RR (relative risk) of the 36 regions in Manawatu of 10 time periods.

It can be seen that each region exhibits different patterns of change, and the magnitude of change varies. Therefore, it is necessary to expand the original spatial model into a spatiotemporal model in order to achieve modelling at a higher temporal resolution.

We will use the four spatiotemporal interaction dependency modes (Knorr-Hold, 2000) to create spatiotemporal interaction variables $\boldsymbol{\delta_{ij}}$ and apply them to spatial-temporal modeling of Campylobacteriosis cases in the Manawatu region of New Zealand during the period Mar 2005-Sep,2016. We divide the entire research time into ten equal parts, and each period is about a year and a half long. The spatio-temporal interactions are plotted in the following figures. In this study, the temporal structure is set by RW2. From left to right, the first row, represents periods 1-5 respectively, and the second row represents periods 6-10 respectively. The function is shown in Eq 3.21.

Type I



.

Figure 4.11 Posterior mean of the spatio-temporal interaction $\boldsymbol{\delta_{ij}}$ for Campylobacteriosis cases at each period in the Manawatu under the type I

Type II

Figure 4.12 Posterior mean of the spatio-temporal interaction $\boldsymbol{\delta_{ij}}$ for Campylobacteriosis cases in the Manawatu under the type II

Type III

Figure 4.13 Posterior mean of the spatio-temporal interaction $\boldsymbol{\delta_{ij}}$ for Campylobacteriosis cases in the Manawatu under the type III

Type IV

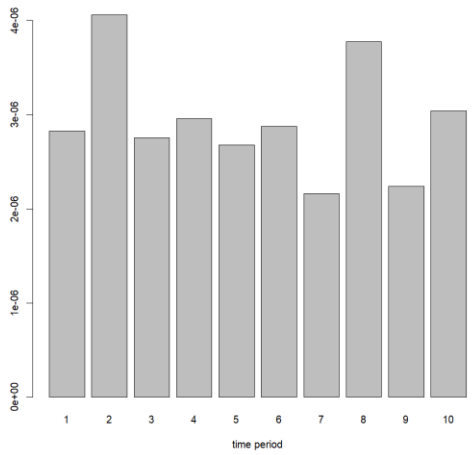
Figure 4.14 Posterior mean of the spatio-temporal interaction $\boldsymbol{\delta_{ij}}$ for Campylobacteriosis cases in

the Manawatu under the type IV.

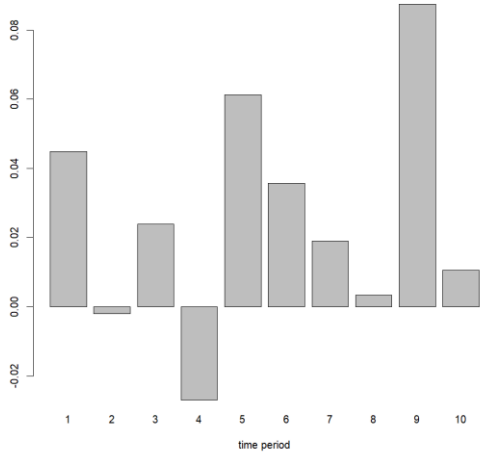
The model in pattern 1 has the least clear spatial-temporal pattern. We cannot observe any apparent differences, which clearly demonstrates that the assumption that different configurations of spatiotemporal parameters don't influence the determination of spatiotemporal dependency structures is unreasonable here. Compared to Pattern 1, Patterns 2, 3, and 4 have clear spatial patterns. Patterns 2 and 3 are very similar, while Pattern 4 is slightly different.

We calculate the mean of the above results by region and year according to four spatiotemporal interaction variables respectively.

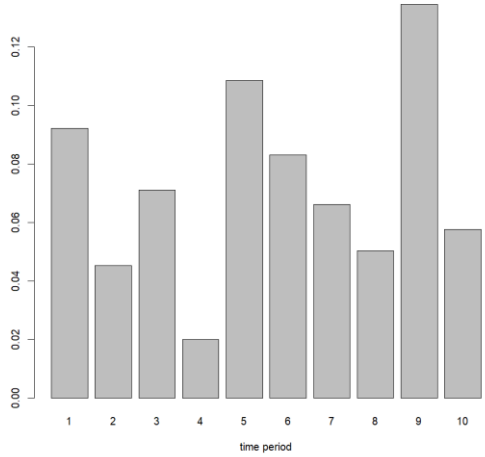
By year:



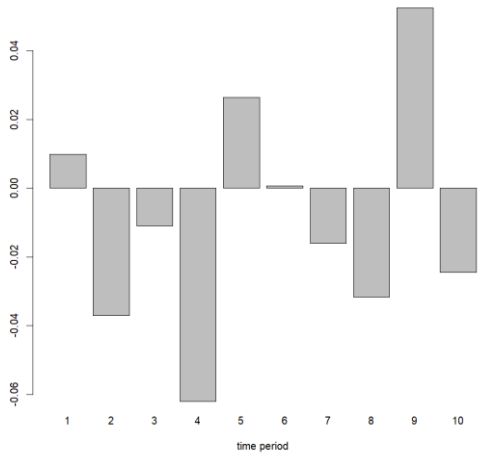
Type I



Type II



Type III



Type IV

Figure 4.15 Bar plots of the averaged spatio-temporal interaction by year.

By region:

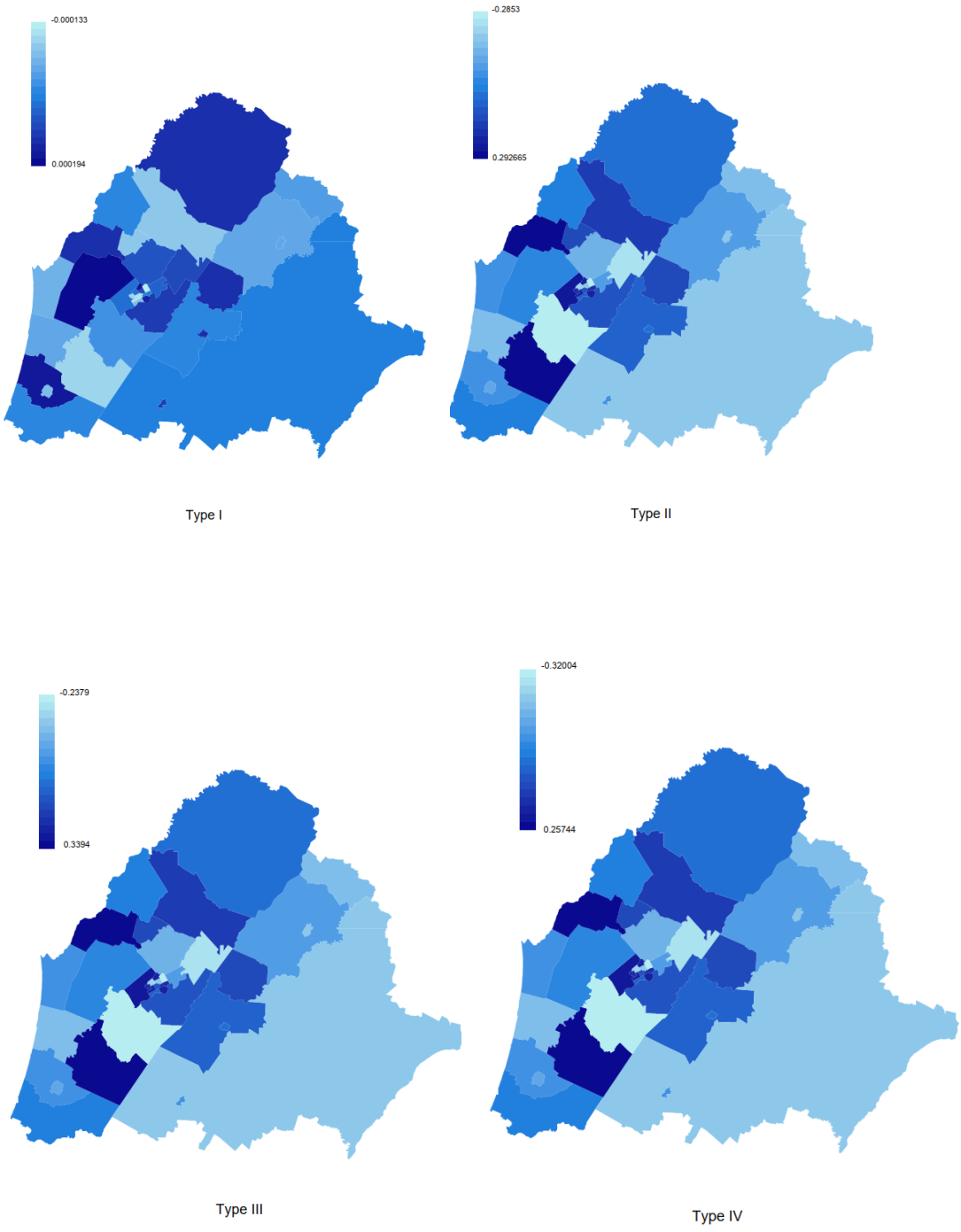


Figure 4.16 The averaged spatio-temporal interaction value by region.

From the years, it can be seen that except for type I, in types II, III, and IV, the spatiotemporal alternation variable reaches its peak in the 9th period and its lowest point in the 4th period. From the regions, it can be observed that the darker the colour, the higher the value of the spatiotemporal interaction variable. Except for type I, types II, III, and IV shared similar spatial patterns for the spatiotemporal interaction variables and only slight differences in values. We should pay more attention to these regions with high values, which means they have high relative risk.

Chapter 5
Spatial Point Process Models

In this chapter, we focus on introducing Bayesian statistical models applied to spatial point processes. Spatial point data can be classified into point pattern data and point reference data (Cressie,1993). In point pattern data, we introduced the log-Gaussian Cox model, which is used in spatial epidemiology to analyze the spatial distribution of case events. It is an extension of the Cox proportional hazards model, often used for survival analysis (Simpson et al., 2016). Point reference data could follow a continuous distribution, such as the Gaussian distribution. It can also follow a discrete distribution, such as the Poisson or binomial distribution. We first introduced the linear spatial geostatistical model, suitable for response variables mutually independent of Gaussian conditional distributions. Then, with the help of the generalized linear model, we expand the linear geostatistical models to the generalized linear geostatistical model, which is used to model independent non-Gaussian measurements at different locations. We extensively discussed the Poisson and binomial distributions, which are very practical in spatial epidemiological research. Finally, we introduced using SPDE methods to establish Gaussian random fields for continuous fields. SPDE methods represent Gaussian random fields as solutions to stochastic partial differential equations, allowing for flexible modeling of Gaussian random fields and simultaneously using INLA for Bayesian posterior inference, offering computational advantages. In the next chapter, we present a non-stationary linear geostatistical model in a case study involving the creation of a non-stationary Gaussian spatial field using the SPDE method. More details about the spatial point process in epidemiology can be found in Diggle (2013a), Diggle (2019), and Lawson et al. (2016). More details about the INLA-SPDE method can be found in Krainski et al. (2018).

5.1 Epidemiological Data with Specific Location

In recent years, there have been significant improvements in applying Bayesian models to small area (lattice) count data. This is partly due to the availability and accessibility of such data and its ready application (Lawson et al., 2016), as we described in the previous chapter. However, there has yet to be much progress in creating equivalent or similar models specifically for point pattern data. Point pattern data in epidemiology, also known as case event data, is found where a point location represents the residential address of a case of disease (Cressie, 1993). Conventionally, epidemiological studies typically focused on spatial risk factors at broader geographic levels, like comparing disease risk estimates across various countries or within administratively defined regions. If the information available for cases is limited to a postal code, zip code, or other arbitrary area, typically, only the disease count within that specific area can be analyzed. This is because, in general, the scale of postal or census regions is relatively large, and the point-referenced or point pattern data are found where studies focus on relatively small spatial windows (Diggle, 2013).

However, with the emergence of more precise postal code systems and the incorporation of postal code details regarding the place of birthplace, residence, or death into disease registries and census data, it became possible to account for much more detailed spatial variations patterns in disease risk (Diggle, 2013). For example, in the UK, they have Lower Super Output Areas (LSOAs), which are geographic hierarchies designed by the Office for National Statistics (ONS) to improve the reporting of small area statistics in England and Wales. In urban areas, the code system is theoretically precise enough to designate a specific street (Pearce et al., 2007). Therefore, individual studies focusing on case-event data and applying spatial point process concepts in epidemiological data analysis have been carried out, particularly in examining the observed disease pattern concerning potential environmental risks. However, it still faces some challenges, including the following: 1. Assigning individuals to a specific location is just a convenient mathematical assumption. 2. When ecological regression involving covariates is performed, covariates with the same high spatial resolution cannot be obtained, because demographic factors are only accessible at a broader geographic level, such as administrative regions. (Diggle, 2014).

In addition to the point pattern data applied to individual studies, there is also a kind of count data with specific location, which is the point reference data. Recalling how Cressie (1993) defined point pattern data and point-referenced (geostatistical) data, as we introduced earlier in Chapter Three. Both have specific geographic locations, and the difference lies in their observed attribution. We use $Z(s)$ which denotes the attribution that we observe at generic locations s. In point referenced data (Geostatistical data), $Z(s)$ could be any kind of observed data, in each site s_i, $Z(s_i)$ could be a continuous distribution, such as a Gaussian distribution. It can also be a discrete distribution, like the Poisson distribution, suitable for rare events, and binomial distribution is available for non-rare events. They are practical in epidemiological research. However, in point pattern data, $Z(s)$ takes a simpler form and usually the value equal to one, which means the occurrence of random case or event. Case event data has higher spatial resolution than count data; simply put, count data are a spatially aggregated form of case event data.

In Bayesian modeling of count data, defining spatial contiguity by employing a spatial weight matrix is often necessary. However, when point locations are available, an entirely different form of analysis method must be performed (Diggle, 2019). So, in this chapter, we will introduce the Bayesian approaches for spatial epidemic modeling methods based on point pattern data and point-referenced data, respectively.

5.2 Point Pattern Data

5.2.1 Frequentist Approaches for Point Pattern Data

Here briefly introduce a frequentist method for point pattern data analysis, the most representative of which is the HEPP model. Heterogeneous Poisson Process (HEPP) techniques proposed by Lawson (1993) provide a way of modeling the spatial distribution of point events. Based on the Epidemiological research, the HEPP model is restricted to three assumptions (Pfeiffer et al., 2008):

1. Case events are distinctive as they happen as single, spatially separate events.
2. Individuals in a defined study population exhibit independent behavior regarding their disease susceptibility.

3. From which cases emerge, the population at risk demonstrates a continuous spatial distribution.

Point pattern data models aim to measure how a group of explanatory factors impacts the presence of events across a study area, considering the spatial interdependence involved. Usually, these methods have been used in scenarios involving potential hazard sources, where the count of disease occurrences within a specified period and the spatial distribution of cases concerning a presumed hazard are examined in a study area (Lawson, 2018).

5.2.1.1 Computational Method

"splancs" (Bivand et al., 2017) from the R library in software R provides computation for the hybrid model of Lawson and Williams (1994) and the binary regression model of Diggle and Rowlingson (1994) by direct maximum likelihood methods. SpatStat from the R library in software R provides computation for a variety of point process models (Baddeley and Turner, 2000, 2003) by a novel integral approximation method proposed by Berman and Turner (1992), which involves using the Dirichlet tile areas or associated Delauney triangle areas of data points as weights in the approximation. The computation of most of these kinds of likelihood parameter estimations is done by using MCMC.

5.2.2 Bayesian Approaches for Point Pattern Data

The Bayesian approach can also be applied in point pattern process modelling. There is an approach to apply geostatistical modelling methods to point pattern data. As discussed earlier, the definitions of these two types of spatial data are very similar, with the only distinction being that point pattern data has a simpler observed attribution (Cressie, 1993). Therefore, applying geostatistical modelling methods (to be introduced in the next chapter) to point pattern data is also natural. We will also discuss the Log-Gaussian Cox Processes (Møller, Syversveen, and Waagepetersen, 1998), a commonly used Bayesian model for analyzing spatial point pattern data.

5.2.2.1 Log-Gaussian Cox Process Approaches

The log-Cox point process model refers to the log intensity of a Cox process, which is characterized by a Gaussian linear predictor. When the log-Cox process is specifically modeled with a Gaussian linear predictor, it is recognized as a log-Gaussian Cox process (LGCP) (Møller et al. in 1998). A Cox process is just a name for a Poisson process with varying intensity, and it is still followed by a Poisson likelihood. It is a hierarchical combination of a Poisson process at the first level and a Gaussian Process at the second level.

The Cox process is a Poisson process with intensity λ(s) that varies in space. Given some area A, the probability of observing a certain number of points in that area follows a Poisson distribution with intensity:

$$\lambda_A = \int_A \lambda(\mathrm{s}) ds \tag{5.1}$$

The intensity of the LGCP could be modeled as:

$$\log\big(\lambda(\mathbf{s})\big) = a + \mathbf{X}(\mathbf{s})\beta + S(\mathbf{s}) \tag{5.2}$$

Where a serves as the intercept and can be seen as a global average value, like the basic intensity on the spatial process. $S(\mathbf{s})$ a spatially structured random effect. $\mathbf{X}(\mathbf{s})$ is the covariates vector at position s and β is the regression coefficient vector.

An alternative Cox process model is short-noise cox process (SNCPs) (Møller, 2003). $S(\mathbf{s})$ which is a Gaussian spatial process with Matérn covariance and zero mean like:

$$\mathrm{Cov}[S(s), S(s+h)] = \sigma^2 \text{ Matérn } \left(\frac{|h|}{\phi}, \nu\right) \tag{5.3}$$

where $\phi, and\ \nu$ are the range and roughness parameters respectively. |h| is the Euclidean distance between two points on the ground. In SNCPs, $(\mathbf{s})$ is the sum of basic functions with random centers. This is a rich class of Cox process models which includes Neyman-Scoot process, Poisson-Gamma process, and short noise Cox process as special case (Møller, 2003).

