Bayesian Modeling of Health Data in Space and Time

Cici Xi Chen Bauer

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Jon C. Wakefield, Chair

Steve Self

Vladimir Minin

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Bayesian Modeling of Health Data in Space and Time

Cici Xi Chen Bauer

Chair of the Supervisory Committee:
Professor Jon C. Wakefield
Statistics

In recent years spatial-temporal modeling has become increasingly popular in the field of public health and epidemiology. Motivated by two datasets, we address three issues in the Bayesian modeling of health data in space and time.

The first motivating example is provided by data from the Behavioral Risk Factor Surveillance System (BRFSS). In a survey sampling context we develop a method for incorporating the sampling weights in a complex survey design, within a spatial smoothing model. A simulation study is presented to demonstrate the performance of the proposed approach and to compare results from models with and without the sampling weights. The results show that mean squared error can be greatly reduced using the proposed model, when compared with standard approaches. Bias reduction occurs through the incorporation of sampling weights, with variance reduction being achieved through hierarchical spatial smoothing.

The second motivating example concerns the surveillance data for Hand-Foot-Mouth disease (HFMD) collected in China between 2009 and 2010. The overall strategy we take is to decompose the log relative risk of disease into three components: a large-scale temporal trend, a large-scale spatial trend and a spatial-temporal interaction. We fit the model in a Bayesian framework and the structure of the interaction between space and time is imposed through a prior on the coefficients of the basis functions, which are constructed as a tensor product of cubic B-splines. This model is amenable to prediction through the use

of Gaussian Markov Random Field (GMRF) space-time priors.

Finally, we consider the situation in which a disease can be caused by multiple virus strains. The data we analyze again concern HFMD in China and contain total disease counts along with a limited amount of strain-specific information gathered on a subset of individuals. We propose a Bayesian hierarchical model that provides a coherent approach to estimating the total number of cases by strain. When data is available for multiple areas and time points, the spatial and temporal variability can again be modeled via smoothing priors. The model can also be extended to accommodate multiple virus strains or multiple clinically-diagnosed severity categories.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Any individual accomplishment worthy of even the most modest of accolades can only be achieved with the help and support of near countless people. I am no different and my past is filled with those who have cheered me on, given me confidence, and eased my doubts. Although all of these people have my eternal gratitude, I would like now to thank some of the most important.

To my advisor, Dr. Jon Wakefield, thank you for your patience and good humor. The value of your critical insights into my work cannot be overstated and I can honestly say I look forward to future collaboration.

To Dr. Steve Self, thank you for providing me access to such an important and interesting dataset. None of this would have been possible without your generous funding of my research with the China CDC project.

Graduate school is a uniquely stressful and rewarding experience and there are far too many individual friends to list, but our time together commiserating on our failures and celebrating our triumphs has made my time here truly unforgettable, so to all my fellow grad students, just plain thanks.

And to my family, thank you for all your support and guidance through the years. You will always be my true source of inspiration and joy and I love you all dearly.

DEDICATION

To Chouchou, QQ and Paopao.

Chapter 1

INTRODUCTION AND MOTIVATING EXAMPLES

1.1 Background

In recent years spatial-temporal modeling has become increasingly popular in the field of public health and epidemiology. The development and interest in spatial-temporal models can be attributed to the availability of health and population data with geographic and temporal indices, the widespread accessibility of Geographical Information Systems (GIS) and various public health databases, and public and scientific interest in questions with a spatio-temporal component. Our motivation for the Bayesian spatial-temporal modeling of public health data comes from one dataset on diabetes in the 2006 Washington Behavioral Risk Factor Surveillance System (BRFSS) and another dataset on Hand-Foot-Mouth disease (HFMD) in China between 2009 and 2010. In both examples, the number of cases with the outcome of interest is collected from areas within a certain region so that we have spatially aggregated data. We are interested in investigating the extent of spatial variability in the data and describing the nature of any existing spatial pattern. Further, for the HFMD data there is limited strain-specific information that we wish to utilize to understand how different strains evolve and interact in space and time. Therefore, the models we develop for these two datasets share a common component, which is to account for spatial variability. However, there are also different and unique aspects of the two examples: in the BRFSS, the data are sampled with unequal probabilities and therefore we should address the sampling scheme in models with which we make inference. On the other hand, the China HFMD data is collected over two years from different administrative divisions. Therefore, in addition to the spatial trend, we are also interested in the temporal trend and the spatial-temporal interaction. Such difference requires us to develop a specific model for each example, and these models should be able to accommodate the particular characteristics of each dataset.

In the remainder of this chapter, we first give detailed descriptions in Section 1.2 of our

two motivating examples and present some exploratory analyses for each dataset. Subsequently in Section 1.3 we give a literature review of the spatial-temporal modeling. Finally, we describe the outline of this dissertation in Section 1.4.

1.2 Two Motivating Examples

1.2.1 Behavioral Risk Factor Surveillance System (BRFSS) Data

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone health survey conducted by the Centers for Disease Control and Prevention (CDC) that has tracked health conditions and risk behaviors in the United States and its territories since 1984. The objective of the BRFSS is to provide uniform, state-specific estimates of the prevalence of risk behaviors. In the BRFSS survey, interviewees (who are 18 years or older) are asked a series of questions on their health behaviors and provide general demographic information, such as age, race, gender and the zip code in which they live. In this paper we focus on the survey conducted in Washington State in 2006, and on the question, "Have you ever been told you have diabetes?", with interviewees responding with either a "Yes"or a "No". Therefore the response variable is a binary indicator of the presence of diabetes, and our objective is to estimate the number of 18 or older individuals with diabetes, by zip code, in Washington state. The CDC currently publishes coarser, county-level prevalence estimates using the model of Malec et al. (1997), at `http://apps.nccd.cdc.gov/DDTSTRS/`.

In 2006, the survey used land-lines only (from 2008, a small number of cell phones supplement the landlines), and a disproportionate stratified random sample scheme with stratification by county and "phone likelihood". Under this scheme in each county, based on previous surveys, blocks of 100 telephone numbers are classified into strata that are either "likely" or "unlikely" to yield residential numbers. Telephone numbers in the "likely" strata are then sampled at a higher rate than their "unlikely" counterparts. Once a number is reached the number of eligible adults (aged 18 or over) is determined, and one of these is randomly selected for interview. The sample weight, `Sample Wt`, is then calculated as the product of four terms:

$$\texttt{Sample Wt} = \texttt{Strat Wt} \times \frac{1}{\texttt{No Telephones}} \times \texttt{No Adults} \times \texttt{Post Strat Wt} \qquad (1.1)$$

Table 1.1: Summary statistics for population data and 2006 Washington BRFSS diabetes data, across zip codes.

	Mean	*Std. Dev.*	*Median*	*Min*	*Max*
Population	12570	12931	7208	11	55700
Sample sizes	46.9	55	30	1	384
Diabetes cases	4.6	6	3	0	38

where `Strat Wt` is the inverse probability of a "likely" or "unlikely" stratum being selected in a particular county, `No Telephones` represents the number of residential telephones in the respondent's household, `No Adults` is the number of adults in the household, and `Post Strat Wt` is the post-stratification correction factor. The latter is given by the number of people in strata defined by gender and age, with the 7 age groups 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75+. The other source of data we use are zip code population estimates for 2006.

Basic summary statistics for the 2006 Washington BRFSS diabetes data by 498 zip codes, and a map of the observed sample size are presented in Table 1.1 and Figure 1.1 respectively. A total of $23,379$ individuals answered the diabetes question in the survey. There is large variability in each of the populations, sample sizes and numbers of diabetes cases across zip codes. About 20% of the areas have sample sizes less than 10, so the diabetes prevalence estimates are highly unstable in these areas in particular. A map of the observed diabetes prevalence by zip codes is presented in Figure 1.2.

In this example, spatial models can be employed to provide more reliable estimates for the diabetes prevalence estimation. This is because, in spatial models, the estimation of one area "borrows" information from the neighboring areas, rather than uses data from that area alone. The application of spatial models has been popular in small area estimation. However, the sampling weights that are required to reflect complex surveys are rarely considered in these models. Motivated by the BRFSS data, we aim to develop a spatial model which acknowledges the sampling weights of the binary data when estimating, for example, small area proportions or predicting small area counts.

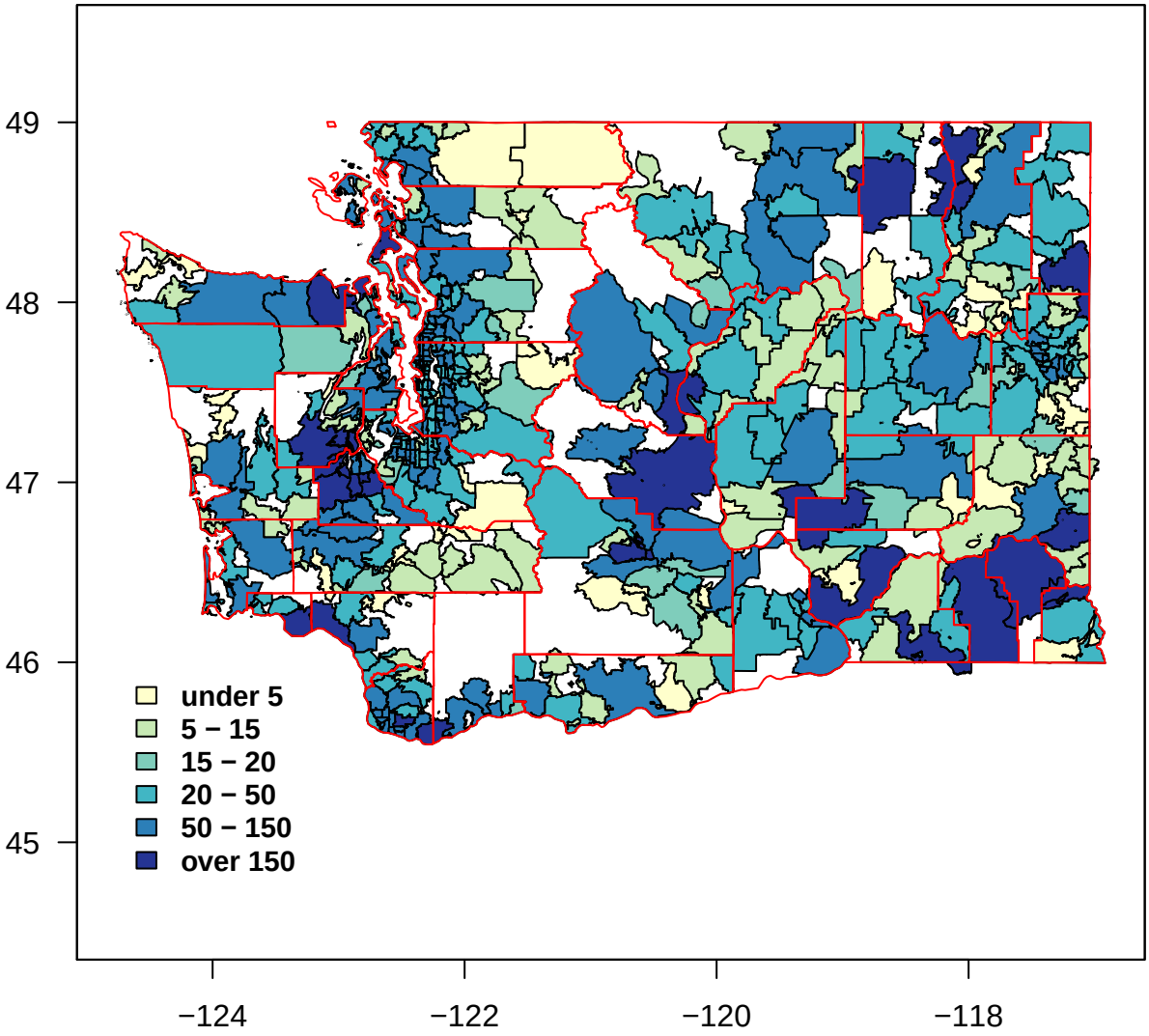


Figure 1.1: Observed diabetes sample size at zip code level using 2006 Washington BRFSS data.