Discrete spatial models are typically constructed using aggregated data, assuming a constant risk within census areas and overlooking the specific locations of individual cases. One of the principles of disease modeling is that when particular locations are known, the use of aggregated data should be avoided. The log-Gaussian process is the natural extension of the Besag-York-Molli´e (BYM) model, commonly employed in discrete spatial modeling. Therefore, when precise geographic locations are available, modeling with LGCP becomes more natural. Considering this characteristic, a series of articles have discussed this approach. Li et al. (2012) presented an example of modeling rare disease data by the log-Gaussian Cox model in the scenario of the population census areas that keep changing during the study period. Diggle et al. (2019) described the log-Gaussian Cox processes (LGCPs) as frameworks for representing spatial and spatiotemporal point process data. They modeled a set of data using discrete spatial models and LGCP models, respectively. Konstantinoudis et al. (2020) employed actual residential locations to mimic the incidence rate and found that, in terms of recovering risk surfaces and identifying high-risk areas, LGCPs models outperformed BYM models.

5.2.2.2 Computational Method

Log-Gaussian Cox processes fit naturally within the hierarchical modelling framework and are latent Gaussian models. So, it is natural to perform the computation of LGCP by integrated nested Laplace approximation (INLA) for fast approximate inference. Simpson et al. (2016) proposed a method in their paper, "Going off-grid: computationally efficient inference for log-Gaussian Cox processes," which directly utilizes the Stochastic Partial Differential Equation (SPDE) for modeling log-Gaussian Cox processes. With the ability to define a mesh flexibly, observations could be represented by considering their precise locations rather than grouping them into cells. Meanwhile, this approach could consider direct approximation with a good theoretical inference. The modelling could be done

by R package “inlabru” (Yuan et al. 2017).

5.3 Point-Referenced (Geostatistical) Model

5.3.1 The linear Geostatistical Model

Each point within a study area could be considered as a mathematical function for set of individual areas. The model for the response variable observed at each point can vary depending on its characteristics (Pfeiffer et al., 2008). Meanwhile, the purpose of modeling is typically to describe data and explain the variance present in the model. Usually, we choose mixture models for modeling, which include both fixed variables and random variables. In spatial modeling, Gaussian random fields are commonly added to the model as random variables with spatial structure. Consider the simplest and most mainstream linear regression model:

$$Y_i = \alpha + \beta d(x_i) + S(x_i) + U_i \; i = 1, \dots, n \tag{5.4}$$

Where:

Y_i is the i-th response variable at location x_i.

α is the overall intercept.

$d(x_i)$ is the value of explanatory variable associated with location x_i ,and β is the corresponding coefficient.

$S(x_i)$ and U_i are spatially structured and unstructured random effects respectively. For $S(x_i)$, it is Gaussian random fields. For U_i, it is mutually independent random variable followed by a zero mean Gaussian distribution with the variance σ^2.

For a mixed effects model, what we need to do is specify the joint probability distribution of the random variables. In the linear geostatistical model, it could be obtained by assuming conditional on location X, and the corresponding variable Y_i are mutually independent with Gaussian conditional distributions and the multivariate Normal distribution is fully characterized by its mean vector and

covariance matrix. In the model introduction stage, we uniformly consider using the simplest stationary and isotropic Gaussian random field for $S(x_i)$. At the same time, we tend to assume that the mean of each spatial random variable $S(x_i)$ is 0, because we are using a mixture model, and non-zero mean is contributed by the fixed effects in the model. So, by defining the covariance function is sufficient to fully specify any Gaussian process. So, The LGM (linear geostatistical model) obtained conditional on location S with variable Y_i are mutually independent with Gaussian conditional distribution could be rewritten as (Lawson et al., 2016; Pfeiffer et al., 2008):

$$\begin{aligned} Y_i \mid S(x_i), \theta &\sim \mathrm{N}[u(s_i), \tau^2] \\ u(s_i) &= \beta d(s_i)^T + S(x_i) + U_i \\ \mathrm{cov}\big[S(x_i), S\big(x_j\big)\big] &= \sigma^2 \rho(u; \phi) \end{aligned} \tag{5.5}$$

We already know that the variance of the spatially structured component S, a stationary and isotropic Gaussian random field, is only defined by the distance u. Additionally, we need to discuss the variance of the spatially unstructured component U_i, which follows a zero-mean Gaussian distribution. Due to the uncertainty about the unstructured component, they are often assigned an exchangeable prior during the modeling phase.

First, the residual component $S(x_i) + U_i$ captures the variation in Y_i which is not explained by the explanatory variables $X(s_i)^T$. The variation in residuals is split into two parts: one displaying spatial correlation, while the other does not exhibit such correlation. When the explanatory variables we add lack spatial structure, the spatial structure of the response variable Y is entirely captured by $S(x_i)$. Conversely, when all covariates explain the spatial structure of the response variable Y, the spatial correlation $S(x_i)$ appears as a redundant component. Next, let's explain the random variable U_i, people refer to this unstructured random variable as nugget effect and provides two explanations for its uncertain (Diggle, 2019; Cressie,1993), either as spatial variation at a level smaller than the shortest observed distance or as intrinsic random variation caused by measurement error. Intrinsic random variation caused by measurement error is easy to understand. It is a kind of inevitable error. In terms of spatial variation, describing spatial variation on a scale smaller than the minimum

observed distance be clarified by defining U as a spatial process $U(x)$, so

$$\mathrm{Cov}\{U(x), U(x')\} = \begin{cases} \tau^2 & : \|x - x'\| = 0 \\ \tau^2 \alpha \delta(u) & : \|x - x'\| > 0 \end{cases} \tag{5.6}$$

Where $\delta(u)$ is a correlation function, and $0 < \alpha < 1$. u is the Euclidean distance between locations x and x', and

$$\delta(u) = 0 \text{ When u} \geq u_0$$

If the smallest distance between any two data points location x_i and x_j is larger or equal to u_0, the correlation between any two values $U(x_i)$ and $U(x_j)$ is zero. So, when Z is uncorrelated Gaussian noise with no spatial structure, the likelihood of this model would not change. It can also be understood that at this point, due to a sufficiently large spatial distance, the variation is captured by the Gaussian random field $S(x_i)$. As Diggle (2019) said, it's impossible to identify the parameter α and the correlation function unless the data includes pairs of sampling locations whose Euclidean distance u is smaller than u_0 or multiple separate measurements at the same location. This is called nugget effect, which indicates the variable being analyzed has short scale variability that is random and unpredictable (Stein, 2012; Diggle, 2019; Cressie,1993). In traditional geostatistics, understanding the nugget effect is crucial for determining the optimal sampling points (Stein, 2012; Cressie,1993). Therefore, U_i is involved in the model to capture the unstructured variation which is caused by the effects of unmeasured explanatory variables that either have no spatial structure or have spatial structure but could not be identified because of the nugget effect (Diggle, 2019).

5.3.2 Generalized Linear Model

The Generalized Linear Model (GLM) was initially introduced and developed by John Nelder and Robert Wedderburn in the 1972 (Nelder and Wedderburn, 1972). GLM extends traditional linear regression models by utilizing a linear predictor (composed of a linear combination of predictors) and mapping this linear predictor to the range of the response variable through a link function, providing a framework for modeling and analyzing data where the distribution of the response variable could

be different probability distributions, such as Poisson distribution, binomial distribution, etc., and does not need to follow a normal distribution. These types of distribution is useful in epidemic modeling.

A regression that has a binary response variable is one of many generalized linear models and is called a logistic regression or a logit model. Recall a generalized linear model is made of a linear predictor and link function:

$$g(\mu_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{5.7}$$

The left $g(.)$ is link function, and the right is linear component.

We first move the probabilities μ_i to the odds. And then take logarithms, calculating the logit or log-odds, we get:

For Binomial:

$$g(\mu_i) = logit(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \tag{5.8}$$

For Poisson:

$$g(\mu_i) = \log(\mu_i) \tag{5.9}$$

So, the generalized linear model for Binomial and Poisson distribution could be written like:

$$\log\left\{\frac{p_i}{1-p_i}\right\} = d(x_i)$$
$$Y \sim Binomial(\mathrm{n}, p_i) \tag{5.10}$$

$$\log\{\lambda_i\} = d(x_i)$$
$$Y_i \sim Poisson(\lambda_i) \tag{5.11}$$

5.3.3 Generalized Linear Geostatistical Model

We have obtained binomial and Poisson link functions in generalized linear models. Naturally, we can apply them to replace the identity link function of the linear Geostatistical Model, then we get generalized linear geostatistical models with Possion and binomial likelihood.

1) Binomial likelihood

In epidemiology, the binomial distribution is often used to model the outcome of events with a fixed probability (e.g., infection rate) following exposure or transmission in a population (e.g., number of infected individuals).

Diggle and Giorgi (2019) illustrate a common scenario, in low-to-middle-income-countries (LMICs) where disease registries are either absent or lacking in geographic completeness, household surveys are a practical and valuable method for tracking the impact of infectious diseases. The format of the data sets produced in such studies can be represented as follows: x represents the position of a selected community or village within the population being studied. n_i represents the count of individuals who are chosen and examined to ascertain the presence or absence of the disease under the survey. Y_i represents the number of people whose result of testing are positive. A suitable model for the distribution of Y_i getting from sampling is a Binomial distribution with n_i attempts with a positive test probability $p(x_i)$. The Y_i form a collection of independent Bernoulli trials and random variables are a binary variable with only two values: 1 or 0. $Y_i = 1$ when the i-th individual tests positive for the disease, and 0 for negative.

$$\log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} = d(x_i)^\top \beta + S(x_i) + U_i \quad (5.12)$$

$$i = 1, \dots, n$$

2) Poisson likelihood

In epidemiology, the Poisson distribution is commonly used to describe the distribution of rare events (e.g., disease incidence) within a given time or space unit, such as describing the occurrence rate of a disease in a specific area. The Poisson linear geostatistical model assumes that the random variables Y_i referenced to the location x_i represent a collection of independent Poisson random variables.

$$\log\{\lambda(x_i)\} = d(x_i)^\top \beta + S(x_i) + U_i \quad (5.13)$$

$$i = 1, \dots, n$$

5.4 INLA (Integrated Nested Laplace Approximation) with SPDE (Stochastic partial differential equation) for geostatistics modelling.

Over the past decade, one of the most significant advancements in spatial statistics is the stochastic partial differential equation (SPDE) method applied to Gaussian fields (GFs). The SPDE method in geostatistics modelling is proposed by Lindgren et al. (2011). The main idea is to represent GFs with the Matérn covariance function as finite element solutions to SPDEs by an explicit link between GFs and GMRFs for r any triangulation of $\mathbb{R}^d$. This method reduces the computational cost of Bayesian inference and sampling by a Gaussian Markov random field (GMRF) approximation. The explicit construction of the GMRF representation involves utilizing a particular stochastic partial differential equation (SPDE). This SPDE equation has GFs with Matérn covariance functions as solutions when driven by Gaussian white noise. Approximation of the solution of the SPDE using a finite element method (FEM), a numerical tool for solving partial differential equations. This method yields a computationally efficient solution for inference by substituting a GRF with a dense covariance matrix with a GMRF that operates with a sparse precision matrix. Based on this method, it is easy to extant

a non-stationary model, which is for the covariates to be added to the model. Moreover, it can create separable and non-separable spatial–temporal interaction structures to achieve spatial-temporal modeling (Lindgren et al., 2022). More details could be found at Lindgren et al (2011) and Blangiardo, M., & Cameletti, M. (2015).

5.4.1 SPDE

The Matérn covariance function is commonly used in diverse scientific domains, especially as a widely used spatial covariance function. However, the key connection we will discuss is that a Gaussian random field (GFs) featuring the Matérn covariance serves as a solution to the linear fractional stochastic partial differential equation (SPDE). This is the work of Whittle (1954, 1963), which is also the original idea of SPDE comes from:

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}(\tau X(s) = \mathcal{W}(s) \tag{5.14}$$

It a Gaussian random field with Matérn covariance function, where $s \in R^d$, τ controls the variance, k is a positive scale parameter,

Δ is the Laplacian, where $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$.

$\mathcal{W}(s)$ is a Gaussian spatial white noise process.

$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}$ is a pseudodifferential operator, using Fourier transform definition of the fractional Laplacian in R^{d}, it is like:

$$\left\{\mathcal{F}(\kappa^2 - \Delta)^{\frac{\alpha}{2}}\phi\right\}(\mathbf{k}) = (\kappa^2 + \| \mathbf{k} \|^2)^{\frac{\alpha}{2}}(\mathcal{F}\phi)(\mathbf{k}) \tag{5.15}$$

And the marginal variance is:

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{\frac{d}{2}}\kappa^{2\nu}} \tag{5.16}$$

We shall name any solution to the equation above a Matérn field, and The Matérn fields are the only stationary solutions to the SPDE. It's worth highlighting that as κ or v approaches 0, the solutions to the SPDE do not exhibit Matérn covariance functions. Despite this, the SPDE still admits solutions when κ or v equals to 0, resulting in well-defined random measures. More detail about this can be seen in Appendix C.3 of Lindgren, (2011). The implicit assumption here involves having suitable boundary conditions for the SPDE. This is because, for $\alpha \geq 2$, the null space of the differential operator is non-trivial. The solution of this SPDE is the stationary GFs with Matérn covariance function given by

$$C(\boldsymbol{u}, \boldsymbol{v}) = \frac{\sigma^2}{2^{\lambda-1}\Gamma(\nu)} (\kappa \parallel \boldsymbol{u} - \boldsymbol{v} \parallel)^{\lambda} K_{\lambda}(\kappa \parallel \boldsymbol{u} - \boldsymbol{v} \parallel) \tag{5.17}$$

where $\|\boldsymbol{u} - \boldsymbol{v}\|$ is the spatial Euclidean distance between two generic locations u, and v $\in R^d$.
σ^2 is the marginal variance.
K_{λ} denotes the modified Bessel function of the second kind and order $\lambda > 0$ which is used to describe the degree of smoothness.

The relationship between the model smoothness parameter λ and the marginal variance σ^2 is defined by here when for R^d d=2 which is a two-dimensional plane.

$$\begin{cases} \lambda = \alpha - 1 \\ \sigma^2 = \dfrac{\Gamma(\lambda)}{\Gamma(\alpha)(4\pi)\kappa^{2\lambda}\tau^2} \end{cases} \tag{5.18}$$

In R-INLA the default value for the smoothness parameter is α = 2, back to the marginal variance function, d is equal to 2 and corresponding smoothness parameter λ equal to 1. The default representation for the SPDE parameters internally is $\log(\tau) = \theta_1$ and $\log(\kappa) = \theta_2$, where θ_1 and θ_2 are assigned a joint Normal independent distribution. More detail about prior specification can be

found in Lindgren et al. (2011). r is the range of the Matérn covariance function, which is defined by the smoothness parameter and the scale parameter r:

$$r(s) = \frac{\sqrt{8\lambda}}{\kappa} \tag{5.19}$$

and the covariance which is valid in stationary case is:

$$\sigma^2(s) = \frac{1}{4\pi\kappa(s)^2\tau(s)^2} \tag{5.20}$$

Afterward, given an SPDE, Lindgren et al. (2011) show how to examine the approach to estimate the solution of the stochastic partial differential equation (SPDE) presented in the equation by employing a finite element method (FEM). This numerical technique relies on a basis function representation defined over a triangulated grid in three dimensions, serving as a computational tool for addressing partial differential equations in the given domain. Here is the approximation way for the stationary and isotropic Matérn GF $\xi(s)$ which represented as the solution of the SPDE:

$$\xi(s) = \sum_{g=1}^{G} \varphi_g(s)\tilde{\xi}_g \tag{5.21}$$

G is the total number of vertices (nodes) of the triangulation (non-intersecting triangles).

$\tilde{\xi}_g$ weights vector distributed by zero mean Gaussian.

φ_g is the set of basic functions defined on a triangulation of the domain. To obtain a Markov structure, the basic functions are chosen to have a local support and to be piecewise linear in each triangle.

Using Neumann boundary conditions, the precision matrix Q for the weights vector $\tilde{\xi}_g$ distributed by zero mean Gaussian is given by:

$$\boldsymbol{Q} = \tau^2(\kappa^4\boldsymbol{C} + 2\kappa^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}) \tag{5.22}$$

Where:

C is diagonal matrix $C_{ii} = \int \varphi_g(s) d(\boldsymbol{s})$, and G is sparse positive semi-definite matrix with $\boldsymbol{G}_{ij} = \int \nabla\varphi_g(s)\varphi_g(s) d(\boldsymbol{s})$.

The precision matrix Q here depends on the parameters τ and κ.

The approximated solution to the SPDE $\xi(s)$ is a GMRF followed by zero mean gaussian distribution with precision matrix Q.

5.4.2 Non-stationary SPDE

Just as we mentioned earlier, there is a demand for modeling non-stationary models to deal with more complex situations. A particularly remarkable expansion within the SPDE framework is its capacity to model non-stationarity. Numerous practical applications demand the incorporation of non-stationarity in the correlation function. One significant benefit of employing the SPDE approach lies in its adaptability. Instead of relying on a covariance function, the characteristics of the random field are described by an SPDE. This flexibility allows us to adjust the SPDE itself, rather than the covariance function, facilitating the creation of Gaussian Random Fields (GRFs) with dependence structures different from the stationary Matérn covariance. Another benefit is that the resulting non-stationary Gaussian field becomes a Gaussian Markov random field (GMRF), which benefits more flexibility, structure, and computational efficiency.