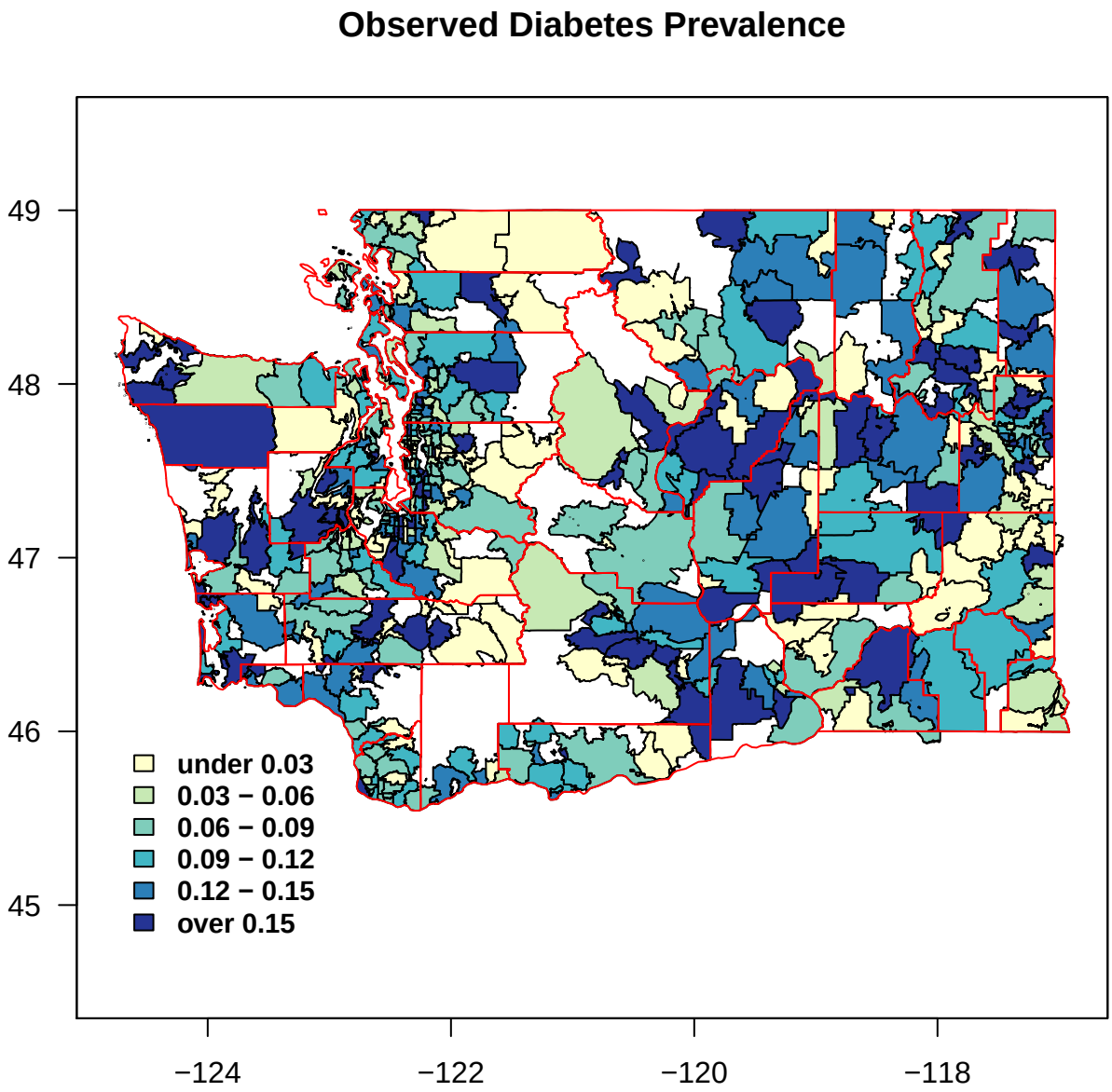


Figure 1.2: Observed diabetes prevalence at zip code level using 2006 Washington BRFSS data.

1.2.2 China Hand-Foot-Mouth Disease (HFMD) data

Hand-Foot-Mouth disease (HFMD) is an acute contagious viral infection that has caused large-scale outbreaks in Asia during the past decade (Tong and Bible, 2009). HFMD is caused by enterovirus (EV) pathogens and often involves mild or moderate symptoms such as fever, oral ulcer or rashes on the hand and foot. Very little is known about the etiology of the enterovirus, the factors associated with its spread, or an effective means of public health intervention. Answers to these questions will greatly benefit authorities charged with policy making to control this infectious disease.

Since the severe acute respiratory syndrome (SARS) epidemic outbreak in China in 2003, China CDC has established a disease surveillance system which regulates the reporting of 39 notifiable infectious diseases including EV71-related HFMD. The purpose of the surveillance system is to monitor epidemics of infectious diseases, identify high case occurrence areas, predict and control epidemics, and provide information for formulating policy. The current surveillance system covers the entire population living in all provinces, prefectures, and counties in mainland China. Over successive years, HFMD in China has intensified in severity. In particular, $1,155,324$ cases of HFMD were reported with $13,834$ cases suffering severe neurological damage and 353 deaths in 2009, and 1,960,929 cases were reported with $30,790$ severe cases and 952 deaths in 2010.

Depending on the clinical symptoms, a HFMD case can be classified as mild or severe. In some regions worst hit by the HFMD, a clinically diagnosed severe case is required to receive a lab test which isolates the enterovirus. After the lab test, the specific strain of the enterovirus that causes the HFMD can be identified as EV71, Coxsackie A16, or other.

Each reported case from the Chinese infectious disease surveillance system consists of the patient's geographical location (up to the level of township), gender, age, the symptom onset date, and the time of death (if applicable). Therefore, the disease surveillance system from China provides an extensive data resource for spatial-temporal modeling. In addition, the lab test information, combined with the surveillance data, provides the opportunity to investigate both the strain-specific HFMD epidemic and how the different strains may interact with each other.

When we deal with aggregated data, we should take account of confounders, in particular the differences in age, gender and race distributions across areas and time points. Ignoring such difference may lead to inappropriate inference. An example of Simpson' paradox in a space-time context was presented by Knorr-Held and Besag (1998). One way to include age, gender or race information in a Poisson model is to use the expected number of cases calculated from indirect standardization. Because this is the approach we take throughout the dissertation when we use Poisson models, we next give a description of this procedure.

Indirect Standardization

For area $i = 1, \ldots, I$ and time period $t = 1, \ldots, T$, we assume the total number of HFMD cases Y_{it} follow a Poisson distribution with relative risk μ_{it}

$$Y_{it}|\mu_{it} \sim \text{Poisson}(E_{it}\mu_{it}), \qquad i = 1, \ldots, I, \quad t = 1, \ldots, T, \tag{1.2}$$

where E_{it} is the expected number of cases in area i and time period t evaluated using estimated age by gender reference risks. Besides two gender groups for HFMD, we choose 5 age groups for the standardization, with the age group defined as < 1, $[1, 3)$, $[3, 6)$, $[6, 10)$ years and 10+ years old. Therefore, we have a total of 10 age-gender strata.

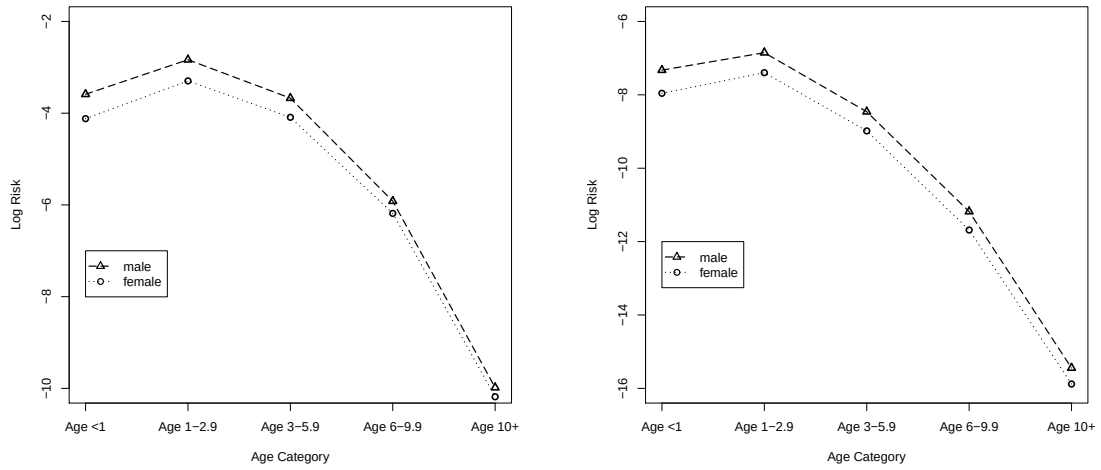
To calculate the expected number of cases in area i we first calculate the reference rate in stratum j, which for internally standardized rates is defined as:

$$\hat{q}_j = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T} Y_{ij}}{\sum_{i=1}^{I} N_{ij}}, \qquad j = 1, \ldots, J, \tag{1.3}$$

where N_{ij} is the population in area i and stratum j and is assumed to be constant over the study period. The reference rate per unit of time can be defined as $\hat{q}_{jt} = \hat{q}_j/T$, where $T = 104$ if we work with weekly events.

Figure 1.3 shows the weekly log risks by age-gender stratum for all HFMD cases and severe cases in China, with combined data from 2009 and 2010. The overall patterns in the two plots are similar: the risks are higher for males than females for every age group, and the risk is the highest for age group $[1, 3)$. For children older than 3, the risks of contracting HFMD decrease as they get older.

To obtain the expected number of cases E_i in area i, we make the assumption that the probability of contracting the disease in area i and stratum j is proportional to the reference



(a) all cases

(b) severe cases

Figure 1.3: Log risks by age-gender stratum for all HFMD cases and severe cases in China, with combined data from 2009 and 2010.

risk q_j. This proportionality assumption removes the need to estimate area×stratum-specific risks, which typically cannot be achieved as there are insufficient data. However, care needs to be taken when making such an assumption as it may not be appropriate in some applications.

Following the assumption, we obtain the expected number of cases as $E_i = \sum_{j=1}^{J} N_{ij}\hat{q}_j$. Note that here we treat the expected number of cases as constant for each time point t, i.e., $E_{it} = E_i$. It is easy to see that in (1.2), the maximum likelihood estimation (MLE) $\hat{\mu}_{it}$ of μ_{it} is Y_{it}/E_{it}, which is called the *standardized mortality/morbidity ratio (SMR)*. This is an important quantity in disease mapping: $\hat{\mu}_{it}$ greater than 1 implies that we observe more cases than expected and indicates elevated spatial/temporal risk. The variance of the SMR is μ_{it}/E_{it}, so that areas with small E_{it} have high associated variance. Therefore, inference made from the observed SMR is highly unreliable and we need spatial-temporal models that can provide more reliable estimation.

In Figures 1.4 and 1.5 we show the map of the log SMR of the China HFMD data aggregated over 2009 and 2010, and the temporal trend in the log SMR aggregated over

all prefectures in China. In the map we see clusters of areas with high log SMR along the east coast of China. As for the temporal trend, there appears to be a cycle of a high peak around week 18, followed by a second but much smaller peak around week 40 in both years. A model with spatial and temporal components is clearly needed for analyzing the China HFMD data.

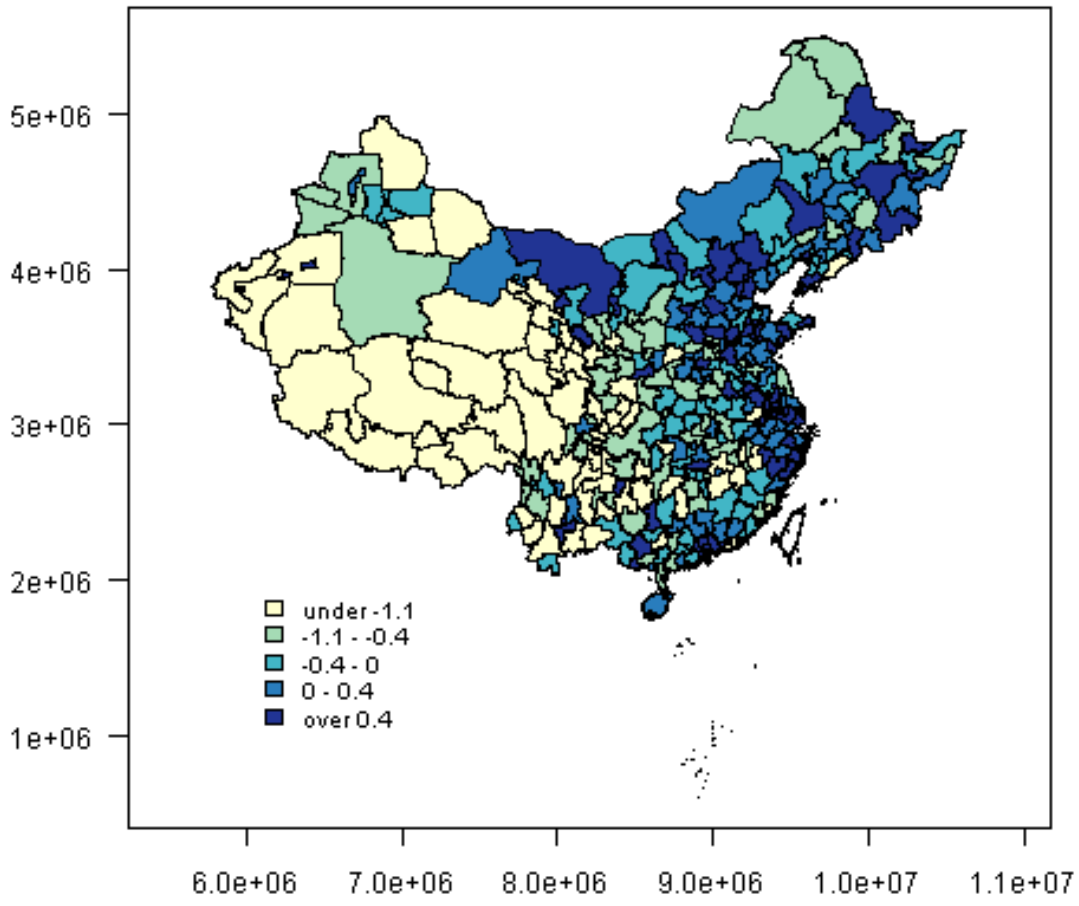


Figure 1.4: The spatial pattern of log SMR using China HFMD data between 2009 and 2010 aggregated over all weeks.