In geostatistical models, we typically tend to associate the variation in parameters, specifically those related to non-stationarity, with spatial coordinates or positions. This way, we can incorporate demographic variables that vary by region into the model. In the previous equation stationary SPDE, the scale parameter k and the error variance τ are constant in space (Eq. 5.19). Non-stationarity is achieved when one or both parameters are not constant. This is a crucial part of implementing non-stationary models. Generally, we have the flexibility to enable both parameters to depend on the coordinate, which represents the continuous spatial field 's'. Therefore, we express Eq. 5.17 in this context (Lindgren et al. 2011).

$$(\kappa^2(s) - \Delta)^{\frac{\alpha}{2}}\big(\tau(s)X(s)\big) = \mathcal{W}(s)$$

(5.23)

Especially noteworthy is when their variation with 's' occurs gradually. In a low dimensional representation, it is like this:

$$\log\{\kappa^2(\mathbf{s})\} = \sum_i \beta_i^{(\kappa^2)} B_i^{(\kappa^2)}(\mathbf{s}),$$

$$\log\{\tau(\mathbf{s})\} = \sum_i \beta_i^{(\tau)} B_i^{(\tau)}(\mathbf{s}) \qquad (5.24)$$

Where:

$B_i^{(.)}(\mathbf{s})$ is the deterministic spatial smoothness basis functions over the study region.

By incorporating slowly varying parameters $\kappa^2(\mathbf{s})$ and $\tau(\mathbf{s})$, the fundamental concept of equation as a Matérn field remains consistent. However, the specific form of the non-stationary correlation function achieved in a non-constant state is still being determined. This is because The SPDE automatically handles the task of 'combining all local Matérn fields into a consistent global field (Lindgren,2011). Technically, introducing space-dependent τ and κ alters only the components of the precision matrix Q in the Markov representation of the Gaussian Random Field (GRF). Rewrite the formular 5.22, It should look like:

$$Q = T(K^2CK^2 + K^2G + GK^2 + GC^{-1}G)T \qquad 5.25$$

Where $\boldsymbol{T} = \mathrm{diag}\big(\tau(\boldsymbol{s}_i)\big)$, and $\boldsymbol{K} = \mathrm{diag}\big(\kappa(\boldsymbol{s}_i)\big)$. $\boldsymbol{T}$ is a diagonal matrix with elements corresponding to τ at the locations $\boldsymbol{s}_i$; $\boldsymbol{K}$ is a diagonal matrix with elements corresponding to κ at the locations $\boldsymbol{s}_i$. Each $\boldsymbol{s}_i$ is a vertex in the mesh.

The relationship between the model parameters and the marginal variance given by (Eq 5.19) is valid in the stationary case but not in the non-stationary case. By disregarding the spatial interaction between the non-stationary parameter fields, we obtain nominal approximations. The relationship between the model parameters and the marginal variance, as outlined in equation above, holds true under stationary conditions but not in the non-stationary scenario. However, it is possible to achieve

rough approximate values for the marginal variance and range by ignoring the spatial correlation interaction among the non-stationary parameter fields. When $\alpha = 2$, which is by default setting in R-INLA, the approximation is like:

$$r(s) \approx \frac{\sqrt{8}}{\kappa}$$
$$\sigma^2(s) \approx \frac{1}{4\pi\kappa(s)^2\tau(s)^2} \tag{5.26}$$

These approximations are valid for slowly varying κ(s) and can be used for straightforward interpretation of the parameters.

5.4.3 Space-Time Inseparable Framework

The space-time inseparable framework of geostatistical spatial-temporal modelling is still achieved through an additive hierarchical model, as we talked about before, but it's worth noting that there are two categories of spatiotemporal interaction variables. They fall into two categories according to whether they are separable.

5.4.3.1 The First Category

As the spatial-temporal interaction structure in discrete spatial models was discussed earlier (Knorr-Held 2000), we could still create a space-time inseparable modeling framework for point-referenced data to be applied in hierarchical models. According to whether the spatial-temporal covariance function of this interaction structure could be separable or not, they could be divided into two categories. The first category is a separable space–time interaction structure, usually constructed as Kronecker products, resulting in separable structures, where the space-time covariance function is a product of a spatial and a temporal covariance function.

In this kind of separable structure model, as illustrated by Cameletti et al. (2011), they modeled the temporal dynamics as a spatially constant latent process. It can be expressed as the production of a distinctly spatial and a distinctly temporal covariance function:

$$S_{ij} = aS_{i(j-1)} + \xi_{ij}$$

$$|a| < 1\ , S_{i1} \sim \text{Normal}\left(0, \frac{{\sigma_e}^2}{(1-a^2)}\right) \tag{5.27}$$

ξ_{ij} is a zero-mean Gaussian field, assumed to be temporally independent and characterized by the following spatial-temporal covariance function.

$$\text{Cov}(\xi_{it1}, \xi_{jt2}) = \begin{cases} 0 & \text{if} \quad t1 \neq t2 \\ \text{Cov}(\xi_i, \xi_j) & \text{if} \quad t1 = t2 \end{cases} \tag{5.28}$$

Where:

$$\text{Cov}(\xi_i, \xi_l) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}}(\kappa\|s_i - s_l\|)^\lambda K_\lambda(\kappa\|s_i - s_l\|) \tag{5.29}$$

It is the solution to SPDE with Matèrn covariance function as we mentioned previously.

In addition to the covariance function defined above, Cameletti et al. (2011) also introduced another method:

$$\text{Cov}(\xi_{it1}, \xi_{jt2}) = \text{Cov}(|t1 - t2|)\text{Cov}(\|s_i - s_l\|) \tag{5.30}$$

Where:

$|t1 - t2|$, which is the temporal lag between time point t1 and t2.

The space-time correlation function factors into a purely temporal component and a purely spatial component.

5.4.3.2 The Second Category

The second category is inseparable (for covariance function) space–time interaction structure. Although it is challenging to construct non-separable non-stationary covariance functions explicitly, non-separable SPDE models can be obtained with relative ease using locally specified parameters. The simplest non-separable Stochastic Partial Differential Equation (SPDE) suitable for the Gaussian Markov Random Field (GMRF) method is the diffusion equation (Lindgren,2011).

$$\left\{\frac{\partial}{\partial t} + (\kappa^2 + \mathbf{m} \cdot \nabla - \nabla \cdot \mathbf{H}\nabla)\right\} x(\mathbf{s}, t) = \mathcal{E}(\mathbf{u}, t) \tag{5.31}$$

Where:

$x(\mathbf{u}, t)$ is Gaussian random field at location s and time t.

m is a transport direction vector

H is a positive definite diffusion matrix

$\mathcal{E}(\mathbf{u}, t)$ is a stochastic space–time noise field.

It is evident that even with this stationary formulation, non-separable fields are generated, as evidenced by the spatiotemporal power spectrum of the solution:

$$R_x(\mathbf{k}, \omega) = R_{\mathcal{E}}(\mathbf{k}, \omega)\{(\omega + \mathbf{m} \cdot \mathbf{k})^2 + (\kappa^2 + \mathbf{k} \cdot \mathbf{H}\mathbf{k})^2\}^{-1} \tag{5.32}$$

Where:

R is a kind of covariance function, and ω could be written as $\omega(s, t)$, which is space–time white noise. A Gaussian Markov Random Field (GMRF) representation can be derived by initially employing the standard spatial approach and subsequently discretizing the interconnected system of temporal stochastic differential equations, using methods such as Euler's method, for instance. Allowing all the parameters to vary with location in space (and possibly in time) generates a large class of non-separable nonstationary models (Lindgren, 2011).

More details about this method can be found in (Lindgren et al., 2020) and (Lindgren, 2011). They

released a new R package, "INLAspacetime," which is friendly for users to create inseparable spatial-temporal interaction based on the diffusion function. A non-separable space-time interaction can be established using the Diffusion-based Extension of the Matern Field (DEMF). This involves Gaussian Matern fields derived from a fractional and stochastic version of the physical diffusion equation originally studied by Whittle (1954, 1963). There are two kinds of models based on the different parameters of DEMF, and they are critical diffusion DEMF (1,2,1) and iterated diffusion DEMF (2,2,0).

Chapter 6
Case Study About the Aotearoa SARS-CoV-2 Wastewater Surveillance Programme

A case study, “A Bayesian spatio-temporal model for misaligned catchment wastewater concentrations data of SARS-CoV-2 for community level surveillance.”

A Bayesian Spatio-Temporal Model for Misaligned Catchment Wastewater Concentrations Data of SARS-CoV-2 for Community Level Surveillance

6.1 Introduction

Wastewater based epidemiology (WBE) monitoring, or sewer monitoring, refers to analyzing wastewater to identify biological or chemical substances of interest to public health monitoring (Phil M et al., 2018). Since the outbreak of COVID-19 in 2020, people have been exploring the possibility of using wastewater viral concentration to monitor and predict the COVID-19 epidemic. This is because COVID-19 is caused by SARS-CoV-2 (World Health Organization, 2021). The SARS-CoV-2 RNA could be detected in the wastewater through bodily fluids (such as feces and urine) and respiratory excretions from individuals who have contracted the COVID-19 virus, whether they are symptomatic or asymptomatic (Jones et al., 2020; Wolfel et al., 2020).WBE has been acknowledged as a cost-effective method for monitoring disease surveillance, implemented in many countries during the COVID-19 pandemic (Manuel et al., 2022). Meanwhile, compared to traditional community nucleic acid testing, WBE surveillance avoids its inherent selection bias (Li et al., 2023). This means that wastewater-based epidemiology (WBE) can be used as an effective tool to monitor COVID-19 infections at the community scale. The most important thing is that using wastewater data can help the public health system predict a surge of COVID-19 cases in advance, which makes it more significant (Hewitt et al., 2022; Medema et al., 2020). To assist public health responses to COVID-19, wastewater-based epidemiology (WBE) is being utilized internationally to monitor SARS-CoV-2 infections at the community level (Sims et al., 2020). Medema et al. (2020) from Kaderrichtlijn Water (KRW) in the Netherlands reported the detection of SARS-CoV-2 in sewage on February 20, 3 weeks before the first Dutch case was reported. They proved that, even when COVID-19 prevalence is low, viral RNA could be detected in sewage, and they show that monitoring the virus's spread in the population through wastewater surveillance is a sensitive approach and could serve as an early warning tool of (re)emergence of COVID-19 in cities. Later, others quantified the timeliness between

wastewater signals and traditional epidemiological indicators. Peccia et al. (2020) used a linear model to show that wastewater concentration can provide a warning for the potential outbreak of cases 4 to 7 days in advance. This demonstrated the effectiveness of wastewater as an early warning tool for outbreaks. Nelson et al. (2022) came up with similar results but also found different lags in time between wastewater loading and case reports across different areas. This may be related to the different socio-demographic characteristics across the areas.

In Aotearoa, in 2020, the SARS-CoV-2 wastewater surveillance program was launched by the Institute of Environmental Science and Research (ESR) to monitor and track potential cases of COVID-19. This has aided Aotearoa's ability to identify potential outbreaks and is now being used to track viral load in communities and monitor variants. Hewitt et al. (2022) estimated the total number of COVID-19 cases required for the detection of SARS-CoV-2 RNA in wastewater and showed that SARS-CoV-2 RNA detections at the wastewater treatment plants were associated with increasing COVID-19 cases in corresponding regions of New Zealand. They also showed that predictions can be made with wastewater even at low prevalence. Nicoll et al. (2022) used Auckland as an example to discuss the feasibility and challenges of using wastewater epidemiology to monitor SARS-CoV-2 and predict potential outbreaks at a community scale. Some of the significant challenges are population mobility and sampling design. Some people have done statistical modelling of wastewater SARS-CoV-2 concentrations for prediction. Regarding temporal modelling, Vaughan et al. (2023) use machine learning methods for time series forecasting the wastewater-based epidemiological data at the community level. The results show that using time series to predict the virus concentration of wastewater in the next period seems unreliable, and there are still many challenges. In terms of spatial-temporal modelling, Li et al. (2023) established a spatial-temporal model that predicts wastewater concentration at an acceptable spatial-temporal resolution covering the United Kingdom by R-INLA. They also studied the relationship between changes in the wastewater viral concentration and the prevalence rate, and they found that using weekly averaged data to assess leads or lags between the signals from wastewater and traditional epidemiological metrics is inappropriate and that daily reported data is more suitable for this. As nucleic acid tests became increasingly costly and hard to implement, simultaneously, reported cases were likely to become less reliable as a measure of actual case rate through time, and potential cases may be challenging to investigate, having a rigorous model

with an appropriate resolution of a known good source (wastewater) will become increasingly important. We propose a dynamic spatial-temporal statistical model of wastewater viral concentration that can be applied in Aotearoa, New Zealand. The purpose is to use the proposed spatial-temporal model to predict wastewater viral concentration, thus achieving community-level (Statistics Area level 2 (SA2): 1000-4000 people) surveillance and forecasting of the whole Aotearoa, New Zealand.

6.2 Data and Method

6.2.1 Data

The data we use in this study is from the SARS-CoV-2 wastewater surveillance program undertaken by Environmental Science and Research (ESR) in Aotearoa. The methods used for the identification of SARS-CoV-2 in wastewater are described in Hewitt et al (2022). In this wastewater surveillance program, a total of 134 sampling points have been used between June 2020 and August 2023. It is worth noting that new sampling sites may be added over time, and sampling may reduce in frequency or cease for other sites. Thus, the data are non-uniform in space and time, caused by irregular sampling patterns in locations and frequency. The location and sampling frequency of sampling points is shown in Figure 6.1. However, the sampling location is not where the wastewater is generated. Each sampling point represents a wider catchment corresponding to the wastewater network upstream of each sampling location. Part of the catchment areas are presented in Figure 6.2.

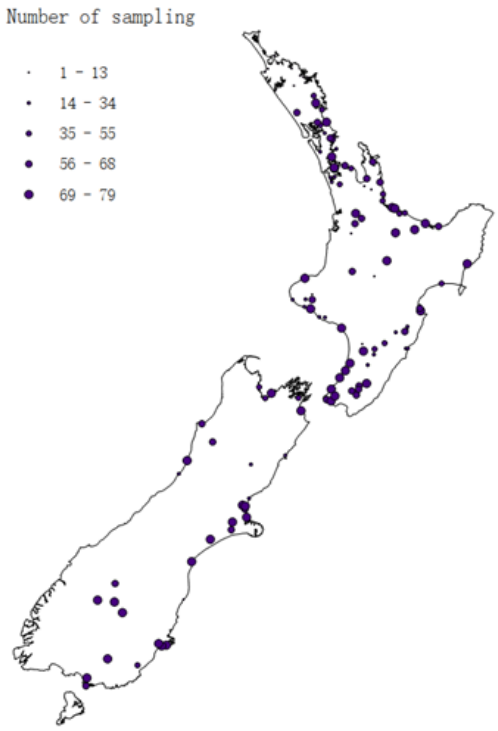


Figure 6.1: Number of samples during the study period at each sampling location.

Figure 6.2: Parts of the catchment areas polygons in Aotearoa wastewater surveillance program.

The wastewater sampling is irregular because of economic advantages. Wastewater sites are chosen considering various factors such as population and geographic coverage. The sampling frequency also varies at different catchment areas; typically, larger wastewater catchments are

sampled twice a week, while smaller towns and communities are sampled once a week, which is also subject to change. The temporal resolution of virus concentration in wastewater data is weekly averaged. Meanwhile, we also use the unit flow concentration, genetic copies per liter (GC/L). So, in this study, the virus concentration unit is weekly-averaged GC/L. The raw data contains many zeros, and most are concentrated at the beginning of the surveillance project, which was in the early stages of the epidemic. Incorporating sparse data from the epidemic's early stages into the model would be meaningless for future prediction, so we restrict our study period from "2022.02.27" to "2023.08.27". Zero has two meanings: first, the virus does not exist, and second, the virus concentration is below the lower detection limit. For ease of handling, any zero records in the study period are replaced with the lowest detection limitation which is 200 GC/L (Hewitt et al., 2022).

We also introduced some covariates into the spatiotemporal model. The basic spatial unit of spatial and spatio-temporal covariates we used is Statistical Area 2 (SA2), consisting of 1000-4000 individuals, for which a range of demographic information is available.
These covariates are:

- Log Population Density: from STAT NZ census 2018.
- Age structure variable: Age proportion below 65 and age proportion above 15 from STAT NZ census 2018.
- Ethnic proportion: Proportion of Non- European from STAT NZ census 2018.
- SDI: Socioeconomic Deprivation Indexes (SDI) 2018.

There is also a temporal variable:

- School terms: school terms and holiday terms for state and integrated schools from ministry of New Zealand.

The primary spatial unit of observation (wastewater catchments) typically spans more than one SA2 area, and thus, we recalculate the population density, age structure, and ethnic proportion of wastewater catchments based on the SA2 and the population grid. The SDI from SA2 was reaggregated to wastewater catchments and weighted by population where appropriate.

6.2.2 Modelling framework

Our aim is to achieve wastewater virus concentration surveillance at the community level and prediction at the catchment level, both from a national perspective using existing catchment wastewater SARS-CoV-2 concentration data. In the sampling scheme, only some catchment areas will be sampled at every point in time due to economic efficiency. Additionally, because of the concentrated distribution of population, catchment areas are not necessarily geographically contiguous. One of our objectives is to model the catchment areas within the wastewater sampling framework nationwide during the study period. Therefore, our main challenges will be dealing with misaligned data caused by irregular sampling and defining spatial correlation among non-contiguous polygons (catchment areas). We prefer Bayesian modeling methods for misaligned data caused by irregular sampling because they have a good tolerance for missing values in the data. For modelling discontinuous areas, several potential modelling frameworks could be utilized:

1. Traditional small area (lattice) studies.

When there are gaps between catchment areas, this makes traditional lattice small-area studies impractical. There is an adaptive CAR model that allows the spatial weight matrix W to be modelled as binary random variates within CAR formulation. They could be used for modelling risk for nonadjacent areas. Although they are not flexible in modelling locally varying spatial dependencies and lack interpretability of the final weight matrix W, they offer an adaptive spatial dependence parameterization process. Initially, these models were motivated for adjacency modelling and boundary detection and were shown to be effective (MacNab, 2022).