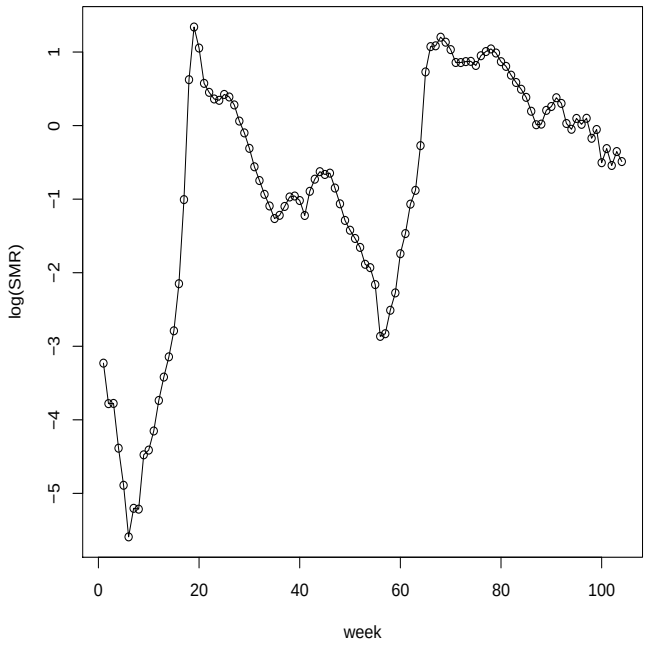


Figure 1.5: The temporal pattern of log SMR using China HFMD data between 2009 and 2010 aggregated over all prefectures.

1.3 Bayesian Models for Space-time Aggregated Data

Many of the models suggested for space-time aggregated data have been developed for disease mapping, with the purpose being the description of the geographical and temporal pattern of disease burden. The benchmark for describing the spatial pattern is probably the popular intrinsic conditional autoregressive (ICAR) model introduced by Besag et al. (1991). In this model, the disease risk was decomposed as the sum of two components. The first component was a spatially dependent effect, with the dependency requiring the specification of a neighborhood structure. The second component was an independent non-spatial effect which allows the different regions to have distinct behaviors. However, the ICAR model in Besag et al. (1991) did not consider the temporal evolution of the risks and therefore only provides static risk estimates in time.

To include the temporal component in the model, Bernardinelli et al. (1995) proposed an ecological regression model with area-specific intercept and temporal trend. Both the

intercept and the temporal trend were treated as random effects, with ICAR models as the priors. In this approach, the temporal trend of the disease risk for every region depended on the temporal trend of its neighbors, but only a linear time trend was allowed. Assunçaõ et al. (2001) took a similar approach but allowed a quadratic term for the time trend. In both cases, the risk evolution over time took a parametric form, which may be too strong an assumption.

A different approach was taken by Waller et al. (1997) and Xia and Carlin (1998), where the risks for every time period was modeled as time-independent spatial random effect with a time-specific precision parameter for the spatial variability. In Knorr-Held (2000), both spatial and temporal trends were treated as the random effects with structured and unstructured components. The interaction between space and time can take four different forms, based on the combinations of structured or unstructured time and space assumptions.

For modeling multiple diseases, Richardson et al. (2006) proposed a temporally independent multivariate conditional autoregressive (MCAR) model for lung cancer mortality, jointly studied for both males and females. In Knorr-Held and Best (2001), the authors proposed a shared compartment model where the underlying risk for each disease was separated into a shared and a disease-specific component.

In recent years, spline models have become increasingly popular in the spatial-temporal setting. For example, MacNab and Dean (2002) used a spline model for the temporal variation in each region, in a study of the infant mortality rate. This work was later extended in MacNab and Gustafson (2007) to a broader class of regression spline models with spatial-temporal relative risks being modeled as spatially varying or randomly varying regression B-splines. Other types of spline bases may be adopted with, for example, Ruppert and Wand (2000) and Wand (2000) using truncated power functions. Kammann and Wand (2003) introduced so-called geoadditive models, which combined additive models and geostatistical kriging models into a mixed model representation. This model was applied to study the geographical variability of birth weight data in Upper Cape Code, Massachusetts. French and Wand (2004) used a very similar model to produce a cancer incident map in the same region, with the focus being on how to handle missing covariate values.

The spatial-temporal models developed for non-infectious diseases have also been ap-

plied to infectious diseases. However, the usual disease mapping methods are often used as purely descriptive models for infectious diseases (Lawson (2009), Chapter 11), see for example, Mugglin et al. (2002) for flu epidemics and Knorr-Held and Richardson (2003) for meningococcal disease. Held et al. (2005) proposed a framework for the analysis of infectious disease surveillance counts, where in its simplest form, the model can be viewed as a Poisson branching process model with immigration. Let Y_{it} be the number of disease counts in area i and time period t. The Held et al. (2005) model assumed Y_{it} follow a Poisson distribution with mean risk being decomposed into an endemic component λ_{it} and an epidemic component μ_{it}. The endemic component included a linear trend and a cyclic seasonal trend:

$$\log(\lambda_{it}) = \alpha_i + \beta t + \sum_{s=1}^{S} \big(\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)\big),$$

where S is the number of harmonics to include and ω_s are the Fourier frequencies, for example, $\omega_s = 2s\pi/52$ for weekly data. The epidemic component was modeled as

$$\mu_{it} = \lambda y_{i,t-1} + \phi \sum_{j \sim i} y_{j,t-1},$$

where $j \sim i$ indicates that area j and i are neighbors. So the epidemic component of the mean had a contribution from the number of counts in the previous time period in the same area and a contribution from neighboring areas. The parameters λ and ϕ determine the relative contributions of cases in the same and in neighboring regions. Paul et al. (2008) adopted the same idea of separating the mean risk into endemic and epidemic components, but used a negative binomial model to allow for overdispersion. They also extended the model to have area-specific λ_i and ϕ_i, and different weights contributed by the neighboring areas in the epidemic component. This model was applied to influenza and meningococcal disease data. Wang et al. (2011) modeled weekly counts of the HFMD in prefectures of China using a similar negative binomial model, with the mean decomposed into epidemic and endemic components similar to Paul et al. (2008), but both components were modeled as functions of area-level covariates.

1.4 Organization of this Dissertation

In this dissertation, we aim to develop spatial-temporal models that can answer the questions raised by the motivating examples. The organization is as follows: in Chapter 2, we describe Gaussian Markov Random Fields (GMRFs) and their connections to common spatial and temporal models. This chapter serves as the theoretical foundations of the models we develop. In Chapter 3 we introduce a Bayesian hierarchical approach that is appropriate for data that have associated complex sampling weights. The performance of our approach is demonstrated via a series of simulations. We also present results for the Washington BRFSS data using the approach we develop. Chapter 4 presents a Bayesian penalized spline model that can be used to describe the spatial and temporal pattern of an infectious disease, with different space-time interaction models being considered. In Chapter 5, we introduce a Bayesian hierarchial model that combines both the surveillance data and the lab test data to obtain strain-specific estimates of the counts of HFMD in China. Finally, in Chapter 6 we state the conclusions and discuss future work.

Chapter 2

SPATIAL-TEMPORAL MODELS AND THEIR CONNECTIONS TO GAUSSIAN MARKOV RANDOM FIELD (GMRF) MODELS

Many spatial, temporal or spatial-temporal models assume that a set of latent variables have a Gaussian distribution, for example, the popular intrinsic conditional autoregressive (ICAR) model introduced in Besag et al. (1991) for spatial modeling takes this form. Such models belong to the general family of Gaussian Markov Random Fields (GMRFs). In this chapter we give details about GMRFs, and give their connections to popular models in spatial-temporal modeling. We also describe different sampling schemes appropriate for GMRFs when using Markov Chain Monte Carlo (MCMC) computation methods in Bayesian analysis. The GMRFs are a foundation for the development of penalized spline models that will be carried out in Chapter 4.

The chapter is organized as follows. In Section 2.1, we give a general description of GMRFs. In Section 2.2, we focus on a popular subclass of GMRFs, which have been used extensively in spatial-temporal modeling and describe their connections to GMRFs. In Section 2.3, we pay attention to the computation aspect of intrinsic GMRFs and describe different MCMC sampling schemes for use in Bayesian analysis. In Section 2.4, we briefly introduce a newly-developed, alternative method to MCMC for making Bayesian inference, which is known as the Integrated Nested Laplace Approximation (INLA) method. We then analyze data on lung cancer incidence in the London Health Authority region to compare the results of different MCMC sampling schemes and the INLA method. The comparison is summarized in Section 2.5.

2.1 Gaussian Markov Random Fields (GMRFs)

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be an n-dimensional normal variable. The GMRFs we focus on are built upon the assumption of conditional independent properties of elements of $\boldsymbol{x}$: if x_i and x_j are conditionally independent given $\boldsymbol{x}_{-ij}$, $i \neq j$, we write $x_i \perp x_j \mid \boldsymbol{x}_{-ij}$ where $\boldsymbol{x}_{-ij}$

indicates elements of $\boldsymbol{x}$ excluding x_i and x_j. This conditional independence can be fully specified by the precision matrix $\boldsymbol{Q}$ of $\boldsymbol{x}$: $\boldsymbol{Q}_{ij} = 0$ iff $x_i \perp x_j \mid \boldsymbol{x}_{-ij}$ (Rue and Held, 2005). With such a precision matrix $\boldsymbol{Q}$ and mean vector $\boldsymbol{\mu}$, the density of a random variable $\boldsymbol{x}$ from a GMRF is

$$\pi(\boldsymbol{x}) = \frac{|\boldsymbol{Q}|^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{2.1}$$

Note that the matrix $\boldsymbol{Q}$ here must be positive definite and of full rank. In many applications, GMRFs are used for latent variables which have zero mean. Then the distribution in (2.1) is simplified to

$$\pi(\boldsymbol{x}) = \frac{|\boldsymbol{Q}|^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}\boldsymbol{x}\right) \tag{2.2}$$

Sampling from a GMRF with full-rank precision matrix $\boldsymbol{Q}$ in (2.2) involves the following three steps:

1. Bandwidth reduction of the precision matrix $\boldsymbol{Q}$. The bandwidth of a matrix is defined as the maximum value of $|i-j|$ such that the ij-th entry is nonzero. Bandwidth reduction can be achieved by permuting the row (or column) indices of the matrix $\boldsymbol{Q}$. Reducing bandwidth is not a prerequisite for sampling from a GMRF. However, this step can highly accelerate the computation speed for the next step.

2. Cholesky decomposition of the precision matrix $\boldsymbol{Q}$

   $$\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T,$$

   where $\boldsymbol{L}$ is the lower triangular factor of the Cholesky decomposition.

3. Solve $\boldsymbol{x}$ from $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\boldsymbol{I}_n$ is the n-by-n identity matrix. The resulting $\boldsymbol{x}$ is one sample from the n-dimensional GMRF with mean zero and precision matrix $\boldsymbol{Q}$.

This algorithm is particulary efficient in repeated sampling from a GMRF, as the bandwidth reduction and Cholesky decomposition of the precision matrix $\boldsymbol{Q}$ only needs to be computed once. Then, for each set of new samples, we just generate a new $\boldsymbol{z}$ from the standard normal distribution and back solve the linear equation in step 3.

2.2 Intrinsic Gaussian Markov Random Fields (IGMRFs)

When the precision matrix $\boldsymbol{Q}$ does not have full rank, GMRFs are called the intrinsic Gaussian Markov Random Fields (IGMRFs). Many popular spatial-temporal models in the epidemiology literature are IGMRFs. We describe one such model, the intrinsic conditional autoregressive (ICAR) model introduced in Besag et al. (1991) that is formulated as follows:

- At the first stage we have the data model (i.e., the likelihood)

$$\begin{aligned} Y_i|\mu_i &\sim \text{Poisson}(E_i\mu_i), \\ \log(\mu_i) &= \alpha + u_i + v_i. \end{aligned} \tag{2.3}$$

  Here, Y_i is the number of cases with the outcome of interest in area i ($i = 1, 2, \ldots, I$), and E_i is the expected number of cases. The relative risk μ_i is decomposed into three components: the overall level α, a spatially-structured random effect u_i, and a non-spatial random effect v_i.

- At the second stage we assign an ICAR prior to the spatial component u_i, and an i.i.d normal prior to the non-spatial component v_i

$$\begin{aligned} u_i|u_j, j \neq i &\sim \text{N}\left(\frac{1}{k_i}\sum_{j\sim i} u_j, \frac{\sigma_u^2}{k_i}\right), \\ v_i|\tau_v &\sim_{iid} \text{N}(0, \tau_v^{-1}), \end{aligned} \tag{2.4}$$

  where $j \sim i$ denotes that area j and i are neighbors and k_i is the number of neighbors of area i.

- At the last stage, the precision parameter τ_u and τ_v are assigned Gamma distributions as the hyperpriors. The intercept parameter α is assigned an improper flat prior:

$$\begin{aligned} \tau_u &\sim \text{Gamma}(a_u, b_u), \\ \tau_v &\sim \text{Gamma}(a_v, b_v), \\ \alpha &\sim flat. \end{aligned} \tag{2.5}$$

An alternative way to write the ICAR prior distribution of $\boldsymbol{u}$ is to use the joint distribution:

$$\pi(\boldsymbol{u}|\tau_u) \propto {\tau_u}^{(I-1)/2} \exp\left(-\frac{\tau_u}{2}\sum_{i\sim j}(u_i - u_j)^2\right). \tag{2.6}$$

Notice that this distribution is only defined by the pairwise difference of the elements of $\boldsymbol{u}$ and so the overall level is unspecified.

In this example, the prior used for the spatial component $\boldsymbol{u}$ can be represented in the same form as the GMRF in (2.2) via

$$\pi(\boldsymbol{u}|\tau_u) \propto {\tau_u}^{(I-1)/2} \exp\left(-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{Q}\boldsymbol{u}\right), \tag{2.7}$$

where $\boldsymbol{Q} = \tau_u\boldsymbol{K}$ with the structure matrix $\boldsymbol{K}$ given by

$$K_{ij} = \begin{cases} k_i & i = j, \\ -1 & i \sim j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

In the specification of $\boldsymbol{K}$, k_i is again the number of neighbors of area i.

Note that the structure matrix $\boldsymbol{K}$ has rank $I - 1$ because the sum of each row (or column) is zero, hence the ICAR prior distribution used for the spatial random effect $\boldsymbol{u}$ is improper (i.e., it is not a legitimate probability distribution). To make the random effects u_i identifiable in the posterior distribution, we need to either exclude the intercept, or include an intercept in the model with a flat prior and a sum-to-zero constraint on the spatial random effect. Both approaches reduce the number of parameters by 1, which is required for identifiability. The second approach was recommended in Besag et al. (1991) and we follow their recommendation throughout the dissertation when we use the ICAR prior.

Some popular models for analyzing temporal trends can also be represented as GMRFs. For example, let Y_t be the number of cases with the outcome of interest at equally-spaced time points t ($t = 1, 2, \ldots, T$), and E_t be the expected number of cases. A temporal model can be formulated as follows:

- At the first stage we assume

$$\begin{aligned} Y_t|\mu_t &\sim \text{Poisson}(E_t\mu_t), \\ \log(\mu_t) &= \alpha + \gamma_t + \phi_t, \end{aligned} \tag{2.9}$$

  with a temporally-structured random effect γ_t, and a non-temporal random effect ϕ_t.

- At the second stage we assign a first-order random walk prior, or in abbreviation a RW(1) prior, to the temporal random effect γ_t. The non-temporal component ϕ_t is assigned an i.i.d normal prior

$$\begin{aligned} \gamma_t|\gamma_{t-1} &\sim \text{N}\left(\gamma_{t-1}, \sigma_\gamma^2\right), \\ \phi_t|\tau_\phi &\sim_{iid} \text{N}(0, \tau_\phi^{-1}). \end{aligned} \tag{2.10}$$

- At the last stage, the precision parameter τ_γ and τ_ϕ are assigned Gamma distributions as the hyperpriors. The intercept parameter α is assigned an improper flat prior.

The joint distribution of $\boldsymbol{\gamma}$ with the RW(1) prior is

$$\pi(\boldsymbol{\gamma}|\tau_\gamma) \propto {\tau_\gamma}^{(T-1)/2} \exp\Big(-\frac{\tau_\gamma}{2} \sum_{t=2}^{T} (\gamma_t - \gamma_{t-1})^2 \Big). \tag{2.11}$$

Notice that the RW(1) model is also a limiting form of a first-order autoregressive model, or in abbreviation an AR(1) model.

The RW(1) prior employed here can be written in the form of GMRF as in (2.2) via

$$\pi(\boldsymbol{\gamma}|\tau_\gamma) \propto {\tau_\gamma}^{(T-1)/2} \exp\Big(-\frac{1}{2}\boldsymbol{\gamma}^T \boldsymbol{Q} \boldsymbol{\gamma} \Big),$$

where $\boldsymbol{Q} = \tau_\gamma \boldsymbol{K}$ with the structure matrix $\boldsymbol{K}$ given by

$$\boldsymbol{K} = \begin{pmatrix} 1 & -1 & & & & & & \\ & 1 & -2 & 1 & & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & & 1 & -2 & 1 \\ & & & & & & 1 & -1 \end{pmatrix}, \tag{2.12}$$

where the blank spaces indicate zero entries. The rank of the matrix $\boldsymbol{K}$ is $T-1$ because the overall level of this temporal effect is unspecified.

Sometimes we prefer the second-order random walk (or RW(2)) to RW(1) for stronger temporal smoothing. Modification of a RW(1) model to a RW(2) model for structured temporal component $\boldsymbol{\gamma}$ is straightforward and gives

$$\pi(\boldsymbol{\gamma}|\tau_\gamma) \propto {\tau_\gamma}^{(T-2)/2} \exp\Big(-\frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{Q}\boldsymbol{\gamma}\Big),$$

where $\boldsymbol{Q} = \tau_\gamma \boldsymbol{K}$ with the structure matrix $\boldsymbol{K}$ given by

$$\boldsymbol{K} = \begin{pmatrix} 1 & -2 & 1 & & & & & & & \\ & -2 & 5 & -4 & 1 & & & & & \\ & & 1 & -4 & 6 & -4 & 1 & & & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & & & 1 & -4 & 5 & -2 \\ & & & & & & & 1 & -2 & 1 \end{pmatrix}, \tag{2.13}$$

where $\boldsymbol{Q}$ has rank $T-2$. The rank 2 deficiency here indicates that the RW(2) distribution is invariant to the addition of any line $\alpha + \boldsymbol{\gamma t}$ to $\boldsymbol{\gamma}$. To make the temporal random effects γ_t identifiable, we need to put both $\sum_t \gamma_t = 0$ and $\sum_t t\gamma_t = 0$ constraints on $\boldsymbol{\gamma}$ and include an intercept α in the model.

In spatial-temporal modeling, besides components to account for spatial and temporal correlations, very often we also need spatial-temporal interactions. Some popular spatial-temporal interaction models can also be represented in the form of (2.2). For example, in Knorr-Held (2000), the number of cases Y_{it} in area i and time t is assumed to follow a Poisson distribution with parameter μ_{it}

$$Y_{it}|\mu_{it} \sim \text{Poisson}(E_{it}\mu_{it}),$$

where E_{it} is the expected number of cases. The log linear predictor μ_{it} is then decomposed additively as

$$\log(\eta_{it}) = \alpha + u_i + v_i + \gamma_t + \phi_t + \delta_{it} \tag{2.14}$$

In (2.14), u_i and v_i are the spatial and non-spatial random effects, and are assigned priors using the models in (2.4). The temporal components γ_t and the non-temporal random effect ϕ_t are assigned priors using the models in (2.10). A model containing only the terms $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ is called a the main effect model. Although they provide information on the spatial-temporal variation in the log relative risk, main effect models lack the ability to reflect the fact that epidemics may have different temporal dynamics in different regions. Therefore, a space-time interaction term $\boldsymbol{\delta}$ is needed. The form of $\boldsymbol{\delta}$ can be modeled as the combination of one of the spatial main effects (i.e. $\boldsymbol{u}$ and $\boldsymbol{v}$) and one of the temporal main effects (i.e. $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$), which results in four types of interaction term $\boldsymbol{\delta}$, as suggested by Clayton (1996):

- Type 1 interaction assumes that after accounting for the main effects, the residuals do not have structure in space and time:

$$\pi(\boldsymbol{\delta}|\sigma_\delta^2) \propto \exp\left(-\frac{1}{2\sigma_\delta^2} \sum_{i=1}^{I} \sum_{t=1}^{T} \delta_{it}^2 \right).$$

- Type 2 interaction assumes the temporal trends differ between areas but do not have spatial structure:

$$\pi(\boldsymbol{\delta}|\sigma_\delta^2) \propto \exp\left(-\frac{1}{2\sigma_\delta^2} \sum_{i=1}^{I} \sum_{t=2}^{T} (\delta_{it} - \delta_{i,t-1})^2 \right).$$

- Type 3 interaction assumes the spatial patterns differ between time points but do not have temporal structure:

$$\pi(\boldsymbol{\delta}|\sigma_\delta^2) \propto \exp\left(-\frac{1}{2\sigma_\delta^2} \sum_{i\sim j} \sum_{t=1}^{T} (\delta_{it} - \delta_{jt})^2 \right).$$

- Type 4 interaction assumes the temporal trends differ between areas but are more likely to be similar in adjacent areas:

$$\pi(\boldsymbol{\delta}|\sigma_\delta^2) \propto \exp\left(-\frac{1}{2\sigma_\delta^2} \sum_{i\sim j} \sum_{t=2}^{T} (\delta_{it} - \delta_{i,t-1} - \delta_{j,t} + \delta_{j,t-1})^2 \right).$$

Such interaction models can also be represented in GMRF form, with the precision matrix being the Kronecker product of the corresponding spatial and temporal precision matrices. For example, the precision matrix for the Type 4 interaction term $\boldsymbol{\delta}$ can be specified as $\tau_\delta \boldsymbol{K}$, where $\boldsymbol{K}$ is the Kronecker product of (2.8) and (2.12). To make the interaction term $\boldsymbol{\delta}$ identifiable, we need to put $\sum_i \delta_{it} = 0$ and $\sum_t \delta_t = 0$ constraints in additional to the constraint $\sum_{it} \delta_{it} = 0$.

The GRMF framework provides a coherent representation for spatial, temporal and spatial-temporal modeling. The interaction model in Knorr-Held (2000) is our motivation for the development of the penalized spline model in Chapter 4, where we give further details of constructing spatial-temporal interactions using GMRFs.

2.3 Markov Chain Monte Carlo (MCMC) Sampling Schemes

To make Bayesian inference for the models we described earlier we describe the use of Markov Chain Monte Carlo (MCMC). The most essential yet difficult part of the MCMC for such models is to efficiently draw samples from the conditional distributions of the latent variables, when these latent variables are assigned GRMFs as priors. In particular, to gain improved convergence, block updating in which sets of parameters are jointly updated is required. In this section we describe different sampling schemes that can be employed in the MCMC method. In addition, we also introduce a modified version of block updating, where we take advantage of the fast computation that results from sampling from a sparse matrix. Rather than describe the procedure in a generic way, we take the spatial ICAR model as an example to make our discussion easier to follow.

The overall sampling scheme can be grouped into three classes: single-site updating for each latent variable and the hyperparameter separately, block updating all latent variables and the hyperparameter separately; and jointly updating all latent variables and the hyperparameter together. In Besag et al. (1991), the authors adopted a single site updating scheme. They updated each spatial random effect u_i given $\boldsymbol{u}_{-i}$ and then updated the hyperparameter τ_u. In Rue (2001), the vector of spatial random effect $\boldsymbol{u}$ was block updated and the hyperparameter τ_u was updated separately. He also described how to use a second-order Taylor approximation to generate proposals to increase acceptance rates. In

Knorr-Held and Rue (2002), the authors compared various block sampling scheme, with different block components and sizes. They concluded that separately updating the latent variables and the hyperparameters can provide misleading results, while joint updating the latent variables and the hyperparameters can significantly improve the mixing of the chain and provide far more reliable estimates.

2.3.1 Single-site Updating of the Latent Random Effect

In single-site updating scheme, we loop through latent random effects and update one latent variable at a time. To update the spatial random effect u_i in (2.3), the MCMC sampling proceeds by first generating a proposal u_i^* based on the current value u_i' from a proposal distribution q. We then calculate the acceptance/rejection ratio

$$r = \frac{\pi(u_i^*|rest)q(u_i'|u_i^*)}{\pi(u_i'|rest)q(u_i^*|u_i')}, \tag{2.15}$$

where $\pi(u_i|rest)$ is the full conditional distribution of variable u_i with $rest$ referring to all other variables that u_i depends upon. The proposal u_i^* is accepted with probability $\min(1, r)$.

A common approach for generating proposals is to use a random walk distribution

$$u_i^* \sim \mathrm{N}(u_i', c),$$

where c is a tuning parameter. This proposal distribution is symmetric, i.e. $q(u_i^*|u_i') = q(u_i'|u_i^*)$ and therefore these terms cancel when calculating the ratio r in (2.15). We may take a trial run to set the turning parameter c for desired acceptance rate, usually around $30\% - 40\%$ (Roberts et al., 1997).

We now discuss an alternative log-normal proposal for e^{u_i}. This approach is motivated by realizing that the usual conjugate prior for a Poisson random variable is Gamma(a, b), with mean a/b and variance a/b^2. In Model (2.3), the intrinsic ICAR prior is not conjugate for the Poisson model. However, we can find a conjugate gamma prior whose first and second moments match those of an ICAR prior. The full conditional distribution with the resulting gamma prior is then a conjugate gamma distribution from which samples can be easily drawn. This is the distribution we use as the proposal density when updating the

random effects. The approach of "matching moments" is a standard practice when using non-conjugate priors and we provide a detailed derivation for Model (2.3) here.

Let $\zeta_i = \exp(u_i)$, then suppose ζ_i has a log-normal distribution with mean $\mu_i = \exp(\bar{u}_i + 1/(2k_i\tau_u))$ and variance $\sigma_i^2 = \exp[2\bar{u}_i + 1/(k_i\tau_u)][\exp(1/(k_i\tau_u)) - 1]$, where $\bar{u}_i = \frac{1}{k_i}\sum_{j\sim i} u_j$ and k_i is the number of neighbors of area i. To match the mean and variance to those of a Gamma(a, b) distribution we set

$$\begin{aligned} \mu_i &= a/b, \\ \sigma_i^2 &= a/b^2 \end{aligned}$$

to yield $a = \mu_i^2/\sigma_i^2$ and $b = \mu_i/\sigma_i^2$. With $y_i|\mu_i \sim \text{Poisson}(E_i\mu_i)$, the posterior distribution of ζ_i is a gamma distribution with parameters $y_i + \mu_i^2/\sigma_i^2$ and $E_i + \mu_i/\sigma_i^2$. The logarithm of this gamma distribution is used as the proposal density in the Metropolis-Hastings algorithm when updating u_i in the single-site updating scheme.