2. Point-referenced data modelling.

Another method involves converting polygons into points, thus transforming lattice data modelling into point-referenced data modelling. One drawback of this approach is the loss of information regarding the shape and size of polygons, namely the catchment areas in this study, during this

conversion process. However, the benefit is that linear geostatistical models can flexibly handle fixed and random variables (including spatiotemporal interaction variables) after converting them into points. Additionally, using SPDE, non-stationary and non-isotropic Gaussian random fields and high-precision grids can be created easily to represent the continuous spatial field.

Compared with the adaptive CAR model, we prefer to use a point-referenced model because we value the interpretability of the model more. In terms of the point-referenced data, we need to consider that our wastewater virus concentration model needs to be based on a nonstationary spatial model. Because wastewater virus concentrations may reveal true COVID-19 prevalence, traditional geostatistical stationary models cannot capture epidemics or outbreaks of infectious diseases. At the same time, we need a high time resolution, which is critical for infectious diseases. We are also interested in fixed effects, or covariates, that may explain the underlying level of risk that existed before the outbreak. For the above, SPDE proposed by Lindgren et al. (2011) is a good choice. First, the modelling can be done in a continuous space of a nonstationary environment. Second, a random field could be characterized by an SPDE rather than a covariance function, enabling us to modify the SPDE to obtain GRFs with the dependence structures we want. Since we need a high time resolution, we can let the dependence structure of the spatial field depend on the time index. In other words, we can create a spatiotemporal interaction variable to capture local departures at different time points. This spatial random variable is used to smooth factors such as population mobility and outbreaks in specific catchment areas. Third, a study on spatial field effects using this method using Bayesian posterior inference has a good tolerance for misaligned and sparse data. Even if there are a large number of missing values in the data, modelling can still be performed without interpolation, thus avoiding introducing extra biases. So, we will use Bayesian method for point-referenced modelling. To better illustrate the spatial location of the catchment (Figure 6.2), we utilize the population-weighted centroid of the corresponding catchment polygons to reflect where wastewater is generated. Since people produce viruses and populations are not uniformly distributed in the catchment areas, it is reasonable to use mathematical function such as population-weighted centroid to represent catchment polygons. This is important because the sampling points (Figure 6.1), that is, wastewater treatment plants, are usually in sparsely populated places far away from the city center, so the characteristics of these sampling areas are generally not representative of the ‘average’ attributes. The

size, shape, and population density of catchment areas vary, and doing so can reduce the bias caused by data transformation to some extent.

6.2.3 Bayesian statistical model based on the Stochastic Partial Differential Equations (SPDE)

The log-transformed number of weekly averaged genetic copies per liter at each catchment area i (i=1…134) and week t (j=1…) follows a Gaussian distribution. n is the number of catchment areas. We consider the model in this form:

$$X_{ij} \sim \text{Gaussian}\left(u_{ij}, \sigma_e\right)$$

$$u_{ij} = b_0 + \sum_{n=1}^{8} \beta_n m_{in} + S_{ij} + L_i + T_j$$

b_0 : Intercept which is the average base viral concentration gc/L of the whole country.

X_{ij}: Log GC/L at location i and time j.

β_n and m_{in}: Covariates and its coefficients based on n which is the number of catchment areas.

In term of three random effect:

L_i: random effect (IID likelihood) for catchments areas index.

T_j: random effect (IID likelihood) for time points index.

S_{ij}: The latent spatial-temporal process at catchment area i and time j which is unobserved level of wastewater viral concentration. Here we set the spatial filed which depend on the week followed by AR1. This is originally introduced by Cameletti et al. (2011)

$$S_{ij} = aS_{i(j-1)} + \xi_{ij}$$

$$|a| < 1\ , S_{i1} \sim \text{Normal}\left(0, {\sigma_e}^2/(1-a^2)\right)$$

ξ_{ij} is a zero-mean Gaussian field, assumed to be temporally independent and characterized by the

following spatial-temporal covariance function.

$$\text{Cov}\left(\xi_{it1}, \xi_{jt2}\right) = \begin{cases} 0 & \text{if} \quad t1 \neq t2 \\ \text{Cov}\left(\xi_i, \xi_j\right) & \text{if} \quad t1 = t2 \end{cases}$$

$$\text{Cov}\left(\xi_i, \xi_l\right) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}}(\kappa\|s_i - s_l\|)^{\lambda} K_{\lambda}(\kappa\|s_i - s_l\|).$$

$\|s_i - s_j\|$ is the Euclidean distance between two generic locations $\boldsymbol{s_i}, \boldsymbol{s_j} \in \mathbb{R}^d$ and σ^2 is the marginal variance. K_{λ} denotes the modified Bessel function of the second kind and order $\lambda > 0$ which measures the degree of smoothness of the process and is usually kept fixed due to poor identifiability. More details on the SPDE approach and SPDE with dependence structure are available at Lindgren et al. (2011) and Rikke et al. (2014)

6.2.4 Implementation and prior specification

All the models above were implemented using the R-INLA package (Rue et al., 2009) (available at "www.r-inla.org"). The INLA method for approximate Bayesian inference on latent Gaussian models has been shown to quickly and accurately estimate posterior margins.

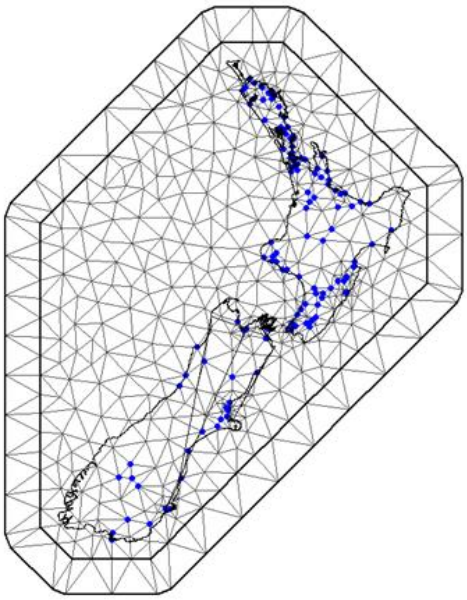


Figure 6.3: The mesh is created by R-INLA for model fitting, and the points are the population weighted centroid of the total 128 catchment areas in the sampling frame.

Like other situations where SPDE is applied, in complex scenarios, the solution to the equation is complicated to express using explicit formulas. Therefore, this method also resorts to numerical calculations. Generally, a grid is needed to approximate the definition space of partial differential equations. The finer the grid, the more accurate the solution obtained. But correspondingly, finer meshes require higher computational costs and larger storage space. The accuracy of the approximations depends only on the numerical integration scheme. New Zealand's South and North Islands are geographically isolated, but population movement is frequent and convenient, so they are not modelled separately or based on a non-isotropic Gaussian random field.

We set a penalised complexity prior (Fuglstad et al., 2019) on the spatial-temporal interaction part. That depends on the mesh. The a priori setting relies on experience and comparison with similar studies. According to the data, the wastewater virus concentration is between 5-15 GC/L. In term of standard deviation of the spatial field. The variation in the model may be caused by the covariates

and spatial variation. We believe that the standard deviation of the spatial effect should be less than 8， so we set sigma0 = 8 and Psigma = 0.05. As for spatial range of the random field, since we are modelling on a national scale, the distance between sampling points even in the same city (e.g., Auckland) is approximately 10 kilometers. From a practical point of view, it is advisable not to favor ranges that are smaller than the resolution of the mesh. We believe that correlation range in the Matèrn covariance is large than 10km, so we set sigma1 = 10 and Psigma=0.05. At the same time, we set the space field pattern followed by AR1 likelihood, and we set the prior of a in S_{Ij} equal to 0.8 or 0.9, and this will be discussed in next section. All covariates have been standardized, so give them a very weak prior by default. Error precision $1/\sigma_e$ is also give a weakly prior by log Gamma (1, 0.00005)

6.2.5 Model comparison and selection

When so many factors are included in the model as fixed effects, confounding will inevitably arise. However, random effects created based on SPDE exhibit high flexibility. Variable selection for fixed effects (covariates) does not necessarily lead to a significant improvement in the model. Therefore, we will not delve into extensive discussion regarding the selection of covariates here. Let's consider several possible models that contain the same fixed effects (covariates) but with different random effects and their prior parameters. The fitting and prediction results were each conducted ten times and then averaged.

$$X_{ij} \sim \text{Gaussian}\left(u_{ij}, \sigma_e\right)$$

Model 1: $u_{ij} = b_0 + \sum_{n=1}^{8} \beta_n m_{in} + S_{ij}$ with the prior of the AR1 parameter ρ=0.9

Model 2: $u_{ij} = b_0 + \sum_{n=1}^{8} \beta_n m_{in} + S_{ij}$ with the prior of the AR1 parameter ρ=0.8

Model 3: $u_{ij} = b_0 + \sum_{n=1}^{8} \beta_n m_{in} + S_{ij} + L_i + T_j$ with the prior of the AR1 parameter ρ=0.8

Model 4: $u_{ij} = b_0 + \sum_{n=1}^{8} \beta_n m_{in} + S_{ij} + T_j$ with the prior of the AR1 parameter ρ=0.8

Model 5: $u_{ij} = b_0 + \sum_{n=1}^{8} \beta_n m_{in} + S_{ij} + L_i$ with the prior of the AR1 parameter ρ=0.8

Fitting result

	Pierce correlation coefficient	Mean absolute percentage error （MAPE）	Root means square error (RMSE)
Model 1	0.796	0.091	1.057
Model 2	0.797	0.091	1.056
Model 3	0.866	0.077	0.882
Model 4	0.806	0.088	1.035
Model 5	0.836	0.084	0.961

Table 6.1: Summary of the Pierce correlation coefficient, Mean absolute percentage error （MAPE）, the Root means square error (RMSE) and the 95% coverage of the fitting by the five potential models

Prediction result

	Pierce correlation coefficient	Mean absolute percentage error (MAPE))	Root means square error (RMSE)
Model 1	0.516	0.099	1.023
Model 2	0.519	0.098	1.020
Model 3	0.559	0.163	1.469
Model 4	0.559	0.137	1.270
Model 5	0.519	0.101	1.041

Table 6.2: Summary of the Pierce correlation coefficient, Mean absolute percentage error (MAPE), the Root means square error (RMSE) and the 95% coverage of the prediction by the five potential models.

	Posterior mean and its 95% CI				
	Model 1	Model 2	Model 3	Model 4	Model 5
Residual precision	79.926 (76.660, 83.300)	79.854 (76.611, 83.249	103.856 (98.396, 109.292)	82.886 (79.418, 86.484)	92.862 (88.604, 97.54)
Correlation range (km) in the Matérn covariance	361.1166918 km (277.057, 457.194)	360.015 km (274.466, 455.073)	45.577 km (37.172, 55.203)	36.110 Km (29.150, 44.151)	262.997 km (206.426, 330.196)
Standard deviation in the Matérn covariance	0.745 (0.548, 0.979)	0.747 (0.534, 0.982)	0.080 (0.072, 0.087)	0.142 (0.128, 0.158)	0.124 (0.108, 0.142)
AR1 coefficient	0.9979518 (0.997, 0.999)	0.998 (0.996, 0.999)	0.828 (0.779, 0.870)	0.980 (0.974, 0.986)	0.908 (0.879, 0.934)

Table 6.3: Posterior estimates of parameter associated with the spatial temporal random effect for five potential models.

Naturally, incorporating two random variables into the model can improve its fitting. However, our intention is to better capture sudden outbreaks of prevalence (i.e., virus concentration in wastewater) at specific spatial locations or time points with the spatial random variable L_i and the temporal random variable T_j respectively. As the spatiotemporal alternating variables based on SPDE are highly flexible, the posterior variance of these two random variables, which follow an IID distribution, always remains small, close to 0. See this in table 6.4.

	Model 3	Model 4	Model 5
Variance of spatial only random effect	0.00962 (0.01335, 0.00719)		0.00982 (0.01351, 0.00737)
Variance of temporal only random effect	0.0052 (0.00766, 0.00353)	0.00351 (0.00544, 0.0023)	

Table 6.4: The posterior mean and its 95% CI for the variance of the spatial only and temporal only random effect in model3, 4, 5 respectively.

To address this, we attempt to impose a restriction on the prior variance of the spatial random variable L_i and the temporal random variable T_j, to make them follow a uniform distribution whose lower limit is at least 1.

$$\text{Model 6: } u_{ij} = b_0 + \sum_{n=1}^{8} \beta_n m_{in} + S_{ij} + L_i + T_j$$

The prior variance of the unstructured random effect is set like this:

```
hyper =list (prec = list (prior = "loggamma", param = c(2.252, 6.698))))
```

For example, through this setup, we enforce the standard deviation of the random variable prior to follow a uniform distribution U[1,3], which is not vague prior. But the results are very poor (we do not show it here); this method is not feasible. This may be because the 'outbreak' occurring only at specific time or space points are not significant or it may be the spatial-temporal interaction created by SPDE is so flexible to capture most changes, includes unstructured variance.

6.2.5.1 Discussion about the model.

Here we primarily observe two outcomes: one is the correlation range (km) in the Matérn covariance,

and the other is the correlation parameter ρ of the autoregression. Firstly, a correlation range (km) in the Matérn covariance that is too small is not desirable; typically, it is a result of overfitting. From Model 3 and Model 4, we can observe that they both exhibit a relatively high fitting rate but an obviously low predictive rate. At the same time, we can note that the posterior mean of the correlation range (km) in the Matérn covariance for Model 3 and Model 4 is significantly low, at 45.577 km and 36.110 km respectively. It is worth noting that we are modeling from a national perspective, and such values are evidently not appropriate.

Next, we consider the correlation parameter ρ of the autoregression. Here, we can observe that regardless of whether its prior is 0.8 or 0.9, the posterior mean of the correlation parameter ρ for the models without spatial-only random effects (models 1, 2, 4) is basically 1 (0.99). This indicates that that there may be a structure difference in the wastewater viral concentration across the space. Why? The reason for the correlation parameter of AR1 of spatial-temporal interaction is so high when we do not include the spatial only random effect is that the region i and the region j have the differing viral concentration. So, the inevitable way to get a good fitting of the potential differences in space when the model includes a spatial-temporal interaction effect is that if the temporal correlation is super high which forces the within-spatial unit viral concentration of wastewater risk to be similar to other time points within the same spatial unit. So, the spatial only random effect is essential to be included in the model. Combine discussed correlation range (km) in the Matérn covariance,

$$\textbf{model 5: } \boldsymbol{u_{ij} = b_0 + \sum_{n=1}^{8} \beta_n m_{in} + S_{ij} + L_i}$$

with the prior of the AR1 (autoregressive with order 1) parameter $\boldsymbol{\rho}$=0.8.

It is the optimal model, and we will use this model for later spatial-temporal interpolation and forecasting.

6.3 Result of the model and model evaluation

6.3.1 Result of the model

We first look at the posterior estimates of parameters associated with the spatial-temporal random effects. Then is the result of fixed effects.

	Posterior mean and its 95% CI
	Model 5
Residual precision	92.862 (88.604, 97.54)
Correlation range (km) in the Matérn covariance	262.997km (206.426, 330.196)
Standard deviation in the Matérn covariance	0.124 (0.108, 0.142)
AR1 coefficient	0.908 (0.879, 0.934)

Table 6.5: Posterior estimates of parameter associated with the spatial temporal random effect for model 5.

Predictors	Mean	SD	Q0.025	Q0.5	Q0.975
Intercept	0.422	0.010	0.403	0.423	0.442
SDI	-0.062	0.022	-0.106	-0.062	-0.018
Log population density	0.041	0.013	0.016	0.041	0.065
Prop 65	0.020	0.016	-0.012	0.020	0.053
Prop 15	0.003	0.017	-0.030	0.003	0.037
Ethnic proportion	0.058	0.024	0.011	0.058	0.104
School term 1	0.013	0.007	-0.0009	0.013	0.026
School term 2	0.010	0.007	-0.003	0.010	0.023
School term 3	0.00008	0.007	-0.013	0.0008	0.013

Table 6.6: Posterior estimates of the fixed effects (Mean, SD, and 95% quantile).

6.3.2 Model evaluation

In terms of model performance, it is evaluated by fitting accuracy, prediction accuracy and cross-validation. Both fitting and prediction were performed 10 times and then averaged. In the case of prediction, this refers to the ability to predict the next period. We performed fitting for periods Week 1-70, Week 1-71, Week 1-72, …, and Week 1-79 respectively, and used the above model to predict its next period, which is the Week 71, Week 72, Week 73…, Week 80.

We also used cross-validation. As mentioned before due to the irregular sampling scheme, the number of sampling records at different sampling locations varies during the study period. So, we perform two types of 10-folds cross-validation respectively. One is to perform cross-validation on all records no matter where or when it was collected, and the other is to perform cross-validation from the perspective of 134 sampling locations.

1. CV by records: Divide all records (regardless of sampling time and location) into 10 parts, 9 of which are used for fitting and 1 for verification, repeated 10 times.

2. CV by catchments: Some sampling catchments have too few records, which may bias the results if they are selected into the validation group. So, the sampling catchments with the number of missing records greater than half of all study periods are always placed in the fitting group. Then, divide the remaining sampling catchments into 10 parts, 9 of which serve as the fitting group together with the previously selected catchments with too few records, and one of which serves as the validation group. All records from the fitting group are used for fitting, and from the validation group are used for prediction and validation.

For: Model 5: $u_{ij} = b_0 + \beta_n m_n + S_{ij} + L_i$

	Pierce correlation coefficient	Mean absolute percentage error (MAPE)	Root means square error (RMSE)
Fitting	0.836	0.084	0.961
CV by records	0.766	0.097	1.140
CV by catchments	0.513	0.123	1.324
Prediction	0.519	0.101	1.041

Table 6.7: Summary of the Pierce correlation coefficient, mean absolute percentage error (MAPE), and the root means square error (RMSE).

The prediction accuracy is affected by the quality of data in the nearby catchment area—the more regular the sampling frequency, the higher the accuracy. We do not do cross-validation by region (one region usually has multiple catchments), and it may not make much sense here since irregular sampling occurs at different frequencies in different areas.