Here the transformation from ζ_i to u_i needs to be taken into account when calculating the acceptance/rejection ratio r. Again let u_i' denote the current value of u_i, and with the proposed value $u_i^* = \log(\zeta_i^*)$ where ζ_i^* is drawn from Gamma$(y_i + \mu_i'^2/\sigma_i^2,\, E_i + \mu_i'/\sigma_i^2)$. The distribution of ζ_i^* is

$$q(\zeta_i^*|u_i') \propto {\zeta_i^*}^{(y_i+\mu_i'^2/\sigma_i^2)-1} e^{-(E_i+\mu_i'/\sigma_i^2)\zeta_i^*}.$$

Because $u_i^* = \log(\zeta_i^*)$ the Jacobian function is $d({\zeta^*}_i)/d(u_i^*) = e^{u_i^*}$. This gives

$$\begin{aligned} q(u_i^*|u_i') &\propto \exp[u_i^*(y_i + \mu_i'^2/\sigma_i^2 - 1) - (E_i + \mu_i'/\sigma_i^2)\exp(u_i^*)]|e^{u_i^*}| \\ &= \exp[u_i^*(y_i + \mu_i'^2/\sigma_i^2) - (E_i + \mu_i'/\sigma_i^2)\exp(u_i^*)]. \end{aligned}$$

The density $q(u_i'|u_i^*)$ can be obtained in a similar fashion. Notice that both u_i' and u_i^* depends on the average value of their neighbors and not themselves. Therefore, the normalizing constant here can be canceled in the proposal ratio because it doesn't depend on the current or proposed u_i.

A third method for forming a proposal is to approximate the full target distribution rather than just to match the first and second moments. The posterior conditional distribution of u_i is

$$\pi(u_i|\boldsymbol{u}_{-i}, v_i, \tau_u, \tau_v, y_i) \propto \exp\left(y_i u_i - E_i \exp(u_i + v_i) - \frac{m_i\tau_u}{2}(u_i - \bar{u}_i)^2\right).$$

The conditional distribution of u_i is not a GMRF because of the $\exp(u_i)$ term. However, we can approximate the distribution by a quadratic Taylor expansion around the current u_i':

$$\begin{aligned}
\exp(u_i^*) &\approx \exp(u_i')\Big(1 + (u_i^* - u_i') + \frac{1}{2}(u_i^* - u_i')^2\Big) \\
&= constant + \exp(u_i')(1 - u_i')u_i^* + \frac{1}{2}\exp(u_i')u_i^{*2} \\
&= A_i + B_i u_i^* + \frac{1}{2}C_i u_i^{*2}, \\
\text{where} \quad A_i &= constant, \\
B_i &= \exp(u_i')(1 - u_i'), \\
C_i &= \exp(u_i')
\end{aligned}$$

The resulting conditional distribution $\pi(u_i|rest)$ is a GMRF and can be used as the proposal distribution in a Metropolis-Hastings step. The proposal density $q(u_i^*|u_i')$ based on the current value u_i' is now

$$q(u_i^*|u_i') \propto \exp\left(y_i u_i' - E_i \exp(v_i) B_i u_i^* - E_i \exp(v_i)\frac{1}{2}C_i u_i^{*2} - \frac{\tau_u}{2}\sum_{i\sim j}(u_i' - u_j')^2\right).$$

Note that the normalizing constant for $q(u_i^*|u_i')$ depends on u_i' and therefore does not cancel.

It is obvious that single-site updating is very time consuming, especially when we have hundreds or even thousands of latent variables. In addition, the high dependence in the latent variables often results in very slow mixing. A natural way to reduce the computation time is to update all of the latent variables at once in a block.

2.3.2 Block Updating of the Latent Random Effect

The Taylor expansion approximation method in the single-site updating can be easily extended to the multivariate case. Let $\boldsymbol{u}^*$ and $\boldsymbol{u}'$ indicate the proposed and current value of the latent variable vector

$$\begin{aligned}
q(\boldsymbol{u}^*|\boldsymbol{u}') &\propto \exp\Big(\sum_i y_i u_i' - \sum_i E_i \exp(v_i) B_i u_i^* - \sum_i E_i \exp(v_i)\frac{1}{2}C_i u_i^{*2} - \frac{\tau_u}{2}\sum_{i\sim j}(u_i' - u_j')^2\Big) \\
&= \exp\Big(-\frac{\boldsymbol{u}'^T(\boldsymbol{Q} + \boldsymbol{K})\boldsymbol{u}'}{2} + \boldsymbol{b}^T\boldsymbol{u}'\Big),
\end{aligned}$$

where $\boldsymbol{K} = diag\big(E_i \exp(v_i) C_i\big)$, $\boldsymbol{Q}$ is the precision matrix for the GMRF of $\boldsymbol{u}$ in (2.7), and $\boldsymbol{b}$ is the vector with entry $b_i = (y_i - E_i \exp(v_i) B_i), i = 1, \ldots, n$. The acceptance/rejection ratio in the MCMC updating is now, in vector form,

$$\frac{\pi(\boldsymbol{u^*}|rest) q(\boldsymbol{u'}|\boldsymbol{u^*})}{\pi(\boldsymbol{u'}|rest) q(\boldsymbol{u^*}|\boldsymbol{u'})}.$$

Similar to the single-site updating case, the normalizing constant of the proposal density does not cancel when the ratio is calculated.

2.3.3 A Modified Block Updating

The modified block updating scheme we develop takes advantage of the ease of sampling from an IGMRF. Sampling from an IGMRF with rank-deficient n by n precision matrix $\boldsymbol{Q}$ involves the following steps (Rue and Held, 2005):

1. Compute the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of matrix $\boldsymbol{Q}$ (ordered from the smallest to the largest), and the corresponding eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$. Suppose $\boldsymbol{Q}$ has rank-deficiency k, then $\lambda_1, \ldots, \lambda_k$ would be 0.

2. For $j = k + 1$ to n, generate $y_j \sim \mathrm{N}(0, 1/\lambda_j)$.

3. Let $\boldsymbol{x} = y_{k+1}\boldsymbol{e}_{k+1} + y_{k+2}\boldsymbol{e}_{k+2} + \ldots + y_n \boldsymbol{e}_n$, then $\boldsymbol{x}$ is the desired samples from the IGMRF with structure matrix $\boldsymbol{Q}$.

When drawing proposed values of a set of spatially-varying random effect $\boldsymbol{u}^*$ based on the current values $\boldsymbol{u'}$, we use the following proposal distribution:

$$\boldsymbol{u}^* \sim \mathrm{N}(\boldsymbol{u'}, c\boldsymbol{K}),$$

where c is a turning parameter and $\boldsymbol{K}$ is the structure matrix for the neighborhood relationship. This proposal distribution is clearly an IGMRF because the matrix $\boldsymbol{K}$ is rank-deficient. Now $c\boldsymbol{K}$ is like the matrix $\boldsymbol{Q}$ in the algorithm above, and we can generate $\boldsymbol{u}^*$ from $\mathrm{N}(\boldsymbol{0}, c\boldsymbol{K})$ and then simply add the current value $\boldsymbol{u'}$ to the generate value. This modified block updating scheme is very fast and efficient, matrix $\boldsymbol{K}$ is pre-defined based on the neighborhood

structure and we only need to compute the eigenvalues and eigenvectors of the matrix $\boldsymbol{K}$ once before the MCMC iterations. Then for each update during the MCMC iterations, we only need the last two steps of the algorithm of sampling from an IGMRF given in Section 2.3.2.

2.3.4 Joint Updating of the Latent Random Effect

In joint updating the latent variable and its precision parameter, we first generate a proposal τ_u^* and then generate $\boldsymbol{u}^*$ based on the proposed τ_u^*. The joint proposal $(\tau_u^*, \boldsymbol{u}^*)$ is accepted or rejected jointly with probability

$$\min\left\{1, \frac{\pi(\tau_u^*, \boldsymbol{u}^*|\boldsymbol{y})}{\pi(\tau_u, \boldsymbol{u'}|\boldsymbol{y})} \frac{q(\tau_u, \boldsymbol{u'}|\tau_u^*, \boldsymbol{u}^*, \boldsymbol{y})}{q(\tau_u^*, \boldsymbol{u}^*|\tau_u, \boldsymbol{u'}, \boldsymbol{y})}\right\},$$

where

$$\frac{q(\tau_u, \boldsymbol{u'}|\tau_u^*, \boldsymbol{u}^*, \boldsymbol{y})}{q(\tau_u^*, \boldsymbol{u}^*|\tau_u, \boldsymbol{u'}, \boldsymbol{y})} = \frac{q(\tau_u|\boldsymbol{u'}, \tau_u^*, \boldsymbol{y})}{q(\tau_u^*|\boldsymbol{u'}, \tau_u, \boldsymbol{y})} \frac{\pi(\boldsymbol{u'}|\boldsymbol{u}^*, \tau_u^*, \boldsymbol{y})}{\pi(\boldsymbol{u}^*|\boldsymbol{u'}, \tau_u, \boldsymbol{y})}$$

If the proposal distribution for the precision parameter τ_u is not symmetric, we will need to explicitly calculate each term in the acceptance/rejection ratio. The joint updating approach provides better mixing of the parameters in general and prevents the MCMC chain getting trapped in long tails of the posterior distributions (Knorr-Held and Rue, 2002).

2.4 Integrated Nested Laplace Approximation (INLA)

The Markov Chain Monte Carlo (MCMC) method for obtaining the posterior distributions and making inference about the parameter of interest is very computationaly expensive, especially for complicated models such as the spatial-temporal models we are interested in. In addition, convergence assessment of MCMC can be difficult to judge. Rue and Martino (2009) introduced the integrated nested Laplace approximations (INLA) as an alternative for making Bayesian inference in hierarchial models. INLA cleverly combines the Laplace approximation and numerical integration strategies, and provides accurate inference about posterior distributions with considerably less computation time. Details about INLA can be found in Rue and Martino (2009) but we give a short review for a simple case. Let $\boldsymbol{y}$ denote a vector of observations that belongs to an exponential family, with likelihood $p(\boldsymbol{y}|\boldsymbol{x})$. The

latent variable $\boldsymbol{x}$ has a zero mean Gaussian distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the unknown hyperparameter with density $\pi(\boldsymbol{\theta})$. The posterior distribution is

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta})$$

and we are interested in the marginal distributions $p(x_i|\boldsymbol{y})$ and $p(\theta_i|\boldsymbol{y})$. These marginal distributions can be written as

$$p(\theta|\boldsymbol{y}) \propto \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \tag{2.16}$$

$$p(x_i|\boldsymbol{y}) = \int p(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \times p(\boldsymbol{\theta}|\boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta} \tag{2.17}$$

The above equations can be evaluated via approximation

$$p(\theta|\boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{x^*(\boldsymbol{\theta})} \tag{2.18}$$

$$\begin{aligned} \tilde{p}(x_i|\boldsymbol{y}) &= \int \tilde{p}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \times \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta} \\ &\approx \sum_k p(x_i|\theta_k, \boldsymbol{y}) \times p(\theta_k|\boldsymbol{y}) \times \Delta_k \end{aligned} \tag{2.19}$$

where $x^*(\boldsymbol{\theta})$ is the mode of the full conditional distribution for $\boldsymbol{x}$ for a given $\boldsymbol{\theta}$. The full conditional distribution $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ can be approximated using Taylor series expansion including a Laplace approximation. The integration over $\boldsymbol{\theta}$ is performed by evaluating a set of points over a grid created with weights Δ_k. The INLA approach gives the approximated marginal distributions that can be used to obtain point estimates as well as credible intervals. The assessment of the accuracy of the approximation can be difficult. In many applications, the INLA approach has been shown to provide an accurate approximation to the MCMC method. However, in some applications especially those involving binary data, the INLA approximation can be quite off (Fong et al., 2010).

2.5 Comparison of MCMC and INLA with London Health Authority Data

In this section, we take the London Health Authority (LHA) data that can be found in the *WinBUGS* set of examples (Spiegelhalter et al., 1998) to compare MCMC and INLA. *WinBUGS* is a free statistical software program for Bayesian analysis using MCMC. Besides *WinBUGS*, there are a variety of other statistical software available for MCMC, for example,

BayesX from Fahrmeir et al. (2004) and *JAGS* from Plummer (2009). For the MCMC method, we run the data set using *WinBUGS*, *BayesX*, and using *R* code we developed with our modified block updating scheme (see Section 2.3.3). For the INLA method, we run the data set using the *R* package *INLA* (Rue and Martino, 2009) and compare the results with MCMC.

The LHA data are simulated observed and expected counts of lung cancer incidence in males aged 65 and over living in 44 wards in the Health Authority region of London. A ward level socio-economic deprivation index is also available as part of the data. We choose the LHA data as the example for its moderate number of geographic regions, and its availability as a ready-to-run example in the *WinBUGS* Manual.

The model we use to fit the LHA data is as follows:

$$\begin{aligned} Y_i|\mu_i &\sim \text{Poisson}(E_i\mu_i) \\ \log(\mu_i) &= \alpha + \beta x_i + u_i + v_i, \end{aligned}$$

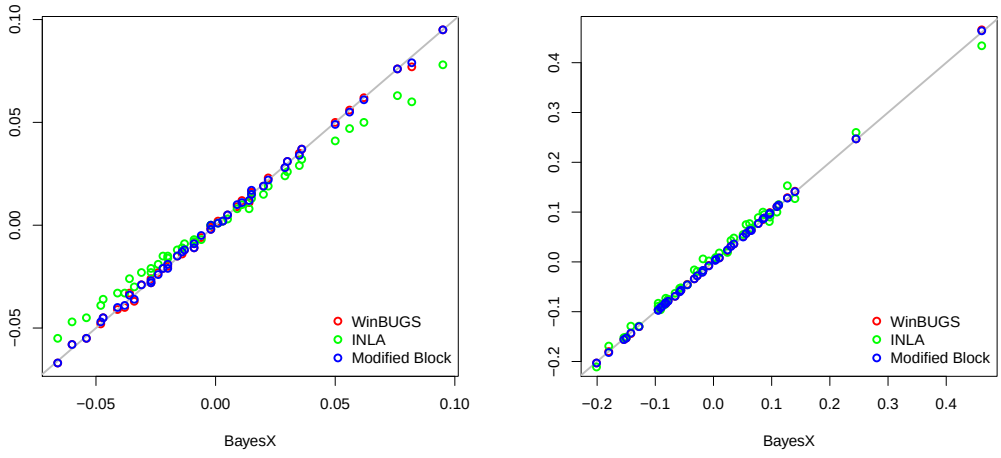
where Y_i and E_i are observed and expected counts of lung cancer incidence. The covariate x_i is the area-level socio-economic deprivation index. We include spatially structured ICAR random effect u_i and non-spatial random effect v_i as described in (2.3). We use Gamma$(1, 0.026)$ priors for the precision parameter τ_u and τ_v based on the prior suggestion in Fong et al. (2010). We then fit the model in both *WinBUGS* and *BayesX* with $550,000$ iterations being carried out. The first $50,000$ iterations are discarded as the burn-in period and inference is made using the remaining $500,000$ MCMC samples. Visual examination of the trace plots show good mixing and convergence of the MCMC chains. For the modified block updating scheme we develop, we run the model under this modified sampling scheme for $550,000$ iterations in an *R* implementation with the first $50,000$ discarded as the burn-in.

The comparison of the parameter estimation using different estimation techniques is summarized in Table 2.1. We use the standard deviation σ in the inference, which is $\tau^{-1/2}$. The parameter estimation is very similar across the different methods, which is reassuring. In Figure 2.1 we present the estimated spatial and non-spatial random effects using different estimation techniques. For the spatial random effect u_i, the results using *WinBUGS*, *INLA* and our own *R* code are almost identical, while INLA is slightly different. However, notice

the estimated spatial random effect is on a very small scale, and the difference between INLA and MCMC is very small in an absolute sense. For the non-spatial random effecst v_i, the estimation difference between INLA and MCMC is negligible. In terms of the computation time, all MCMC methods using *WinBUGS*, *BayesX* or our modified block updating scheme in *R* take about 2 hours while INLA takes only about 3 seconds.

Table 2.1: Comparison of the parameter estimation with different estimation techniques using LHA data.

node	WinBUGS	BayesX	Modified Block	INLA
α	-0.219	-0.219	-0.218	-0.224
β	0.045	0.045	0.045	0.046
σ_u	0.193	0.194	0.193	0.193
σ_v	0.231	0.229	0.231	0.228



(a) Spatial Random Effect (b) Non-Spatial Random Effect

Figure 2.1: Comparison of the estimated random effect using different estimation techniques.

2.6 Conclusion

In this chapter, we have briefly reviewed the GMRF models and made the connection to some of the most popular spatial and temporal models. We then focused on the computational aspect of these models and gave descriptions of different MCMC sampling schemes for Bayesian analysis. Using MCMC in spatial-temporal models is often extremely time consuming, partially due to the high correlation of the latent variables in space and in time. For complicated spatial-temporal models, the computation may be the main obstacle for real applications, especially the ones that require real-time results.

An alternative approach that has gained great popularity is INLA. With an analysis of the LHA data, we have compared MCMC estimation with the alternative INLA method. The analysis shows that INLA can provide good approximation to MCMC with far less computation time, and suggests INLA is a useful tool for spatial-temporal models. However, caution needs to be taken with INLA, as it may not always provide an accurate approximation to the MCMC method, see the Salamander example in Fong et al. (2010).

Chapter 3

THE USE OF SAMPLING WEIGHTS IN BAYESIAN HIERARCHICAL MODELS FOR SMALL AREA ESTIMATION

3.1 Introduction

Small area estimation is a common enterprise in many disciplines, particularly in the social sciences and public health. Often such estimation is based on data arising from complex surveys, for which sampling weights are calculated to account for the disproportionate nature of the sample, by comparison with the target population of interest. Two common forms of bias that are controlled for by weights are non-response and non-coverage bias. In addition to bias control, a second important consideration in small area estimation is variance reduction. As the target areas of interest decrease in size, the uncertainty in estimation is increased, and so the search for smoothing techniques has been popular. From a statistical perspective there are a number of important issues that need consideration when faced with data from a survey. First, one needs to choose whether to use a design- or model-based approach to inference. Unfortunately, though a model-based approach is appealing, the practical difficulties of modeling complex survey design are currently problematic, see Gelman (2007) and the ensuing discussion. Second, one needs to decide on whether the target of inference is a characteristic of the population from which sampling has been carried out (for example an area total), or a characteristic of the superpopulation from which the population was hypothetically sampled (Graubard and Korn, 2002).

Many hierarchical approaches are available for reducing the mean squared error in estimation of small area proportions or counts, and here we provide a brief flavor of developments. Hierarchical models for small area estimation can be traced back to the Fay-Herriot model (Fay and Herriot, 1979), in which an adaptation of the James-Stein estimator was applied to sample estimates of income for small places (in their case, populations less than 1,000). Datta and Ghosh (1991) described a hierarchical Bayes model in which predictive

distributions were derived for the unobserved non-sampled responses, in a normal linear model. Hierarchical Bayes models developed for binary survey data include Nandram and Sedransk (1993), in which Bayesian predictive inference was carried out for a finite population proportion from a two-stage cluster sample. A comprehensive treatment of the Bayesian predictive approach to binary survey was provided by Stroud (1994), and included simple random, stratified, cluster, and two-stage sampling, as well as two-stage sampling within strata. Ghosh et al. (1998) provide a general approach for small-area estimation based on hierarchical Bayes generalized linear models, with extension to a spatial correlation random effect structure. Farrell (2000) describes a logistic hierarchical model with computation via Markov chain Monte Carlo; the model compares favorably with an empirical Bayes technique (Farrell et al., 1997), though design bias is present in both procedures. Malec et al. (1997) describe a hierarchical Bayes model for binary survey data. They examined the use of sampling weights as a covariate in the model and did not find any improvement for their example of county-level data from the National Health Interview Survey. Review articles on small area estimation include Rao (1999) and Pfeffermann (2002), with Rao (2003) providing a comprehensive review of design-based, empirical Bayes, and hierarchical Bayesian methods. In this dissertation we add to this literature by providing a simple hierarchical approach that is appropriate for spatial data with associated sampling weights.

The outline of this chapter is as follows: in Section 3.2 we describe the notation and the conventional methods used for design-based small area analysis. In the next section we present our approach to include the sampling weights in Bayesian hierarchial models. we conduct a series of simulation studies to demonstrate the performance of our approach, and gives the results in Section 3.4. In Section 3.5, we revisit the motivating example of the Washington 2006 BRFSS data and present the results using our proposed method. The conclusion and future work is presented in Section 3.6.

3.2 Notation and the Conventional Methods of Analysis

3.2.1 Notation

Let i index area, j the group classification by which post-stratification is carried out (which is age and gender in our motivating example) and k the individual, $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, N_{ij}$, so that N_{ij} represents the population size in area i and group j. In this section, to simplify notation, we assume there are no strata beyond those used for post-stratification. Let Y_{ijk} denote the binary variable indicating whether the k-th individual in group j from area i has the outcome of interest ($Y_{ijk} = 1$) or not ($Y_{ijk} = 0$). Common small area characteristics of interest are the true proportions, $P_i = \frac{\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} Y_{ijk}}{\sum_{j=1}^{J} N_{ij}}$, or true counts, $T_i = \sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} Y_{ijk}$, in area i, $i = 1, \ldots, I$. In the context of a complex survey, let S_{ijk} denote the binary variable indicating whether the k-th individual in group j from area i is sampled ($S_{ijk} = 1$) or not ($S_{ijk} = 0$). Given N_{ij} we may calculate selection probabilities as $\pi_{ij} = \frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}} S_{ijk}$. In addition, let R_{ijk} be the binary variable indicating whether the sampled individual responds to the survey ($R_{ijk} = 1$) or not ($R_{ijk} = 0$). We let $m_{ij} = \sum_{k=1}^{N_{ij}} R_{ijk}$ denote the sample size of group j from area i. For brevity, we use $m_i = \sum_{j=1}^{J} m_{ij}$ to denote the total sample size from area i, and $N_j = \sum_{i=1}^{I} N_{ij}$ to denote the total population size of group j over the study area. To reflect the sampling design, weights w_{ijk} are attached to each respondent's outcome. For example, (1.1), provides the form of the weights in the BRFSS example. In general, the weights will reflect both the selection probability and post-stratification.

3.2.2 Conventional Methods

The most commonly used direct unbiased estimator of the area proportion in complex surveys is the (post-stratified) Horvitz–Thompson (Horvitz and Thompson, 1952):

$$\hat{p}_i = \frac{\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk} w_{ijk} y_{ijk}}{\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk} w_{ijk}}, \qquad i = 1, \ldots, I, \tag{3.1}$$

where y_{ijk} is the observed response and w_{ijk} is the sampling weight assigned to the k-th person in area i and group j. A common strategy is to calculate weights as the product

of the reciprocal of the sampling probability for selection (the design weight) and the post-stratification weights:

$$w_{ijk} = \pi_{ijk}^{-1} \times \frac{N_j}{\hat{N}_j}, \tag{3.2}$$

for $k = 1, \ldots, N_{ij}$ where $\hat{N}_j = \sum_{i=1}^{I} \sum_{k=1}^{N_{ij}} R_{ijk} \pi_{ijk}^{-1}$, so that $\sum_{i=1}^{I} \sum_{k=1}^{N_{ij}} w_{ijk} = N_j$, the known group totals in the population. Hence, the design weights adjust for the systematic sampling scheme used, while the post-stratification weights attempt to adjust for non-response, by rescaling each group j so that the estimated population total matches the known population total. The computations and the motivation are the same as direct standardization of proportions and rates in epidemiology, although direct standardization is typically used to reduce bias from confounding rather than from non-response. Non-response bias will be removed to the extent that it is predicted by group membership; post-stratification has no impact on differential non-response within groups.

The variance of $\hat{p}_i$ can be expressed as

$$\hat{\text{var}}(\hat{p}_i) = \frac{1}{m_i} \left(\frac{N_i - m_i}{N_i} \right) \hat{\sigma}_i^2, \qquad i = 1, \ldots, I, \tag{3.3}$$

where $\hat{\sigma}_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{J} \sum_{k=1}^{N_{ij}} R_{ijk} (y_{ijk} - \hat{p}_i)^2$ is the empirical variance of y_{ijk}, $j = 1, ..., J; k = 1, ..., N_{ij}$.

The estimator $\hat{p}_i$ defined in (3.1) is unbiased in the absence of non-response, and approximately unbiased when post-stratification is used to correct non-response, but it is very imprecise when the sample size is small. For example, for rare events and a small sample size, the empirical variance $\hat{\sigma}_i^2$ can be 0, which results in a zero variance of the estimated proportion given in (3.3).

In this chapter we show that the bias correction provided by sampling weights and non-response weights can be combined with the reduction in variance provided by Bayesian hierarchical models, to achieve more accurate estimation (in a mean squared error sense) than either technique alone.

3.2.3 Inference

For concreteness we focus on predicting counts. To obtain point and interval estimates of the total count we have

$$\hat{Y}_i = \mathrm{E}(Y_i) = y_i + \hat{p}_i \times (N_i - m_i), \qquad i = 1, \ldots, I. \tag{3.4}$$

with variance estimate

$$\hat{\mathrm{var}}(\hat{Y}_i) = \hat{\mathrm{var}}(\hat{p}_i) \times (N_i - m_i)^2, \qquad i = 1, \ldots, I. \tag{3.5}$$

In most large-scale surveys the sample size is very small compared to the population size, and the overlap between sample and population is often ignored. In survey terminology, the data are analyzed as if they were sampled with replacement. This simplifies the point estimate of the population count to

$$\hat{Y}_i = \mathrm{E}(Y_i) = \hat{p}_i \times N_i, \qquad i = 1, \ldots, I. \tag{3.6}$$

with variance estimate

$$\hat{\mathrm{var}}(\hat{Y}_i) = \hat{\mathrm{var}}(\hat{p}_i) \times N_i^2, \qquad i = 1, \ldots, I. \tag{3.7}$$

This point estimate (3.5) is valid even with large sampling fractions when $\hat{p}_i$ is the Horvitz–Thompson estimator (3.1), because this estimator is based on data from area i only. It is not valid with large sampling fractions for the Bayesian estimators that we describe in the next section, which borrow strength from other regions.

3.3 Sample Weighted Bayesian Hierarchical Models

3.3.1 A Definition of Effective Sample Size

As described in the previous section, the weights (3.2) will correct the mean of a population total for sampling bias, and for non-response to the extent that this is explained by the post-strata. Bayesian modelling requires more than bias correction; we need a likelihood that approximates the distribution of the data/weighted probabilities. Following Korn and Graubard (1998), we model the weighted probability estimates as binomial proportions, with

an "effective sample size" chosen to match the binomial variance to the sampling variance of the estimates. Using the effective sample size rather than the actual sample size allows for the varying information per observation under complex sampling. The precision of an estimate from a complex sample can be higher than for a simple random sample, because of the better use of population data, via stratification and post-stratification. However, the precision can also be lower, either because of correlation within clusters (which reduces information), or because the design was optimized for estimating a specific quantity which is not well correlated with the quantity of interest. The ratio of the effective sample to the actual sample size is the reciprocal of Kish's "design effect" (Kish, 1995), a standard summary of the efficiency of a sampling design.