6.4. Prediction

6.4.1 Potential wastewater concentration

The predicted value of sa2 wastewater virus concentration is divided into two parts; one is the fixed effects, that is, the covariate (shown in table 6.6). The other part is the random effect (shown in table 6.5). The fixed effects explain the likely potential wastewater viral concentration level before any catchment sampling data is available. We plotted maps for each Sa2 area based on the covariates (Figure 6.4), which could be used to explain the potential risk of virus concentration in wastewater before obtaining any sampling data.

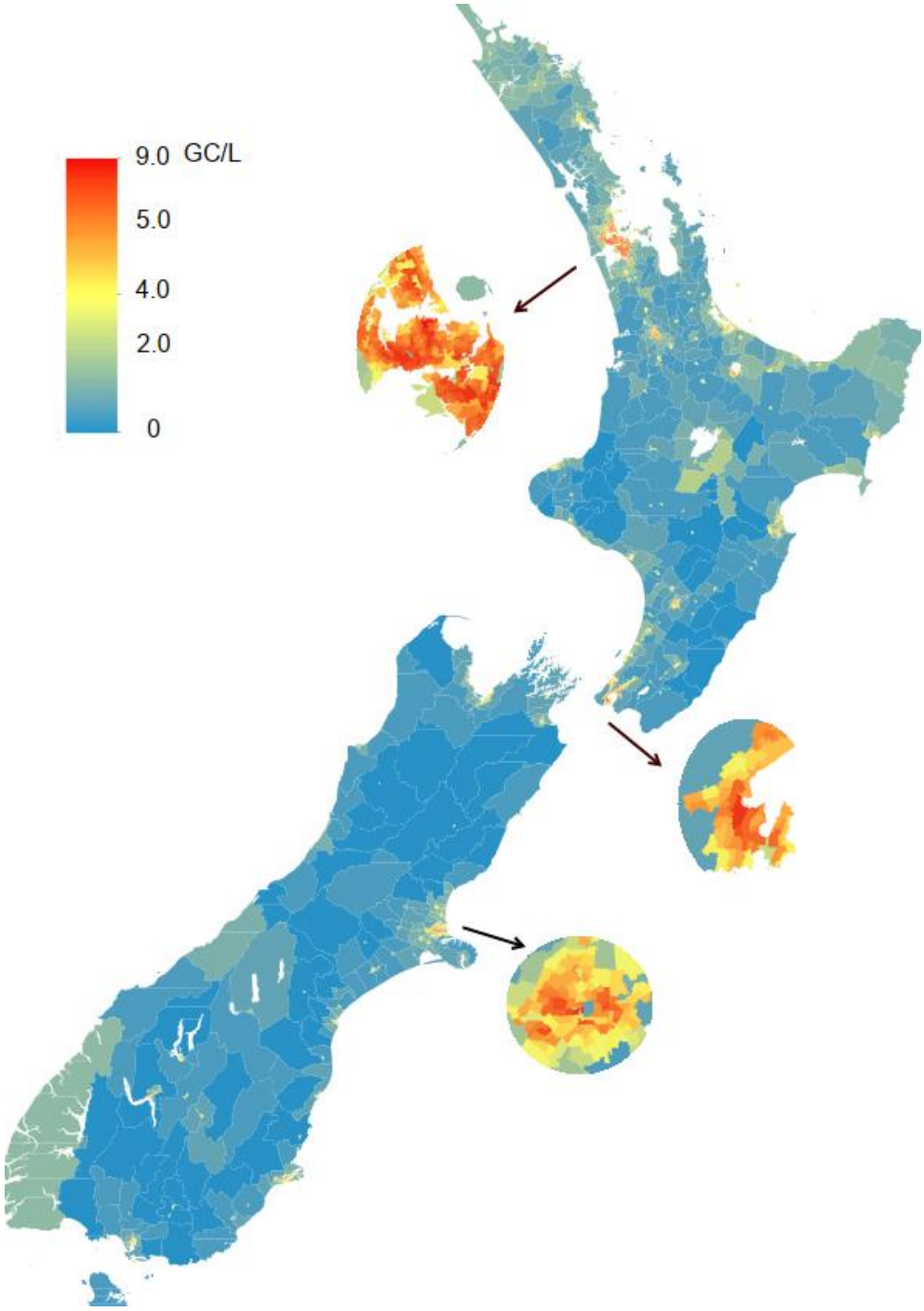


Figure 6.4: Potential wastewater concentration based on the corvariates of each sa2 areas with zoom in of Auckland, Wellington, and Christchurch councils.

While conducting spatiotemporal interpolation of the model and predicting the next period, we also pay attention to demographic covariates (fixed effects) and their impact on the representation of

COVID-19 prevalence rates in wastewater virus concentration. In our results, regarding all fixed effects, we note that three fixed effects are significant, namely SDI (Sustainable Development Index), population density, and ethnic proportion. The result shows a positive relationship between the population density (number of people/km^2) and the wastewater virus concentration (GC/L) between 0.016-0.065, and the posterior mean of the corresponding regression coefficient is 0.041. Similarly, the result demonstrates that the proportion of ethnic people is strongly associated with wastewater virus concentration (GC/L) with a posterior mean of 0.058 and a 95% posterior CI (0.011-0.104). A negative correlation was also identified between the SDI values and the virus concentration in wastewater, with the posterior mean of -0.065 and 95%CI (-1.031, -0.027). For the remaining covariates, the evidence of association with viral concentrations is weaker but still worth discussing. There are two population age structures and three time indicators related to school terms. The posterior mean for both age structures, Prop65 (proportion of people over 65) and Prop15 (proportion of people under 15) is positive. The value for Prop 65 is 0.021, whereas the value for Prop 15 is only 0.0002.

About the variable "school_term." For epidemics, non-pharmaceutical interventions are effective measures when effective treatments are lacking. They include schools, nonessential businesses, restaurants, bar closures, etc. Kids frequently play a crucial role in spreading viral epidemics such as influenza, mainly due to their extended periods of close interaction with other children at school. So, we have incorporated three fixed effects related to the school term into the model. They are all binary variables, where 0 represents holidays, and 1 represents the school term. These are defined as lag periods of one week, two weeks, and three weeks, respectively. Auger et al. (2020) conducted a study on the closure of schools in the United States and the incidence and mortality rates of COVID-19. They found that school closure in the US was temporally associated with decreased COVID-19 incidence and mortality. In states where schools were closed earlier when the cumulative incidence of COVID-19 was lower, the relative reduction in both incidence and mortality was most significant. In our model, school terms with three different lag periods also exhibit a positive correlation with virus concentration in wastewater, meaning that virus concentration in wastewater decreases during holidays (school closure periods). Notably, their correlation is strongest with a one-week lag and weakest with a three-week lag. For population density some studies have

shown a strong correlation between population density and COVID prevalence. Iderus et al. (2022) study the Correlation between Population Density and COVID-19 Cases in Malaysia. Considering absolute population and population density, and they find more populous and densely populated districts have a higher risk of transmission of COVID-19. However, some studies have found that the relationship between population density and COVID-19 prevalence is complex and related to accounting for other confounding factors such as income, health insurance, etc. (Carozzi, 2020). For SDI, some studies suggest a positive correlation between the SDI value and the prevalence of COVID-19. Ossimetha et al. (2021) conducted research on counties in the United States with at least one SARS-CoV-2 case and found that counties with higher social deprivation scores experienced more significant growth in SARS-CoV-2 cases and deaths. In our model, however, SDI presents an opposite result, indicating that areas with lower SDI have higher concentrations of viruses in wastewater. This may be because, in New Zealand, high SDI values typically occur in sparsely populated areas, where the prevalence is relatively low. Of course, this does not rule out spatial confounding caused by multiple covariates.

6.4.2 Spatial interpolation based on SA2.

In order to achieve community-level monitoring, we will predict the wastewater viral concentration in each SA2 area (2018). SA2 areas are usually suburbs or part-suburbs with 2,000 to 4,000 residents. In rural districts, many SA2 areas have populations of fewer than 1,000 residents. In the same way as the catchment area, the prediction points are the population weighted centroid of each sa2 area calculated based on the New Zealand 2018 Estimated Resident Population Statistical Grid (Figure 6.5). The value at this point will represent the wastewater concentration of the entire sa2 area. Since we're looking at the concentration of viruses in wastewater, and viruses are produced by people, it is reasonable to use population-weighted centroid. There are (l=1:2253),2253 sa2 areas, at 79 weeks (j=1:79) to be predicted and they followed by:

$$SA2_{lj} = b_0 + \beta_n m_l + S_{il} + L_i + T_j$$

Where the covariates are the same as described above.

The spatiotemporal interactions at each time index of each location are projected to each SA2 area through the existing catchment sampling data and the mesh. That is to say, predict each sa2 value by computing their predictive distribution:

$$\pi\left(y_{SA2_{lj}} \mid \mathbf{y}_{catchj}\right) = \int \pi\left(y_{mSA2_{lj}}, \boldsymbol{\theta} \mid \mathbf{y}_{catchj}\right) d\boldsymbol{\theta}$$
$$= \int \pi\left(y_{mSA2_{lj}} \mid \mathbf{y}_{catchj}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta} \mid \mathbf{y}_{catchj}\right) d\boldsymbol{\theta}$$

To achieve community-level surveillance, we predict the virus concentration in wastewater in each SA2 area. The community-level surveillance result every 10 weeks is shown in Figure 6.6. Spatial interpolation prediction for each week could be found in attachment file.

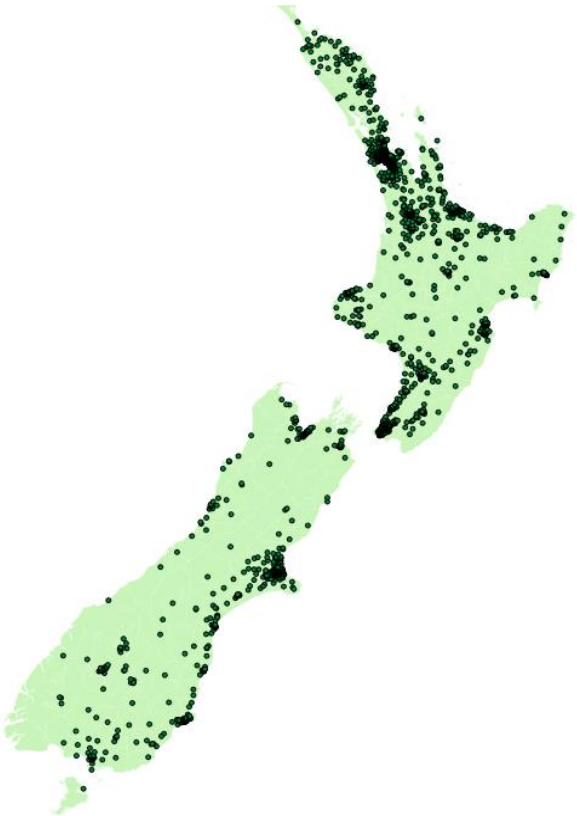


Figure 6.5: Population weighted centroid of each sa2 areas.

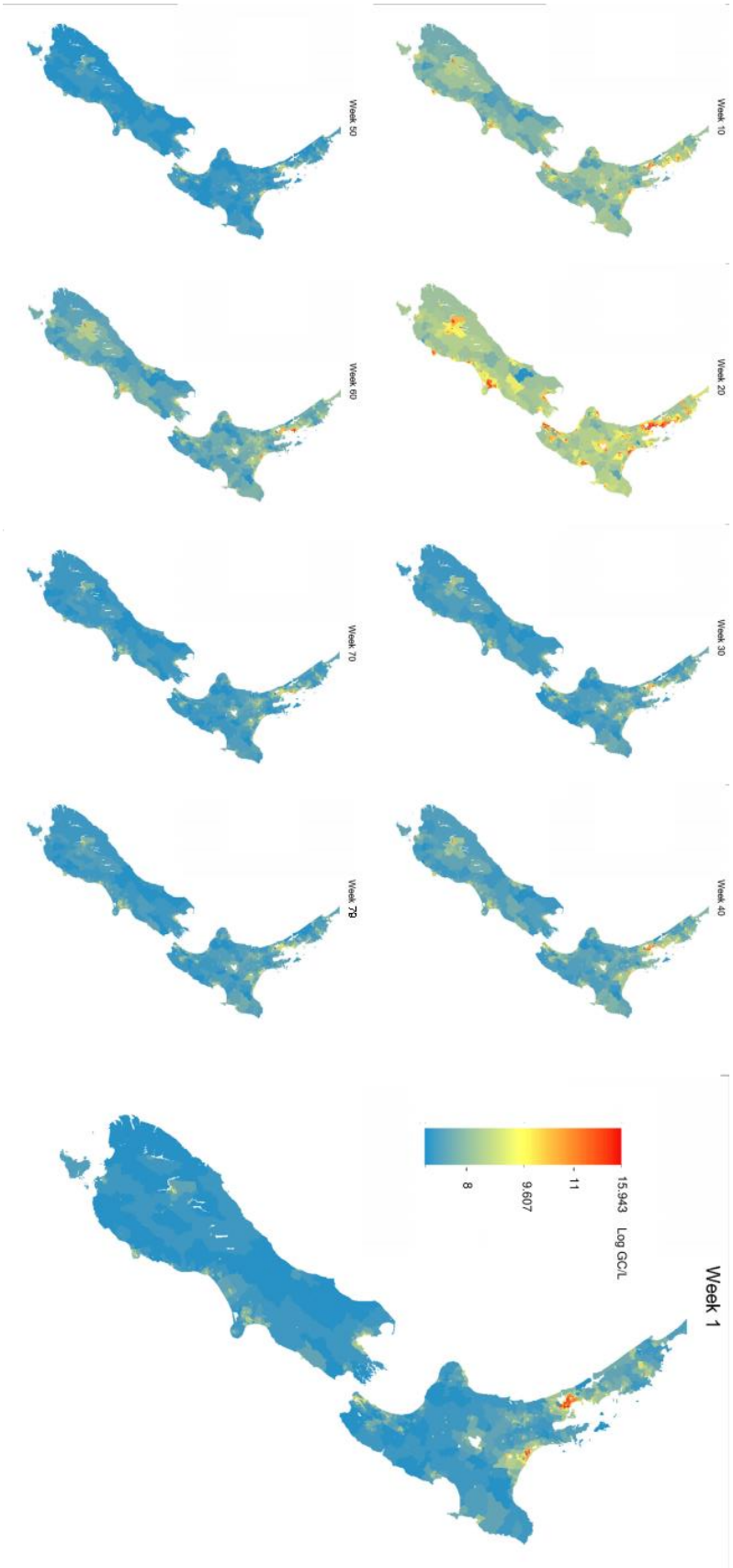


Figure 6.6: Posterior mean of spatial interpolation predictions for each SA2 area every 10 weeks.

The figures of the posterior mean of spatial interpolation predictions of each SA2 areas for each week during the study period could be got in the attachment.

6.4.3 Prediction on next period

Our study period is from “2022.02.27” to “2023.08.27”, there are total 79 weeks. Based on the model proposed, We fit the model using data from the weeks 1-79， and predict the virus concentration of wastewater at each catchment area in the next following period which is the 80th week. The evaluation of the model's performance in predicting the next period can be seen in Table 6.7. The wastewater catchment sampling is irregularly motivated by economic benefit, so there are missing values in the catchment data. There are 44 out of 134 sampling catchments have records on the 80th week (2023.9.07).

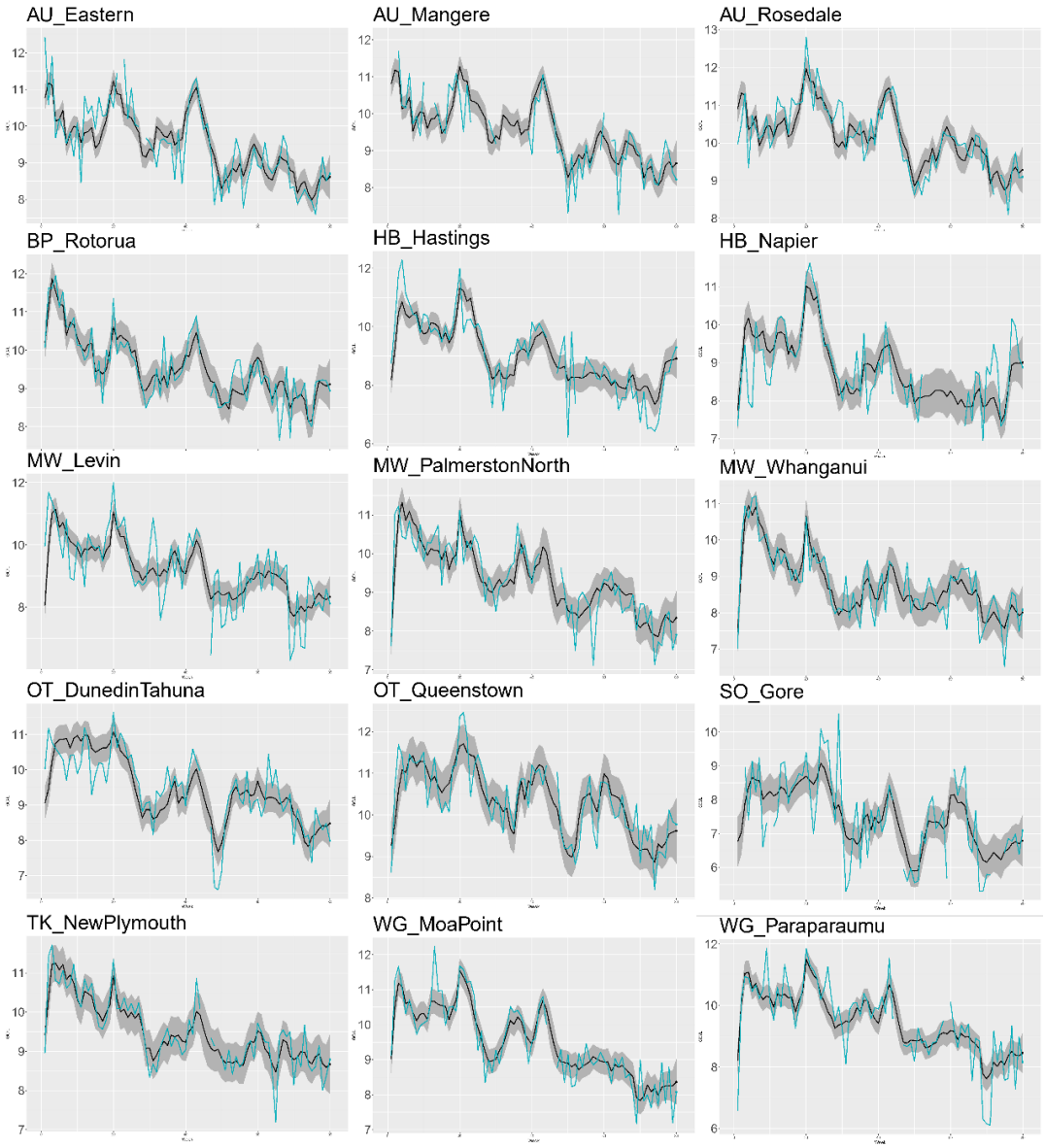


Figure 6.7: The real and the model fitted virus SARS-CoV-2 concentration (GC/L) in wastewater of some catchment areas during the 79 weeks study period, and the prediction for the next time point (week 80). (The blue is for the real records. The black is for the posterior mean of fitting and prediction and the shadow is its posterior standard deviation. The point at the end is the value for the week 80. Graphs of all catchment areas are in supplementary material.)

There are 16 big regions in NZ, and we reaggregate virus concentration from the catchments into regions by population. Due to the irregular sampling of catchment areas, the real values of regions here are only displayed when all catchment areas of the region have sampling records. We also show

values of a dominant catchment in the region, which is usually the catchment with the largest population and the most frequently sampled.

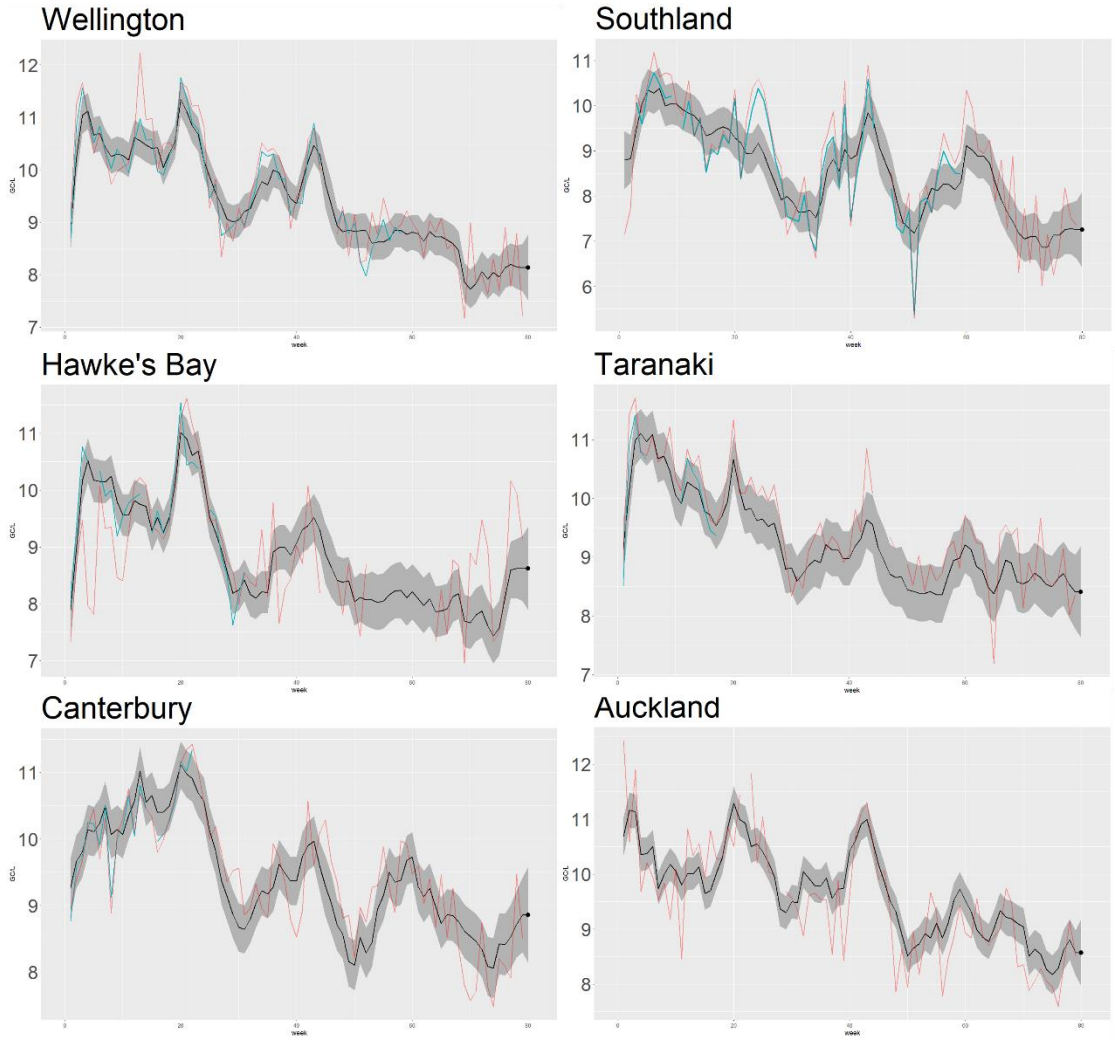


Figure 6.8: Reaggregated virus SARS-CoV-2 concentration (GC/L) in wastewater from catchment areas to regions. (The aggregated true value is in blue, and the red is the true record of one of the catchment areas in the region. The black is the aggregated posterior mean, and the grey is the aggregated posterior standard deviation. They are aggregated by population of catchment areas.

From here, it can be seen that regardless of whether it is from the perspective of the catchment area or a more extensive regional perspective, it appears that different regions share a similar time-series pattern in virus concentration in wastewater. Take the catchment area, for example; there is a noticeable increase in virus concentration in wastewater in the 20th and 45th weeks, followed by a

rapid decline in the 50th week. These changes within catchment areas belonging to the same region are more similar. From a regional perspective, although not as pronounced as in catchment areas, similar temporal patterns could still be observed. From a particular perspective, this may suggest the possibility of using time series to forecast virus concentration and explain the rationality of using irregular sampling for wastewater sampling schemes.

6.5 Discussion:

This study uses Bayesian approach to model the wastewater SARS-CoV-2 concentration from a national perspective using the irregular sampling data from all wastewater catchment areas in NZ. While considering the fixed effects, which are covariates, we also add a spatial random field, which depends on the time index, into the model to capture local departures used to smooth factors such as population mobility. To study from a national perspective, we use population-weighted centroids to convert catchment polygon data into point estimates for study. Similarly, to achieve the prediction of community level, which is sa2 areas, each sa2 area is also replaced by the population-weighted centroid. Then, the fitting and prediction of the model could be done in a continuous spatial field. At the same time, the Bayesian point-referenced model can be used for misaligned data, which is caused by an irregular sampling plan of catchment areas. Naturally, the fewer the sampling points and the farther away they are from the sampling points, the greater the standard deviation of its posterior value will be.

Many discussions have occurred regarding the relationship between virus concentration in wastewater and disease prevalence in its corresponding catchment areas. This study does not discuss the relationship between wastewater concentration and epidemiological indicators and their time lag. This is because the maximum time resolution we can obtain for wastewater virus concentration is weekly averaged, and due to irregular sampling, there are a large number of missing values in the original data. In another study, Li G et al. (2023) attempted to use weekly data to assess the lead or lag between wastewater signals and traditional epidemiological indicators from all catchment areas but realized that daily data is more suitable than weekly data. Some studies to evaluate the lead and

lag between daily wastewater concentration data and epidemiological indicators have been performed on small scales (specific areas, isolated buildings) (Hewitt et al., 2022). However, there are still many challenges in conducting this type of study on a national scale using data from all catchments, like potential weekly periodicity, irregular sampling patterns, and significant variability in the lag between the times of infection and detection (Li et al.,2023). There are also still numerous confounding factors to consider. For example, variations in disease reporting methods, population demography in different catchment areas, and differences in detection equipment. When conducting research on a large scale, one must also consider the impacts of spatial heterogeneity and different aggregation levels. The study by Faraway et al. (2022) focused on utilizing wastewater-based epidemiology (WBE) to quantitatively detect or predict disease prevalence. However, this remains challenging, especially when wastewater data is discontinuous. They found that the spatiotemporal relationship between wastewater and prevalence is dynamically changing. Therefore, this paper only addresses the spatial-temporal interpolation and prediction of wastewater concentration. How to accurately translate wastewater virus concentration data into intuitive information, such as potential prevalence, remains a future endeavor. Similar time series patterns of wastewater virus concentrations in different catchments appear to be noteworthy. A region often has multiple catchments, and its wastewater viral concentration is usually dominated by one or two main catchments (Figure 6.8). The population size of these main catchment areas is relatively large, which is natural. But across the country, we also find that different catchments, even though they are far apart, appear to share the same temporal trends. In Figure 6.7, they are in the plot with different y-axis. Whether the expansion or compression of the time series plots is related to population density, population size, or something else needs to be explored. This inspires us to study the relationship between wastewater viral concentrations of different catchments and the national level, and to consider the possibility of using wastewater virus concentrations from some representative catchment areas to extrapolate its concentrations at national level. The design of wastewater sampling scheme may also be improved from similar studies so that they are both cost-effective and produce high-quality data.

The wastewater sampling plan needs economic advantages, so not all catchment areas are included in the sampling framework. The catchment areas within the sampling framework are typically those with a more concentrated population distribution and are representative. There may be certain biases

when applying the fixed effects from the resulting fitted model for spatial interpolation predictions in other regions. For example, in areas with very low population density, the interpolated prediction of wastewater concentration may also be very low (catchment areas with low population density are not included in the fitting dataset). But is that true? We do not know. The estimation bias introduced by sampling also needs to be considered. Spatial confounding, often discussed as collinearity between fixed and random effects, also needs to be considered. When covariates vary across areas or exhibit spatial correlation, spatial confounding is highly likely to occur. The most significant impact of spatial confounding is on the fixed effects estimation. However, since this paper primarily focuses on spatial interpolation and predicting for the next period, the influence of spatial confounding is relatively minor, thanks to the flexibility of random effects in SPDE-based modeling. Therefore, we did not discuss the issue of spatial confounding in this paper. Another problem is that spatial dependency random filed based on geographical location may not truly express the pattern of population mobility. The way people travel is changing. It is easy for people to commute long distances without stopping through cars, planes, and other kinds of transportation, and this may become the norm. This may cause spatial fields based on geographical location not to work. Modelling long-distance interactions is one of the challenges in spatial epidemic models (Riley S et al. 2015). In future research, we may consider using population flow data to establish a spatial random field instead of a geographical location. In wastewater-based Epidemiology (WBE), we should also pay attention to rainfall or weather data. Combined sewers, which are stormwater + sewerage, present challenges, and the weather effect will also be an issue for catchment areas in cities (Phil et al., 2018). These are the kinds of things that we expect to be addressed in the data collection or modelling.

6.6 Conclusion:

This study used a Bayesian approach to model the wastewater SARS-CoV-2 concentration from a national perspective using the irregular sampling data from all wastewater catchment areas in NZ. According to the results of cross-validation and prediction, the spatial-temporal Bayesian linear model based on SPDE performs well, so we can spatially interpolate wastewater concentrations over small areas at a good temporal resolution while also predicting catchment wastewater concentrations

at the next period with relative accuracy. Although there is a more considerable variance in predicted values where there are fewer sampling locations, irregular sampling is economically beneficial. The model overcomes irregular sampling and maps wastewater SARS-CoV-2 concentrations from a national perspective that will be helpful to public health personnel. Meanwhile, we fitted and predicted the virus concentration in wastewater from two perspectives: catchment area and region. The graphs indicate similar temporal patterns, which provides the basis for using time series to forecast virus concentration in the future. This model can flexibly deal with the changes in sampling locations, meeting the demand for surveillance and prediction of COVID-19 epidemic levels in populations using wastewater virus concentration over long periods and within smaller areas.

Data availability：

The wastewater data is from Aotearoa Wastewater Surveillance Programme (https://github.com/ESR-NZ/covid_in_wastewater).

The catchment polygons are not publicly available. They are the intellectual property of the individual regional and city councils that operate the treatment plants where the samples from.

Sa2 areas polygons are from Statistical Area 2 2018

Population grid we use is New Zealand 2018 Estimated Resident Population Statistical Grid.

Population density, age structure and ethnic proportion are from census 2018 NZ. They all are available at STATNZ (https://www.stats.govt.nz/)

Socioeconomic Deprivation Indexes (SDI) is available from (https://www.otago.ac.nz/wellington/departments/publichealth/research-groups-in-the-department-of-public-health/hirp/socioeconomic-deprivation-indexes-nzdep-and-nzidep-department-of-public-health#2018).

All maps are based on New Zealand Transverse Mercator 2000 (NZTM) (EPSG:2193)

Acknowledgments:

We thank Massey University's subscription to New Zealand eScience Infrastructure (NeSi) allow us use high-performance computing clusters. We are also grateful to Environmental Science and Research (ESR) for providing data support.

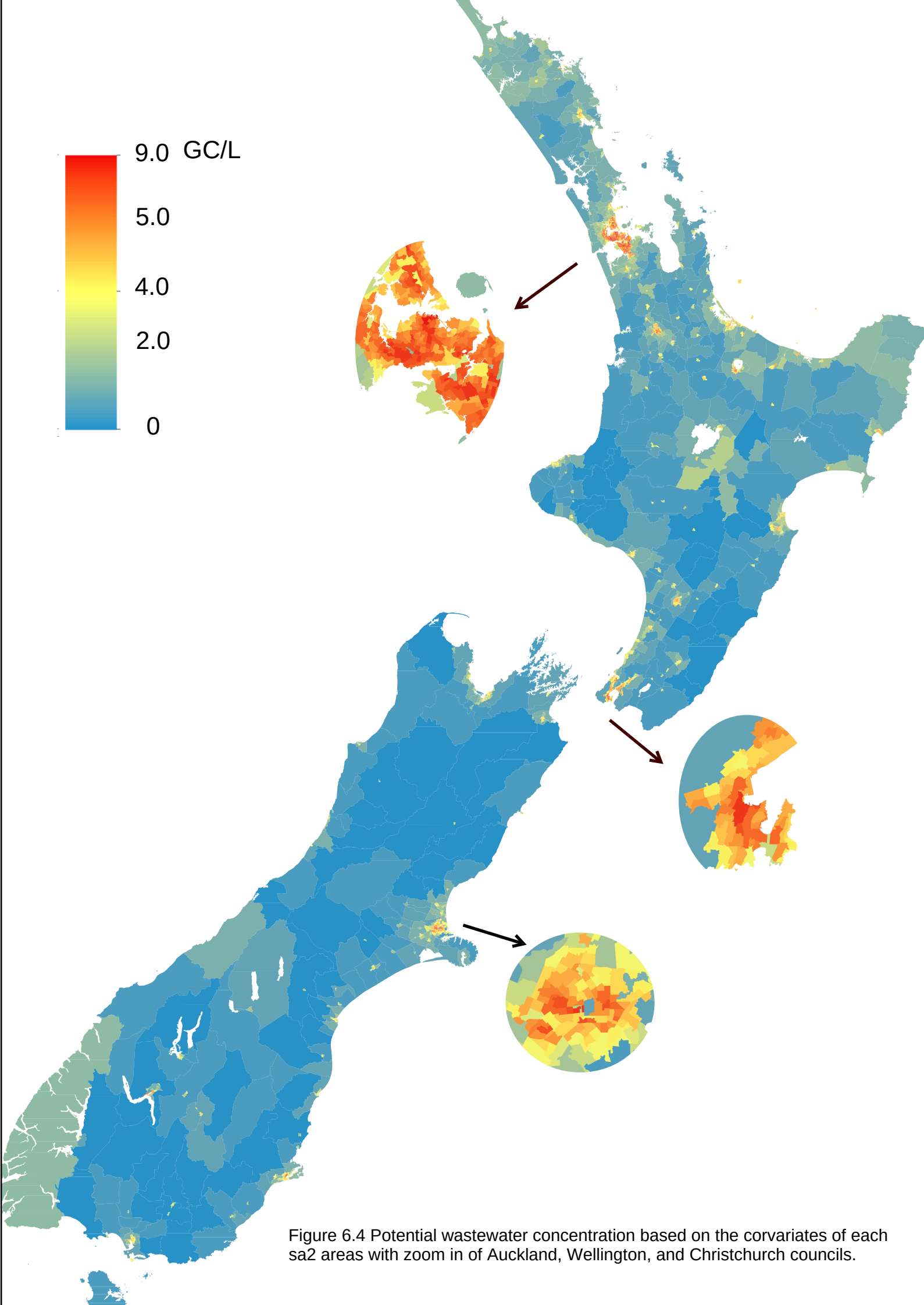


Figure 6.4 Potential wastewater concentration based on the corvariates of each sa2 areas with zoom in of Auckland, Wellington, and Christchurch councils.