To approximate the sampling distribution in the estimator (3.1) we express the sampling variances in terms of the effective sample sizes for simple random samples. In a simple random sample, the estimated variance would be $\hat{p}_i(1-\hat{p}_i)/m_i$. For our approach define

$$\hat{p}_{\cdot j} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{N_{ij}} R_{ijk}w_{ijk}y_{ijk}}{\sum_{i=1}^{I}\sum_{k=1}^{N_{ij}} R_{ijk}w_{ijk}},$$

and $e_{ijk} = y_{ijk} - \hat{p}_{\cdot j}$. The estimated variance for the post-stratified mean is

$$\hat{\text{var}}\,(\hat{p}_i) = \frac{N_i - m_i}{N_i}\frac{1}{m_i(m_i-1)}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk}e_{ijk}^2.$$

The "effective sample size" m_i^* is then obtained by solving

$$\frac{\hat{p}_i(1-\hat{p}_i)}{m_i^*} \quad = \quad \frac{N_i - m_i}{N_i}\frac{1}{m_i(m_i-1)}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk}e_{ijk}^2$$

to give

$$m_i^* = \hat{p}_i(1-\hat{p}_i)/\left(\frac{N_i - m_i}{N_i}\frac{1}{m_i(m_i-1)}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk}e_{ijk}^2\right). \tag{3.8}$$

We define the effective number of cases y_i^* as the product of the effective sample size m_i^* and the estimated proportion $\hat{p}_i$:

$$y_i^* = m_i^* \times \hat{p}_i. \tag{3.9}$$

For small sample sizes, the design-weighted estimator $\hat{p}_i$ in (3.1) can be zero or unity; the estimated variance of the weighted estimator in (3.8) is then zero. To avoid this problem we first use an empirical Bayes model based on a Beta-Binomial model (described in greater detail later in this section) to estimate the proportion, $\widetilde{p}_j$, using data from all areas where $\sum_j \sum_k Y_{ijk}$ is available to estimate the group proportion with the outcome of interest. We then define the weighted estimator of proportions for these area i as:

$$\hat{p}_i = \frac{\sum_{j=1}^{J} m_{ij}\widetilde{p}_j}{m_i}.$$

The weighted estimator defined in this way is moved slightly away from 0 or 1. The calculation of the effective sample size for these areas remains as in (3.8).

3.3.2 Multiple Strata

In this section we extend the model to allow for $h = 1, \ldots, H$ strata. We begin with the simplest situation in which groups, j, and areas, i, are nested within strata, $i = 1, \ldots, I_h$. For example, if h indexes large sampling areas, i indexes smaller areas nested within h, and j age-gender post-stratified groups also within h. In this case we define $e_{hijk} = y_{hijk} - \hat{p}_{h\cdot j}$ with

$$\hat{p}_{h\cdot j} = \frac{\sum_{i=1}^{I_h} \sum_{k=1}^{N_{hij}} R_{hijk} w_{hijk} y_{hijk}}{\sum_{i=1}^{I_h} \sum_{k=1}^{N_{hij}} R_{hijk} w_{hijk}},$$

where R_{hijk} and w_{hijk} are the response indicators and sampling weights. In this nested case $\hat{p}_i = \hat{p}_{hi}$, $N_i = N_{hi}$ and $m_i = m_{hi}$. The variance is then

$$\hat{\text{var}}\left(\hat{p}_{hi}\right) = \frac{N_{hi} - m_{hi}}{N_{hi}} \frac{1}{m_{hi}(m_{hi} - 1)} \sum_{j=1}^{J} \sum_{k=1}^{N_{hij}} R_{hijk} e_{hijk}^2. \tag{3.10}$$

The effective sample size, $m_{hi}^{\star}$, is defined exactly as in the non-stratified case by setting $\hat{\text{var}}\left(\hat{p}_{hi}\right) = \hat{p}_{hi}(1 - \hat{p}_{hi})/m_{hi}^{\star}$ and solving for $m_{hi}^{\star}$. The National Health And Nutrition Examination Surveys (NHANES) are one example of a large survey where post-stratification groups are nested in strata (Mohadjer et al., 1996).

In the non-nested case, the centering by stratum mean and group mean must be done separately. We define $d_{hijk} = y_{hijk} - \hat{p}_{\cdot\cdot j}$ and $e_{hijk} = d_{hijk} - \hat{d}_h$ where

$$\hat{p}_{\cdot\cdot j} = \frac{\sum_h \sum_i \sum_k R_{hijk} w_{hijk} y_{hijk}}{\sum_h \sum_i \sum_k R_{hijk} w_{hijk}}, \tag{3.11}$$

and

$$\hat{d}_h = \frac{\sum_i \sum_j \sum_k R_{hijk} w_{hijk} d_{hijk}}{\sum_i \sum_j \sum_k R_{hijk} w_{hijk}} \tag{3.12}$$

and the variance is

$$\hat{\text{var}}\,(\hat{p}_i) = \frac{N_i - m_i}{N_i} \frac{1}{m_i(m_i - 1)} \sum_h \sum_j \sum_k R_{hijk} e^2_{hijk}. \tag{3.13}$$

where the sums are taken over all (h, j, k) combinations that exist in the population. In (3.11) and (3.12) the summations are over all combinations of indices that occur in the population.

3.3.3 Hierarchical Models

We employ Bayesian hierarchical models that involve three stages for inference. At the first stage, we approximate the sampling distribution using the design-based variance for the survey-weighted estimator, as defined in Section 3.3.1. At the second stage, we model between-area variation using random effect models. Finally, the unknown hyperparameters in the second stage are assigned proper hyperprior distributions at Stage 3.

In the first stage, the data distribution is assumed to be:

$$y_i^* | p_i \sim \text{Binomial}(m_i^*, p_i), \qquad i = 1, \ldots, I,$$

where y_i^* and m_i^* are as defined (3.9) and (3.8), respectively. By construction, the sampling distribution of the commonly used estimator, y_i^*/m_i^*, is unbiased for the population prevalence p_i (under the same conditions as the estimator (3.1) is unbiased, as detailed in Section 3.2) and the reciprocal of the Fisher information is equal to the design-based variance estimate, giving an appropriate indication of precision. As a binomial distribution, it also respects the [0, 1] bounds on p_i. As this data distribution is a good approximation to the mean, variance, and range of the actual data distribution, it should give a reasonable approximation to the likelihood for Bayesian inference. At the second stage, we describe three possible random effect models to account for between-area variation.

Model 1: *Independent beta random effects model, with empirical Bayes estimation*
The small area proportions p_i follow a beta distribution:

$$p_i|a,b \sim_{iid} \text{Beta}(a,b), \qquad i = 1,\ldots,I. \tag{3.14}$$

The marginal distribution is

$$\begin{aligned} \Pr(y_i^*|a,b) &= \int \Pr(y_i^*|p_i) f(p_i|a,b)\ dp \\ &= \binom{m_i^*}{y_i^*} \frac{\Gamma(a+b)\Gamma(a+y_i^*)\Gamma(b+m_i-y_i^*)}{\Gamma(a)\Gamma(b)\Gamma(a+b+m_i^*)}. \end{aligned} \tag{3.15}$$

Under a full Bayesian approach a prior is placed on a, b, and these parameters are subsequently integrated over. Unfortunately this integration is not analytically tractable and so, as a simple empirical Bayes alternative a and b can be estimated from the marginal distribution of the data by maximum likelihood to give estimates $\hat{a}, \hat{b}$. Inference for p_i can then be made based on the posterior mean:

$$\hat{p}_i = v_i \times \frac{y_i^*}{m_i^*} + (1 - v_i) \times \left(\frac{\hat{a}}{\hat{a}+\hat{b}}\right) \qquad \text{where } v_i = \frac{m_i^*}{\hat{a}+\hat{b}+m_i^*}. \tag{3.16}$$

The estimated p_i is a familiar form, the weighted combination of the maximum likelihood estimator y_i^*/m_i^* and the mean of the Beta prior distribution $\hat{a}/(\hat{a}+\hat{b})$. Hence an outlying estimate based on a small (effective) sample size will be shrunk towards the posterior mean. This shrinkage of random effect estimators inherently induces bias in estimation but reduces the estimated variance.

Model 2: *Independent normal random effects model, full Bayes estimation*
A commonly generalized linear mixed model is

$$\begin{aligned} \text{logit}(p_i) &= \alpha + v_i \\ v_i|\sigma_v^2 &\sim_{iid} \text{N}(0,\sigma_v^2), \qquad i = 1,\ldots,I, \end{aligned} \tag{3.17}$$

where α is the overall effect and the random effects v_i capture the unexplained log odds ratio of the prevalence in the residuals in area i. When area-level covariates are available, the model can be extended to:

$$\text{logit}(p_i) = \alpha + \boldsymbol{x}_i{}^T\boldsymbol{\beta} + v_i,$$

where $\boldsymbol{x}_i$ is a vector of length p associated with area i and $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects.

Model 3: *Intrinsic conditional autoregressive (ICAR) model, full Bayes estimation*

In general, we might expect areal units that are close to each other tend to share more similarities then units that are far away and we would like to exploit this information in order to provide more reliable estimates in each area. Here we adopt the spatial model introduced by Besag et al. (1991) described in Section 2.2 that supplements the independent normal random effects with spatial terms:

$$\begin{aligned} \text{logit}(p_i) &= \alpha + \boldsymbol{x}_i{}^T\boldsymbol{\beta} + u_i + v_i \\ u_i|u_j, j \neq i &\sim \text{N}\left(\frac{1}{k_i}\sum_{j\sim i} u_j, \frac{\sigma_u^2}{k_i}\right) \\ v_i &\sim_{iid} \text{N}(0, \sigma_v^2), \qquad i = 1, \ldots, I. \end{aligned} \tag{3.18}$$

The parameter σ_u^2 is a *conditional* variance and it determines the contribution of spatial variation. The variance parameters σ_v^2 and σ_u^2 are on different scales, and therefore cannot be directly compared. However, the amount of variation that can be explained by the spatial component can be estimated as the empirical variance, $\text{var}(u_i)$. Again, $\boldsymbol{x}_i$ is the available area-level covariates.

In this model the nature of the spatial dependency is defined by the neighborhood structure. For example, a common approach defines areas i and j to be neighbors if they share a common boundary. Other neighborhood schemes are possible, for example, Cressie and Chan (1989) define the neighborhood structure as a function of the distance between centroids. For the independent normal and spatial models we require priors for β_0 and the random effects variances. A normal hyperprior is typically assumed for the former, and gamma distributions for the latter.

A related model has been suggested by Raghunathan et al. (2007), in the context of combining data from multiple sources. In the context of estimating a proportion, they assume the model

$$\sin^{-1}\sqrt{\hat{p}_i}|p_i \sim_{ind} \text{N}\left(\sin^{-1}\sqrt{p_i}, \frac{1}{4\widetilde{m}_i}\right).$$

The arcsine-squared root transformation stabilizes the variance but may be deficient for areas with small sample sizes. In addition, this model does not constraint the proportion

to be between 0 and 1.

For inference, formulas (3.4)–(3.7) are all still relevant, but with y_i replaced by y_i^* and m_i replaced by m_i^*, and $\hat{p}_i$ a suitable location estimate such as the posterior median. In the simulations and the BRFSS example we used (3.6) as the point estimate, since the survey sample sizes in each zip code were small compared to the populations. So far as the variance is concerned we can use the the posterior, $\text{var}(p_i|y)$. A more precise summary is the complete posterior distribution of Y_i, but this is more computationally complex to obtain.

3.3.4 Implementation

The usual implementation of Bayesian hierarchical models is via Markov chain Monte Carlo (MCMC) described in Chapter 2. However, as mentioned in Chapter 2, the large computational burden can impede the application of Bayesian hierarchical models in practice. INLA is particularly useful for simulation studies, as we demonstrate in the next section. In our simulations the gain in speed is substantial, for example, the analysis of the BRFSS diabetes data in Washington State takes about 5 seconds using INLA, and hours using MCMC.

3.4 Simulation Study

3.4.1 Simulation Scenarios

We examine five sampling scenarios with different response probabilities. For simplicity we consider the single stratum case. In order to investigate the bias introduced by non-response, we let q_{ijk} denote the response probability associated with the k-th individual in group j and area i. Therefore, given that the k-th individual is sampled, the response indicator R_{ijk} follows a Bernoulli distribution, $R_{ijk}|S_{ijk} = 1, q_{ijk} \sim \text{Bernoulli}(q_{ijk})$. We consider five data generating scenarios:

Scenario 1:

In scenario 1 we assume there is no non-response in the survey. In other words, $q_{ij} = 1$ for all sampled individuals. This is the ideal situation. The prevalence of diabetes we use

across six gender-age groups are: Female, 18–44, 0.017; Female, 45–74, 0.15; Female, 75+, 0.17; Male, 18–44, 0.014; Male, 45–74, 0.16; Male, 75+, 0.19. These values are chosen based on the National Surveillance Data from the CDC website:

`http://www.cdc.gov/diabetes/statistics/prev/national/menuage.htm`

Scenario 2:

In scenario 2 we consider a more practical sampling situation. We assume that not every sampled individual will respond to the survey and the response rate is different for each group j. However, the response rate in each group is the same for each area. The response rates are: Female, 18–44, 0.55; Female, 45–74, 0.65; Female, 75+, 0.80; Male, 18–44, 0.50; Male, 45–74, 0.60; Male, 75+, 0.75. The groups with older people have slightly higher response rates, which is generally considered to be true in surveys.