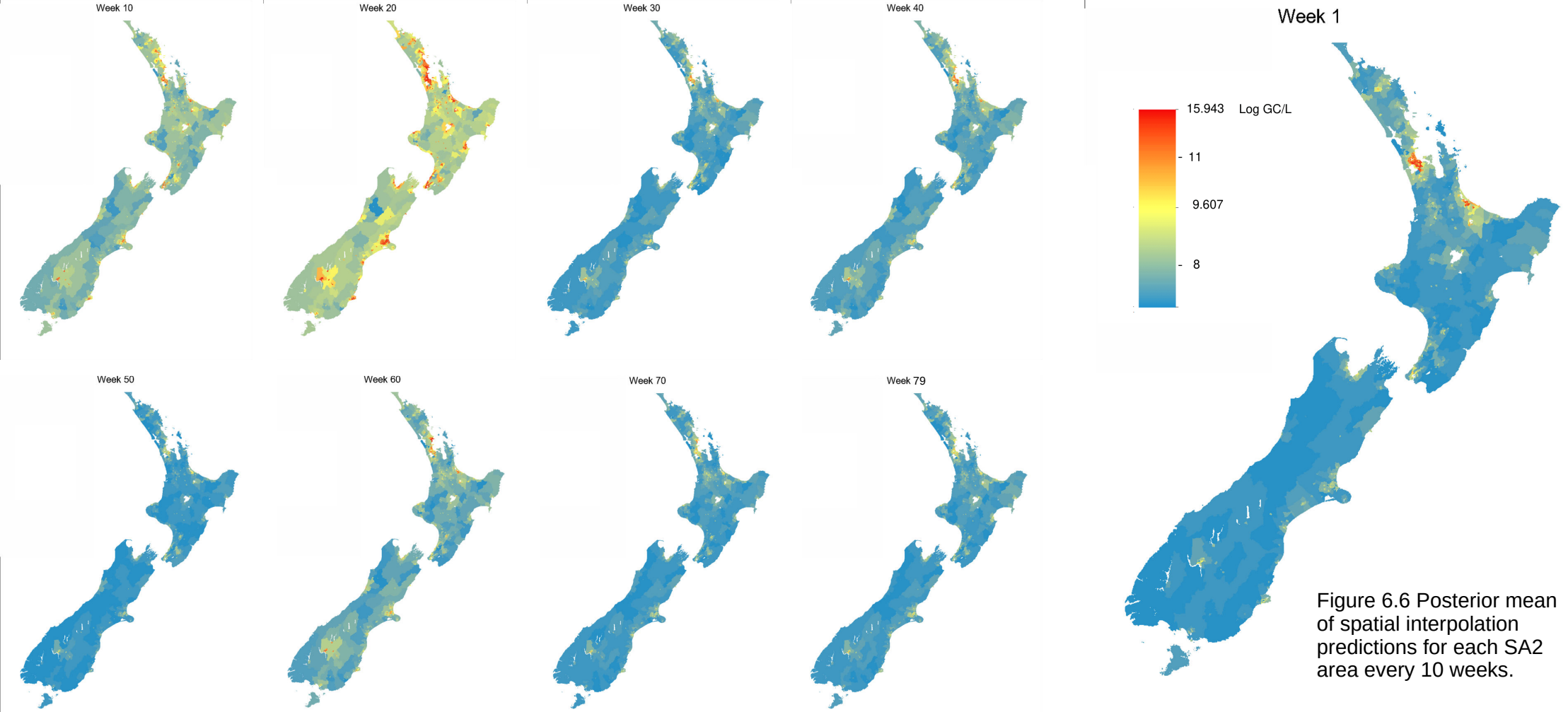


Figure 6.6 Posterior mean of spatial interpolation predictions for each SA2 area every 10 weeks.

Chapter 7: Discussion

In this chapter, we discussed the transformation of data types related to the flexible use of spatial epidemiological data and modeling methods. It is crucial to model flexibly according to the problem we aim to solve in epidemiological modeling rather than adhering to specific data formats corresponding to models.

7.1 Flexibility of Spatial Data Type Conversion

Data regarding human health outcomes are commonly accessible in spatially discrete scales, often represented as counts of cases within small areas like administrative regions. Due to the accessibility and low confidentiality concerns associated with areal data, discrete spatial models in epidemiology have seen widespread and innovative development (Lawson et al., 2016). On the contrary, point pattern data, due to the difficulty in obtaining highly precise geographical locations, issues related to high confidentiality, and controversies surrounding the use of point processes to simulate disease spatial patterns, have been decreasingly utilized in spatial epidemiology and related disease modeling (Diggle,2013a). As for another type of data with precise geographical locations, geostatistical data (point-reference data), although similarly constrained by confidentiality concerns in the health science or epidemiology field, holds a lower priority in usage compared to spatial area discrete models. However, with the application of mobile technology in health science and advanced epidemiological surveillance technology, the use of data with georeferencing to point locations would be more common (Diggle, 2019).

Before introducing the spatial data transformations, let us start with a basic introduction to data with spatial properties. In the spatial models, every response, Y, is linked to a spatial location; it could be a polygon or a point. However, in numerous uses of geostatistical models, the observed response

represents a characteristic of a finite region where it is merely a convenient reference point. For example, each Y refers to a village community that extends over a finite area and serves as a somewhat arbitrary reference location within the village's boundary (Diggle, 2019). Usually, we are interested in information related to people, or in cases where the population is not evenly distributed within this area, we use the population-weighted centroid to represent this point. In traditional geostatistics, the defined area where this information originates from is termed the support of Y (Rossi et al., 1993). Informally, ignoring the support of each response is equivalent to assuming that for all quantities of interest, the variation within the extent of each single support is negligible compared with the variation between different supports (Lawson et al., 2016). This implies that the larger the geographical range of each support, the greater the degree of aggregation and the less believable this assumption becomes.

Regarding spatial resolution, point pattern data has the highest spatial accuracy, followed by point-referenced (geostatistical data) and areal-level data (lattice data). Typically, we aggregate higher-resolution data types into lower-resolution ones, and this process may be completed before the data is available due to confidentiality. At the same time, it is worth our attention to three kinds of spatial data modeling theories for areal-level data (lattice) data, point-referenced (geostatistical data), and spatial point process (point pattern) data. In modeling areal-level data (lattice data) and point-referenced (geostatistical data), both utilize spatial relationships to share information for spatial modeling purposes. In modeling areal-level data (lattice data), spatial modeling is achieved by defining spatial weight matrices W and discrete Markov random fields (CAR structures). However, in point-referenced (geostatistical data) modeling, spatial modeling is achieved through spatial correlation functions to provide a spatially continuous interpretation. In SPDE method (Lindgren et al., 2011), it needs to depend on the mesh. Linear statistical models and discrete spatial models both rely on Gaussian random fields for modelling, while the spatial point process, a statistical model used to analyse the positions of events within a spatial area, its methods and models have developed separately from geostatistics and discrete spatial models, along with their respective methodologies. Diggle et al. (2013b) show several examples of using various spatial modelling methods that do not match their corresponding data types, aiming to better understand situations in modelling that might be misleading. Next, we discuss some specific cases of spatial data transformation and modelling:

1. Transform data from high resolution to a lower resolution.

A) Transformation from point pattern data to point-referenced (geostatistical) data and area-level data.

This is the simplest form of aggregation, as individual health data often requires high confidentiality due to exact geographic locations or needs modeling on a larger spatial scale. Typically, point-pattern data is aggregated to a representative point or area. For instance, within a village, assigning a common location for all cases could be the town's geometric centroid or population-weighted centroid (Diggle, 2019). Alternatively, within small administrative units, simply summing up all the records. Whether aggregating point pattern data to point or region, the responses Y represent case-count data, but they have different supports X.

B) Transformation from point-referenced (geostatistical) data to area-level data.

This method is also quite straightforward, aggregating all point-referenced data located within the same region into an area-level dataset. However, based on the characteristics of the data we discussed earlier, the larger the geographical range of each support, the less believable this assumption becomes. However, geostatistical models should be considered the default option when point locations are available, and fitting area-level models to aggregated points can be avoided. (Lawson, et al., 2016)

C) Without aggregation.

One way is to transform the spatial point study into an area-level data study with a discrete spatial model. Every point within a study area can be seen as a collection of individual subregions, and we could assign a subregion to each point. Discrete spatial modeling requires the definition of a spatial weight matrix. While this kind of directly similar to the use of discrete spatial models for modeling areal-level data is easy to understand, a significant drawback with it is that the spatial weight matrix can become extremely large and complex as the number of point locations in the dataset increases (Pfeiffer et al., 2008). If we construct the spatial correlation matrix W for the lattice on the right side

of Figure 7.1, it will be huge (653 x 653 matrix). If more complex definitions of spatial matrix W are adopted, such as boundary proportions, W will also be very complex.

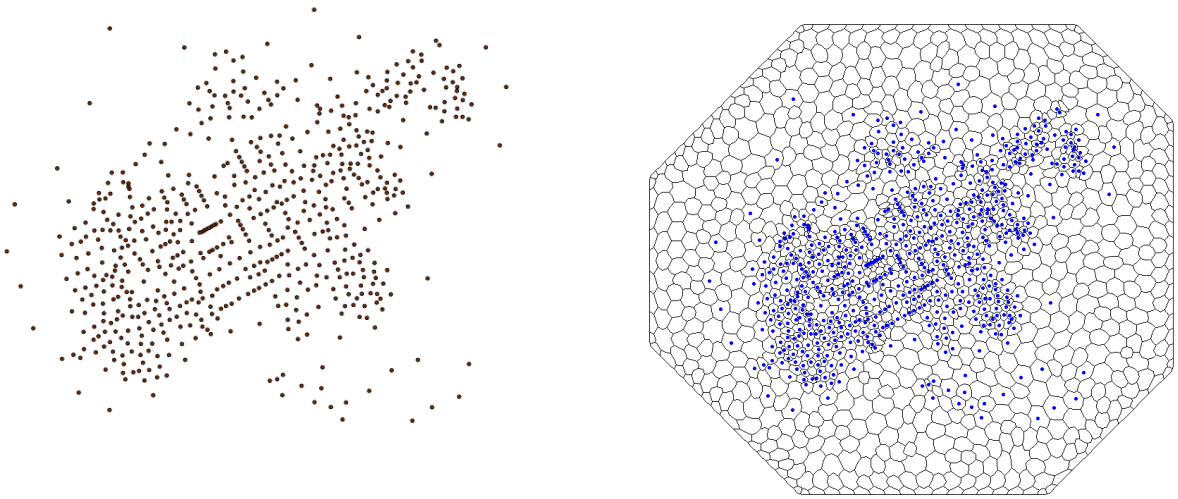


Figure 7.1 Schematic representation of the method C) using the example of mesh blocks of the Palmerston North

2. Transform data from lower resolution to a high resolution.

Transformation from lattice data to point-referenced data.

Usually, this process occurs prior to the disclosure of data due to the high confidentiality of health or disease-related data. Additionally, as lattice data is typically employed for analyzing small-area data and modeled using discrete spatial models, using the same model for spatial modeling of nonadjacent lattice data across large regions is not feasible due to gaps between lattices. There is a kind of adaptive CAR model that allows the spatial weight matrix W to be modeled as binary random variates within CAR formulation (Lee & Mitchell, 2012) that could be used for nonadjacent areas modeling, but the

result is often difficult to interpret. Another viable approach is substituting appropriate points for lattices, thereby enabling geostatistical modeling over larger regions. For instance, in the conclusion of this paper, we presented an example of spatial modeling of COVID-19 prevalence using wastewater epidemiology. Given the gaps between catchment areas or the lack of wastewater sampling in all catchment areas, traditional discrete spatial models cannot be applied. Thus, we use population-weighted centroids of each catchment area instead of the polygons themselves for geostatistical modeling. However, there are some issues to consider: 1) Ignoring the shape and size of the lattice. 2) If adjacent polygons (lattices) share boundaries, should their spatial relationships be redefined when transformed into point-reference data? It is also related to the size and shape of the lattice. These are worthy of consideration.

Chapter 8: Conclusion

8.1 Conclusion

This thesis introduces three modeling approaches for spatial epidemiological data, primarily focusing on regional and point-referenced data. For areal data modeling, we discuss various discrete spatial models based on Conditional Autoregressive (CAR) processes and how to extend spatial modeling to spatiotemporal modeling. Regarding point-referenced data, we cover linear spatial statistical models and generalized linear spatial statistical models. We describe the process of modeling non-stationary statistical models using the INLA-SPDE (Integrated Nested Laplace Approximation - Stochastic Partial Differential Equation) method (Lindgren et al., 2011) and how to create spatial-temporal interaction based on SPDE to extend the spatial model to spatiotemporal model. We apply this method in a case study related to wastewater epidemiological monitoring for potential COVID-19 in New Zealand. In this study, the spatial-temporal Bayesian INLA-SPDE model performs well. We can spatially interpolate wastewater concentrations over small areas at an appropriate temporal resolution while predicting catchment wastewater concentrations in the next period. The model overcomes irregular sampling and maps wastewater SARS-CoV-2 concentrations from a national perspective that will be helpful to public health personnel. Through case studies, we realize that when dealing with spatial epidemiological data, it is essential to flexibly employ modeling methods and transform data types based on the nature of the problem at hand or the desired objectives and should not be restricted by the corresponding methods of data format.

8.2 Current Research and Looking into the Future

Due to the relative accessibility and low confidentiality of areal-level data, discrete spatial models have experienced significant development (Lawson, 2018). From disease mapping, disease mapping

has traditionally focused on identifying patterns of high-risk areas and disease incidence and mortality rates. Apart from mapping diseases and illnesses, some new disease maps have been developed, including but not limited to spatial mapping for rates of healthcare utilization (Nathoo and Ghosh., 2012), for vaccination coverage (Utazi et al., 2018), for gene frequency and the rates of health care utilization and child growth. Extensions have been developed in Conditional Autoregressive (CAR) models, including Multivariate GMRFs, which are MCAR structures for multidimensional disease mapping. For more details about multivariate GMRFs, please see MacNab (2016), Martinez-Beneito (2019), MacNab (2018), and Greco and Trivisano (2019). Another noteworthy development is Adaptive CARs, a spatial statistical modeling technique that incorporates adaptability in capturing spatial dependencies in data. For more detail about Adaptive CARs, please see Gao and Bradley (2019), Rushworth et al. (2017), Lee and Mitchell (2012), and Corpas-Burgos et al. (2020). They represent the latest developments in modeling area-level spatial epidemiology. In the future, the modeling direction will advance further into complex models involving spatial and spatiotemporal, with multivariable aspects and multiple arrays, which may act as covariates. In terms of multivariable aspects, they may include infection, mortality, hospitalization, and vaccination rates. In terms of multiple arrays, multivariable outcomes and dimensions may include age, gender, race, country, and risk factors, which may relate to different health intervention public policies and vaccine accessibility (MacNab, 2022).

In the context of point-referenced data, the INLA-SPDE approach to Bayesian spatial data analysis over high-resolution grids marks a new frontier of Bayesian disease mapping and its applications (MacNab,2011). INLA-SPDE provides the possibility of high-resolution, large-dimensional modeling for linear geostatistical models. Another part of the research focuses on how to create spatial-temporal interactions. Creating distinct spatiotemporal interaction random variables allows us to achieve spatiotemporal modeling with a flexible hierarchical model. A new R package, “INLAspacetime”, provides a way to create a non-separable spatiotemporal interaction random variable based on the diffusion function (Lindgren et al.,2020). Compared to the spatiotemporal separable version, the non-separable one exhibits higher forecasting accuracy and relatively the same interpolation accuracy.

In addition to the latest developments, spatiotemporal epidemiological modeling faces these challenges:

A) large dimension

The advancements in data science and big data initiatives are sparking innovative concepts in Bayesian disease mapping approaches, particularly for managing high-dimensional datasets (MacNab, 2022)—for example, area data or point-referenced data with a large number of observations or time points. One recent example comes from (Orozco-Acosta et al., 2021), which introduces a scalable Bayesian modelling approach for smoothing mortality (or incidence) risks in high-dimensional data when the number of small areas is huge. This approach is based on the idea of 'divide and conquer.' The development of hardware and the emergence of computing methods such as INLA have made Bayesian computations for large datasets easy to implement.

B) Spatial confounding and the scale of the covariates.

Spatial confounding is collinearity between fixed effects (covariates) and random effects in a spatial generalized linear mixed model. It could adversely affect estimates of the fixed effects. As discussed earlier in Chapter 3, there have already been some methods proposed to address and improve such issues; a recent study from Dupont (2022) illustrates a novel method called “spatial+,” which reduces the sensitivity of the estimates to smoothing by replacing the covariates by their residuals after spatial dependence has been regressed away. This is regarded as a potentially innovative approach to spatial confounding.

At the same time, another issue related to modeling with covariates is encountered when conducting ecological modeling-disease mapping involving covariates. Different spatial scales of covariates may have varying impacts on the model. The performance of spatial priors for random effects differs when covariates operate at different spatial scales (MacNab, 2022). For instance, areal-level covariates are typically released based on administrative regions, while point-level covariates, such as most environmental monitoring data, are always based on specific locations. Caution should

be exercised when dealing with diverse modeling objectives. The covariate scale can influence the accuracy and confounding of spatial models, and the quantification of their effects and making optimal choices deserve discussion.

C) Long distance interaction modelling.

Existing geographic-based spatial sharing patterns may not accurately depict spatial interactions. Long-distance interaction has consistently been a challenge in spatial modeling (Riley et al., 2015), especially in studies related to epidemiology, as people's travel patterns evolve. Exploring how to rely on population mobility data to establish relatively accurate models for long-distance interaction in epidemics and comprehensively capture spatial correlations in spatial data is a direction worthy of further research in the future.

Supplementary Material

Code, parts of the data, and attachment files have been uploaded to GitHub but are set private.

References:

[1] Adler, R. J. (2009). Random Fields and Geometry. Springer Science & Business Media.

[2] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723.

[3] Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93-115.

[4] Auger, K. A., Shah, S. S., Richardson, T., Hartley, D., Hall, M., Warniment, A., ... & Thomson, J. E. (2020). Association between statewide school closure and COVID-19 incidence and mortality in the US. *Jama*, *324*(9), 859-870.

[5] Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2015). Hierarchical modeling and analysis for spatial data (2nd ed.). Chapman and Hall/CRC.

[6] Bavaud, F. (1988). Space-Time Models and Geostatistics. Springer.

[7] Bavaud, F. 1998. Models for spatial weights: A systematic look. Geographical Analysis, 30(2):153–171.

[8] Beale, L., Abellan, J. J., Hodgson, S., & Jarup, L. (2008). Methodologic issues and approaches to spatial epidemiology. Environmental health perspectives, 116(8), 1105-1110.

[9] Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 192–236. http://www.jstor.org/stable/2984812

[10] Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics, 43(1), 1-59.

[11] Bivand, R., Rowlingson, B., Diggle, P., Petris, G., Eglen, S., & Bivand, M. R. (2017). Package 'splancs'. R package version, 2-01.

[12] Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-

INLA. John Wiley & Sons.

[13] Boehmer, T. K., DeVies, J., Caruso, E., van Santen, K. L., Tang, S., Black, C. L., ... & Gundlapalli, A. V. (2020). Changing age distribution of the COVID-19 pandemic—United States, May–August 2020. *Morbidity and Mortality Weekly Report*, *69*(39), 1404.

[14] Bolstad, W. M., & Curran, J. M. (2016). Introduction to Bayesian statistics. John Wiley & Sons.

[15] Borgen, N., Fuglstad, G.-A., Martino, S., Richardson, S., Rue, H., & Held, L. (2018). Bayesian Methods in Epidemiology. CRC Press.

[16] Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics*, **22**, 985–996

[17] Carlin, B. P., & Louis, T. A. (2010). Bayesian methods for data analysis (3rd ed.). CRC Press.

[18] Carlin, B.P., Louis, T.A. BAYES AND EMPIRICAL BAYES METHODS FOR DATA ANALYSIS. Statistics and Computing 7, 153–154 (1997). https://doi.org/10.1023/A:1018577817064

[19] Carozzi, F. (2020). Urban density and COVID-19.

[20] Choi, P. M., Tscharke, B. J., Donner, E., O'Brien, J. W., Grant, S. C., Kaserzon, S. L., ... & Mueller, J. F. (2018). Wastewater-based epidemiology biomarkers: past, present and future. TrAC Trends in Analytical Chemistry, 105, 453-469.

[21] Cliff, A. D., & Haggett, P. (1988). Spatial Aspects of Inequality. Pion.

[22] Cliff, A. D., & Ord, J. K. (1981). Spatial processes: models & applications. (No Title).

[23] Cliff, AD and Ord, JK 1973: Spatial autocorrelation. London: Pion.

[24] Corpas-Burgos, F., Martinez-Beneito, M.A., 2020. On the use of adaptive spatial weight matrices from disease mapping multivariate analyses. Stoch Environ Res Risk Assess 34, 531–544

[25] Cressie, N. (1993) Statistics for Spatial Data. John Wiley & Sons, New York

[26] Cressie, N., & Moores, M. T. (2022). Spatial statistics. In Encyclopedia of Mathematical Geosciences (pp. 1-11). Cham: Springer International Publishing.

[27] Cressie, N., & Wikle, C. K. (2015). Statistics for spatio-temporal data. John Wiley & Sons.

[28] DF, R. (2003). Spatial data analysis: theory and practice.

[29] Diggle, P. J. (2013a). Statistical analysis of spatial and spatio-temporal point patterns. CRC press.

[30] Diggle, P. J. and Ribeiro, P. J. (2006) Model-based Geostatistics. New York: Springer

[31] Diggle, P. J., & Gibbons, R. D. (1996). Techniques for detecting extra-Poisson variation in

epidemiological data. Statistics in Medicine, 15(7-9), 639-654.

[32] Diggle, P. J., & Giorgi, E. (2019). Model-based geostatistics for global public health: methods and applications. CRC Press.

[33] Diggle, P. J., Moraga, P., Rowlingson, B., & Taylor, B. M. (2013b). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm.

[34] Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model-based geostatistics. Journal of the Royal Statistical Society Series C: Applied Statistics, 47(3), 299-350.

[35] Dupont, E., Wood, S. N., & Augustin, N. H. (2022). Spatial+: a novel approach to spatial confounding. *Biometrics*, *78*(4), 1279-1290.

[36] Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. Environmental health perspectives, 112(9), 998-1006.

[37] Faraway, J., Boxall-Clasby, J., Feil, E.J., Gibbon, M.J., Hatfeld, O.,Kasprzyk-Hordern, B., Smith, T., 2022. Challenges in realising the potential of wastewater-based epidemiology to quantitatively monitor and predict the spread of disease. Journal of Water and Health 20, 1038–1050.

[38] Fischer, M. M., & Getis, A. (Eds.). (2010). Handbook of applied spatial analysis: software tools, methods and applications (pp. 125-134). Berlin: Springer.

[39] Fuglstad, G. A., Simpson, D., Lindgren, F., & Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, *114*(525), 445-452.

[40] Gao, H., Bradley, J.R., 2019. Bayesian analysis of areal data with unknown adjacencies using the stochastic edge mixed effects model. Spatial Stat 31, 2211–6753

[41] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). Chapman and Hall/CRC.

[42] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). CRC Press.

[43] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). CRC Press.

[44] Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence, (6), 721-

741.

[45] Gomez-Rubio, V. (2020). Bayesian inference with INLA (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315175584

[46] Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

[47] Greco, F.P., Trivisano, C., 2009. A multivariate CAR model for improving the estimation of relative risks. Stat. Med. 28, 1707-1224

[48] Guinness, J., & Baak, C. W. S. (2021). Analyzing Spatial Data With Graph Convolutional Networks. Statistical Science, 36(4), 549-563.

[49] H. Rue, A. Riebler, S.H. Sørbye, J.B. Illian, D.P. Simpson, and S.H. Lindgren. Bayesian computing with INLA: a review. The Annual Review of Statistics and Its Applications, 4(1):395–421, (2016)

[50] Haining, R. (2003b). Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology. SAGE Publications.

[51] Haining, R. P., & Li, G. (2020). Regression Modelling with Spatial and Spatial-Temporal Data: A Bayesian Approach. CRC Press.

[52] Haining, R. P., & Li, G. (2020). Regression Modelling with Spatial and Spatial-Temporal Data: A Bayesian Approach. CRC Press.

[53] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[54] Hoque, N., Islam, S. S., Uddin, M. N., Arif, M., Haque, A. Z., Neogi, S. B., ... & Kabir, S. L. (2021). Prevalence, risk factors, and molecular detection of Campylobacter in farmed cattle of selected districts in Bangladesh. Pathogens, 10(3), 313.

[55] Haviland, A., 1871. The geographical distribution of disease in England and Walses. The British Med J 7, 5–6.

Haviland, A., 1888. The geographical distribution of cancerous disease in the British Isles. Lancet March 3, 412–414

[56] Hewitt J, Trowsdale S, Armstrong BA, Chapman JR, Carter KM, Croucher DM, Trent CR, Sim RE, Gilpin BJ. Sensitivity of wastewater-based epidemiology for detection of SARS-CoV-2 RNA in a low prevalence setting. Water Res. 2022 Mar 1;211:118032. doi: 10.1016/j.watres.2021.118032. Epub 2022 Jan 2. PMID: 35042077; PMCID: PMC8720482.

[57] Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.

[58] Iderus, N. H. M., Singh, S. S. L., Ghazali, S. M., Ling, C. Y., Vei, T. C., Zamri, A. S. S. M., ... & Gill, B. S. (2022). Correlation between population density and COVID-19 cases during the third wave in Malaysia: Effect of the delta variant. *International Journal of Environmental Research and Public Health*, *19*(12).

[59] Jones, D. L., Baluja, M. Q., Graham, D. W., Corbishley, A., McDonald, J. E., Malham, S. K., ... & Wilcox, M. H. (2020). Shedding of SARS-CoV-2 in feces and urine and its potential role in person-to-person transmission and the environment-based spread of COVID-19. *Science of the Total Environment*, *749*, 141364.

[60] Keefe, M. J., Ferreira, M. A., & Franck, C. T. (2018). On the formal specification of sum-zero constrained intrinsic conditional autoregressive models. Spatial statistics, 24, 54-65.

[61] Knorr-Held L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. Statistics in medicine, 19(17-18), 2555–2567. https://doi.org/10.1002/1097-0258(20000915/30)19:17/18<2555::aid-sim587>3.0.co;2-#

[62] Koch, T., & Denike, K. (2009). Crediting his critics' concerns: Remaking John Snow's map of Broad Street cholera, 1854. Social science & medicine, 69(8), 1246-1251.

[63] Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy, 52(6), 119-139.

[64] Kyriakidis, P.C., Journel, A.G. Geostatistical Space–Time Models: A Review. Mathematical Geology 31, 651–684 (1999). https://doi.org/10.1023/A:1007528426688

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

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

[67] Lawson, A. B. (1993). A Deviance Residual for Heterogeneous Spatial Poisson Processes. Biometrics, 49(3), 889–897. https://doi.org/10.2307/2532210

[68] Lawson, A. B. (1995). MCMC methods for putative pollution source problems in environmental epidemiology. Statistics in Medicine, 14(21-22), 2473-2485.

[69] Lawson, A. B. (2013). Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. CRC Press.

[70] Lawson, A. B. (2018). Bayesian disease mapping: hierarchical modeling in spatial epidemiology. CRC press.

[71] Lawson, A. B., Banerjee, S., Haining, R. P., & Ugarte, M. D. (Eds.). (2016). Handbook of spatial epidemiology. CRC press.

[72] Lee, D., Mitchell, R., 2012. Boundary detection in disease mapping studies. Biostatistics 13 (3), 415–426.

[73] Li G, Denise H, Diggle P, Grimsley J, Holmes C, James D, Jersakova R, Mole C, Nicholson G, Smith CR, Richardson S, Rowe W, Rowlingson B, Torabi F, Wade MJ, Blangiardo M. A spatio-temporal framework for modelling wastewater concentration during the COVID-19 pandemic. Environ Int. 2023 Feb;172:107765. doi: 10.1016/j.envint.2023.107765. Epub 2023 Jan 18. PMID: 36709674; PMCID: PMC9847331.

[74] Lindgren, F., Bakka, H., Bolin, D., Krainski, E., & Rue, H. (2020). A diffusion-based spatio-temporal extension of Gaussian Mat\'ern fields. *arXiv preprint arXiv:2006.04917*.

[75] Lindgren, F., Rue, H. and Lindström, J. (2011), An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73: 423-498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

[76] Lindgren, F., Rue, H. and Lindström, J. (2011), An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73: 423-498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

[77] Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). The BUGS book: A practical introduction to Bayesian analysis. CRC press.

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

[79] MacNab, Y.C., 2016a. Linear models of coregionalization for multivariate lattice data: a general framework for coregionalized multivariate CAR models. Stat. Med. 35, 3827–3850.

[80] MacNab, Y.C., 2018. Some recent work on multivariate Gaussian Markov random fields (with

discussions). TEST 27 (3),497–541

[81] Manuel, D., Amadei, C.A., Campbell, J.R., Brault, J.M., Veillard, J., 2022. Strengthening public health surveillance through wastewater testing .

[82] Martinez-Beneito, M.A., Botella-Rocamora, P., 2019. Disease Mapping: From Foundations To Multidimensional Modeling.CRC Press.

[83] Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: new features. Computational Statistics & Data Analysis, 67, 68-83.

[84] Matérn, B. (2013). Spatial variation (Vol. 36). Springer Science & Business Media.

[85] Matheron, G. (1963). Principles of geostatistics. Economic geology, 58(8), 1246-1266.

[86] Matheron, G. (1967). Eléments pour une théorie des milieux poreux. (No Title).

[87] McElreath, R. (2016). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.

[88] Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. Environ Sci Technol Lett. 2020 May 20;7(7):511-516. doi: 10.1021/acs.estlett.0c00357. PMID: 37566285.

[89] Merry, W. J. (1991). Gaussian Random Functions. CRC Press

[90] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. Journal of Chemical Physics, 21(6), 1087–1092.

[91] Moraga, P. (2023). Spatial Statistics for Data Science: Theory and Practice with R. CRC Press.

[92] Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17-23.

[93] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370-384.

[94] Nicoll, T., Jamieson, T., Price, M., & Trowsdale, S. (2022). Neighbourhood-scale wastewater-based epidemiology for COVID-19: opportunities and challenges. Journal of Hydrology (New Zealand), 61(1), 31-43.

[95] Nelson JR, Lu A, Maestre JP, Palmer EJ, Jarma D, Kinney KA, Grubesic TH, Kirisits MJ. Space-time analysis of COVID-19 cases and SARS-CoV-2 wastewater loading: A geodemographic perspective. Spat Spatiotemporal Epidemiol. 2022 Aug;42:100521. doi:

10.1016/j.sste.2022.100521. Epub 2022 May 28. PMID: 35934330; PMCID: PMC9142176.

[96] Openshaw S, Charlton ME, Wymer C, Craft A. A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. Int J Geogr Inf Syst. 1987;1:335–358. [Google Scholar]

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

[98] Ossimetha, A., Ossimetha, A., Kosar, C. M., & Rahman, M. (2021, January). Socioeconomic disparities in community mobility reduction and COVID-19 growth. In *Mayo Clinic Proceedings* (Vol. 96, No. 1, pp. 78-85). Elsevier.

[99] P.D. Congdon. Applied Bayesian Hierarchical methods. Clarendon Press, Boca Raton, (2010)

[100] Pearce, J., Witten, K., Hiscock, R., & Blakely, T. (2007). Are socially disadvantaged neighbourhoods deprived of health-related community resources? International Journal of Epidemiology, 36(2), 348–355

[101] Pearce, N., & Merletti, F. (2006). Complexity, simplicity, and epidemiology. International Journal of Epidemiology, 35(3), 515–519.

[102] Peccia, J., Zulli, A., Brackney, D. E., Grubaugh, N. D., Kaplan, E. H., Casanovas-Massana, A., … Omer, S. B. (2020). Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. Nature Biotechnology, 38(10), 1164–1167. doi:10.1038/s41587-020-0684-z

[103] Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., and Clements, A. C. A. (2008). Spatial Analysis in Epidemiology. Oxford University Press, Oxford.

[104] Phil M. Choi, Ben J. Tscharke, Erica Donner, Jake W. O'Brien, Sharon C. Grant, Sarit L. Kaserzon, Rachel Mackie, Elissa O'Malley, Nicholas D. Crosbie, Kevin V. Thomas, Jochen F. Mueller,Wastewater-based epidemiology biomarkers: Past, present and future,

[105] Rikke Ingebrigtsen, Finn Lindgren, Ingelin Steinsland, Spatial models with explanatory variables in the dependence structure, Spatial Statistics, Volume 8, 2014, Pages 20-38, ISSN 2211-6753, https://doi.org/10.1016/j.spasta.2013.06.002.

[106] Riley S, Eames K, Isham V, Mollison D, Trapman P. Five challenges for spatial epidemic models. Epidemics. 2015 Mar;10:68-71. doi: 10.1016/j.epidem.2014.07.001. Epub 2014 Jul 31. PMID: 25843387; PMCID: PMC4383807.

[107] Riley S, Eames K, Isham V, Mollison D, Trapman P. Five challenges for spatial epidemic models. Epidemics. 2015 Mar;10:68-71. doi: 10.1016/j.epidem.2014.07.001. Epub 2014 Jul 31. PMID: 25843387; PMCID: PMC4383807.

[108] Rossi, R. E., Mulla, D. J., Journel, A. G., & Franz, E. H. (1992). Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological monographs*, *62*(2), 277-314

[109] Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology, 71(2), 319-392.

[110] Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. Annual Review of Statistics and Its Application, 4, 395-421.

[111] Rushworth, A., Lee, D., Sarran, C., 2017. An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. Appl. Stat. 66 (1), 141–157

[112] Rushworth, A., Lee, D., Sarran, C., 2017. An adaptive spatiotemporal smoothing model for estimating trends and step
changes in disease risk. Appl. Stat. 66 (1), 141–157

[113] Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., & Rue, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. Biometrika, 103(1), 49-70.

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

[115] SPENCER, S. E. F., et al. “The Spatial and Temporal Determinants of Campylobacteriosis Notifications in New Zealand, 2001—2007.” Epidemiology and Infection, vol. 140, no. 9, 2012, pp. 1663–77. JSTOR, http://www.jstor.org/stable/23254492. Accessed 27 Feb. 2024.

[116] Stein, M. L. (2012). Interpolation of Spatial Data: Some Theory for Kriging. Springer.

[117] Tobler, W. R. (1993). Three presentations on geographical analysis and modeling. National Center for Geographic Information and Analysis, Technical Report, 93(1).

[118] TrAC Trends in Analytical Chemistry,Volume 105, 2018, Pages 453-469, ISSN 0165-9936,

[119] Use of geographic information systems in epidemiology (GIS-Epi). Bull PAHO, 1996,17:1-

6.

[120] Vaughan L, Zhang M, Gu H, Rose JB, Naughton CC, Medema G, Allan V, Roiko A, Blackall L, Zamyadi A. An exploration of challenges associated with machine learning for time series forecasting of COVID-19 community spread using wastewater-based epidemiological data. Sci Total Environ. 2023 Feb 1;858(Pt 1):159748. doi: 10.1016/j.scitotenv.2022.159748. Epub 2022 Oct 25. PMID: 36306840; PMCID: PMC9597519.

[121] World Health Organization. (2021). WHO-convened global study of origins of SARS-CoV-2: China Part.

[122] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. Biometrika, 57(1):97–109, (1970).

[123] Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41**, 434–449.
Whittle, P. (1963) Stochastic processes in several dimensions. *Bull. Inst. Int. Statist.*, **40**, 974–994

[124] Wolfel, ¨ R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Muller, M.A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., Hoelscher, M., Bleicker, T.,Brünink, S., Schneider, J., Ehmann, R., Zwirglmaier, K., Drosten, C., Wendtner, C., 2020. Virological assessment of hospitalized patients with COVID-2019. Nature 581 (7809), 465–469 https://doi.org/10.1038/s41586-020-2196-x.

[125] Yuan, Yuan, Bachl, E. F, Lindgren, Finn, Borchers, L. D, Illian, B. J, Buckland, T. S, Rue, Håvard, Gerrodette, Tim (2017). “Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales.” Ann. Appl. Stat., 11(4), 2270–2297. doi:10.1214/17-AOAS1078.