Scenario 3:

In Scenario 3 we allow the response rates for each group to vary between areas:

$$\text{logit}(q_{ij}) = \text{logit}(q_j) + b \times \epsilon_i, \qquad i = 1, \ldots, I,$$

where $\epsilon_i \sim_{iid} \text{N}(0, 1)$. The response rates in this scenario are formulated such that the average response rate for each group are the same as in scenario 2. We set $b = 0.35$ to give response rates with 10% and 90% quantiles of 0.46 and 0.81.

Scenario 4:

In Scenario 4 the underlying true prevalence rates include spatial dependency induced by adding a spatially correlated area-level covariate x_i:

$$\text{logit}(p_{ij}) = \text{logit}(p_j) + b \times x_i, \qquad i = 1, \ldots, I.$$

We choose $b = 0.2$ to allow sufficient variation in the prevalence rates between area. Across areas the 10% and 90% quantiles for prevalence p_{ij} are 0.013 and 0.20. The setup in this simulation scenario induces far greater variability in group prevalence than in the other scenarios. To simulate spatially correlated covariates x_i, we employ an ICAR model with mean 0 and conditional variance 1. Details on how to simulate from ICAR models can be found in Section 2.3.4. The purpose of this simulation is to investigate the effect of the underlying spatial dependency on small area estimation when the underlying cause of the

dependency is not observed. In this case, spatial models can be used as a surrogate for the unmeasured covariates. For the analysis using this scenario, we use the spatial model introduced in Section 3.3 but omit any area-level covariates to estimate the true counts.

Scenario 5:

In Scenario 5 we allow the response rate for each group to vary between areas by adding a spatial component to the variation:

$$\text{logit}(q_{ij}) = \text{logit}(q_j) + b \times x_i, \qquad i = 1, \ldots, I,$$

where x_i again is simulated from an ICAR model with mean 0 and conditional variance 1. We let $b = 0.3$ to give 10% and 90% quantiles for the response rate q_{ij} across areas as 0.49 and 0.79.

In Scenarios 1, 2, 3 and 5, the underlying diabetes prevalence p_i is considered to be the same for each area. In Scenario 4, the true prevalence rates exhibit spatial dependency and so a second set of prevalences are needed. The population sample sizes are the same in all scenarios.

To draw samples from the population, the sampling strategy we take is, for a particular zip code, to randomly draw individuals from the population. The sample size m_i is chosen to be the actual number of individuals who responded in the Washington 2006 BRFSS survey. For 9 areas with 1 sample only, we change the sample size to 2 in order to provide variance estimates.

3.4.2 Simulation Results

We analyze each simulated dataset using the empirical Bayes model, and the independent and spatial full Bayesian models. As a baseline, "conventional" estimates are also calculated with the unadjusted version based on y_i/m_i, and the adjusted version being (3.1). At the third stage of the full Bayesian model, we assume an improper uniform prior for β_0, and assign Gamma(0.5, 0.008) distributions to the precision parameters σ_v^{-2} and σ_u^{-2}. This prior gives the 95% of residuals odds in the range of $(0.5, 2.0)$ (Wakefield, 2009). We denote the estimated diabetes counts for each zip code using our proposed method as the "adjusted"

estimates. The diabetes counts are also estimated using observed number of diabetes, y_i and observed sample size, m_i for comparison, which we denote as the "unadjusted" estimates.

We compute three statistics to evaluate the estimates in the simulation study: the estimated squared bias, the estimated variance and the estimated mean squared error (MSE). Denote by S the total number of simulations, and y_i the "true" diabetes count in area i (which is the same across simulations). The summary statistics are calculated as:

$$\begin{aligned} \text{Bias} &= \frac{1}{I}\sum_{i=1}^{I}\left(\overline{\hat{y}}_i - y_i\right), \quad \text{where} \quad \overline{\hat{y}}_i = \frac{1}{S}\sum_{s=1}^{S}\hat{y}_i^{(s)}, \\ \text{Variance} &= \frac{1}{I}\sum_{i=1}^{I}\left(\sum_{s=1}^{S}(\hat{y}_i^{(s)} - \overline{\hat{y}}_i)^2\right), \\ \text{MSE} &= \text{Bias}^2 + \text{Variance}. \end{aligned}$$

Estimators with small MSE are considered superior, although among estimators with comparable MSE those with lower bias are preferred because they lead to interval estimates with improved calibration.

The results are presented in Table 3.1 with all results based on 100 simulations. In scenario 1, the unadjusted conventional estimator is approximately unbiased by construction and therefore has the smallest squared bias. The adjusted conventional estimator has slightly larger estimated bias. This is as expected because when non-response does not occur in the survey, nothing is gained by the adjusted estimator. Estimates from the random effect models (i.e., the empirical Bayes model, and the independent and spatial full Bayesian models) all have substantial bias due to the shrinkage towards the overall population prevalence, but reduced variance and MSE. The variance in the estimates is higher for the adjusted estimators than their unadjusted counterparts in the random effect models, due to the information loss in estimating the additional population group mean $\hat{p}_{\dot{j}}$ during post-stratification. This is true in general for all simulation scenarios.

Scenario 2 is a more practical situation, where there is non-response in the survey and response rates are different by age-and-gender group (but the response rate for each group remains constant across areas). In this case, the unadjusted conventional estimator is highly biased due to non-response and this bias can be reduced by post-stratification; this is the main purpose of post-stratification in large surveys. The reduction in bias carries over to the

empirical Bayes and Bayesian estimators based on adjusted data, and outweighs the increase in variance. The same message is shown again from the simulation results in scenarios 3 and 5.

The simulation setup in scenario 4 is similar to that in scenario 1 but allows more variability in the underlying prevalence. The results show an increase in both the bias and variance estimation under all models, due to the increased variation in the simulated data. However, our proposed method again provides a substantial reduction in estimated bias and also in MSE.

In scenario 4 and 5, we impose spatial dependency in the data but pretend the source of the dependency is unknown to us. The spatial model produces estimates with the smallest MSE. This is because the spatial models can serve as a surrogate for the dependency in the underlying prevalence, especially when the dependency can not be accounted for by adding the area-level covariates. When we include the spatially-varying covariate, the difference in the estimates between independent and spatial full Bayesian models diminishes (results not shown). In general, ignoring the sampling weights produces very poor inference, illustrating that using unadjusted hierarchical models with complex sampling schemes is not a good idea.

3.5 Application to 2006 WA BRFSS Data

We apply the sample weighted Bayesian hierarchical models we developed in Section 3.3 to the Washington 2006 BRFSS data introduced in Section 1.2.1. Sampling weights w_{ijk} are taken to be the final weight used in the BRFSS survey. For those 9 areas with only 1 observation, the effective sample size and effective number of observation are taken to be the same as the corresponding observed values.

Figure 3.1 presents the boxplots of logit-transformed estimated diabetes prevalence by zip code under different approaches. For the conventional unadjusted model, we employ the empirical logit transformation, i.e. $\log[(y_i+0.5/(m_i+0.5)]$. The figure shows a large amount of variation in the unadjusted conventional estimates due to large sampling variability, with the variability of the adjusted estimates being only slightly reduced. With our proposed adjustment, the variability of the estimates is reduced. Boxplots of the prevalence estimates

Table 3.1: Simulation summaries to compare estimated squared bias, variance and mean squared error for five different data generating scenarios, with four different models.

$\text{Bias}^2(\times 10^3)$		*Conventional*	*Emp Bayes*	*Indept Normal*	*Spatial Normal*
Scenario 1	Unadjusted	2	25	25	20
	Adjusted	5	19	19	15
Scenario 2	Unadjusted	16	58	59	49
	Adjusted	6	28	29	22
Scenario 3	Unadjusted	20	56	57	48
	Adjusted	10	26	26	21
Scenario 4	Unadjusted	3	39	40	30
	Adjusted	7	32	33	26
Scenario 5	Unadjusted	13	55	56	45
	Adjusted	6	26	27	21
$\text{Variance}(\times 10^3)$					
Scenario 1	Unadjusted	227	6	5	7
	Adjusted	204	14	14	13
Scenario 2	Unadjusted	252	6	5	6
	Adjusted	201	8	8	8
Scenario 3	Unadjusted	250	6	6	7
	Adjusted	199	9	8	9
Scenario 4	Unadjusted	235	8	8	8
	Adjusted	212	15	15	14
Scenario 5	Unadjusted	247	5	5	6
	Adjusted	197	8	8	8
$\text{MSE}(\times 10^3)$					
Scenario 1	Unadjusted	229	31	30	27
	Adjusted	209	33	33	28
Scenario 2	Unadjusted	268	64	64	55
	Adjusted	207	36	37	31
Scenario 3	Unadjusted	270	63	63	54
	Adjusted	208	35	35	30
Scenario 4	Unadjusted	238	47	48	38
	Adjusted	219	47	48	40
Scenario 5	Unadjusted	260	60	61	51
	Adjusted	203	34	35	29

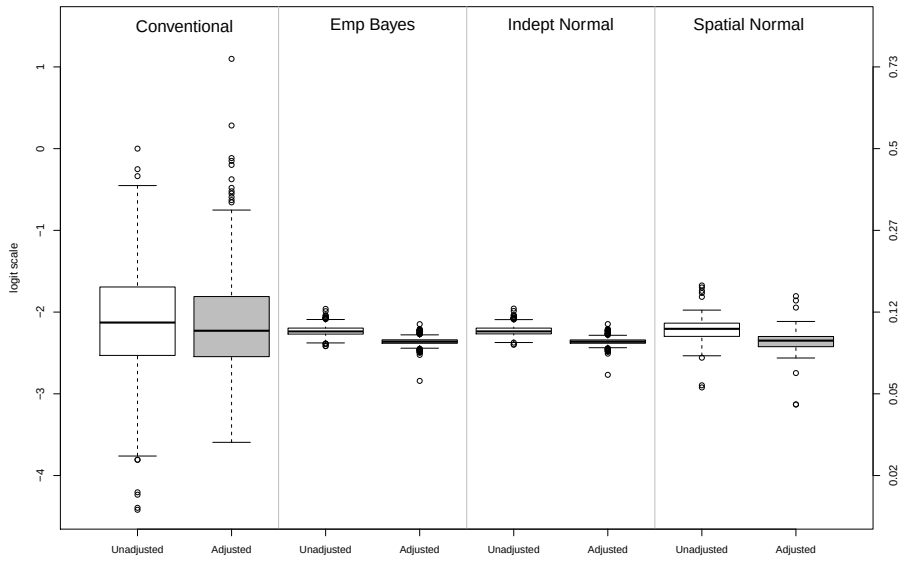


Figure 3.1: Estimated diabetes prevalence by zip code using models in Section 3.3: the left axis is on the logit scale and the right is on the $[0, 1]$ scale.

under the empirical Bayes model and the independent normal random effect are very similar in terms of both the mean and spread. The spatial model gives estimates with increased variation compared to other approaches. The take-home message here is that random effect models can greatly reduce the variance in estimation.

Figure 3.2 gives the map that we would report, based on the adjusted spatial model. There is higher diabetes population around Puget Sound area (the channel running north-south with with many small, highly populated, zip codes to the east) and the central south area. These areas correspond to the King county, Snohomish county, Spokane county and Yakima valley, which are the most populated counties in Washington State.

Figure 3.3 shows the difference in square root transformed total diabetes counts between the adjusted spatial model and the adjusted conventional model. We choose the square root transformation because it will approximately stabilize the variance for binomial counts when the prevalences are relatively small, and it is more interpretable than the arcsine-square root transformation which is variance stabilizing for binomial counts. We see lots

of differences, with a magnitude that is important; the totals in Figure 3.2 have a 10-90% range of (36,2121). There is clear spatial structure in the differences, as we might expect. Figure 3.4 shows the differences in square root transformed total diabetes counts between the adjusted and unadjusted spatial models. The diabetes prevalence estimate is lower for the adjusted spatial model than the unadjusted in almost all areas, as we saw to a greater or lesser extent with all models in Figure 3.1.

3.6 Discussion

In this chapter we have described a pragmatic approach to small area estimation, that allows spatial smoothing, and incorporates sample weights to acknowledge the design. By using the sample weights to adjust the data before estimation we separate the design-based survey computations and the model-based Bayesian shrinkage, allowing both components to be modified as the problem requires. The simulation study demonstrated much better performance in bias and variance reduction using our proposed approach under a number of difference scenarios.

To illustrate the effect of post-stratification we adopt in our method, we now provide some examples that compare the observed sample size and the effective sample size in the analysis. For zip code areas with moderate sample size and somewhat balanced samples in each age-by-gender groups, the effective sample sizes and effective number of cases defined in our approach should be close. Take zip code 98022 for example, the sample size for each age-by-gender group is Female, 18–44, 6; Female, 45–74, 6; Female, 75+, 2; Male, 18–44, 10; Male, 45–74, 4 and Male and 75+, 1. The effective sample size is 25 while the observed sample size is 29. The consequent effective number of cases is 1.7 while the observed is 2. There are other areas which show large differences. For example, zip code 98433 has an observed sample size 17. However, all these samples are from two age-by-gender groups with the lowest diabetes prevalence. The number of observed diabetes cases in this zip code is zero. After the adjustment with our proposed method, the effective sample size for this area is estimated as 475.2 with an effective number of cases of 12.2. The observed ratio of effective number of cases to sample size gives a naive estimate of the prevalence as 0.026, which is quite different to 0, which is the estimate based on the observed values.

Rao and Wu (2010) have recently proposed another way of combining survey design information and Bayesian models, through a version of empirical likelihood with a similar rescaling by effective sample size. They considered only whole-population mean estimation, but an extension of their approach to small-area estimation would be of interest.

In this chapter we focus on the bias that comes from non-response. There is another type of bias that comes from selection. In the future work, the selection bias will be our focus. Details about the plan for this can be found in Chapter 6.

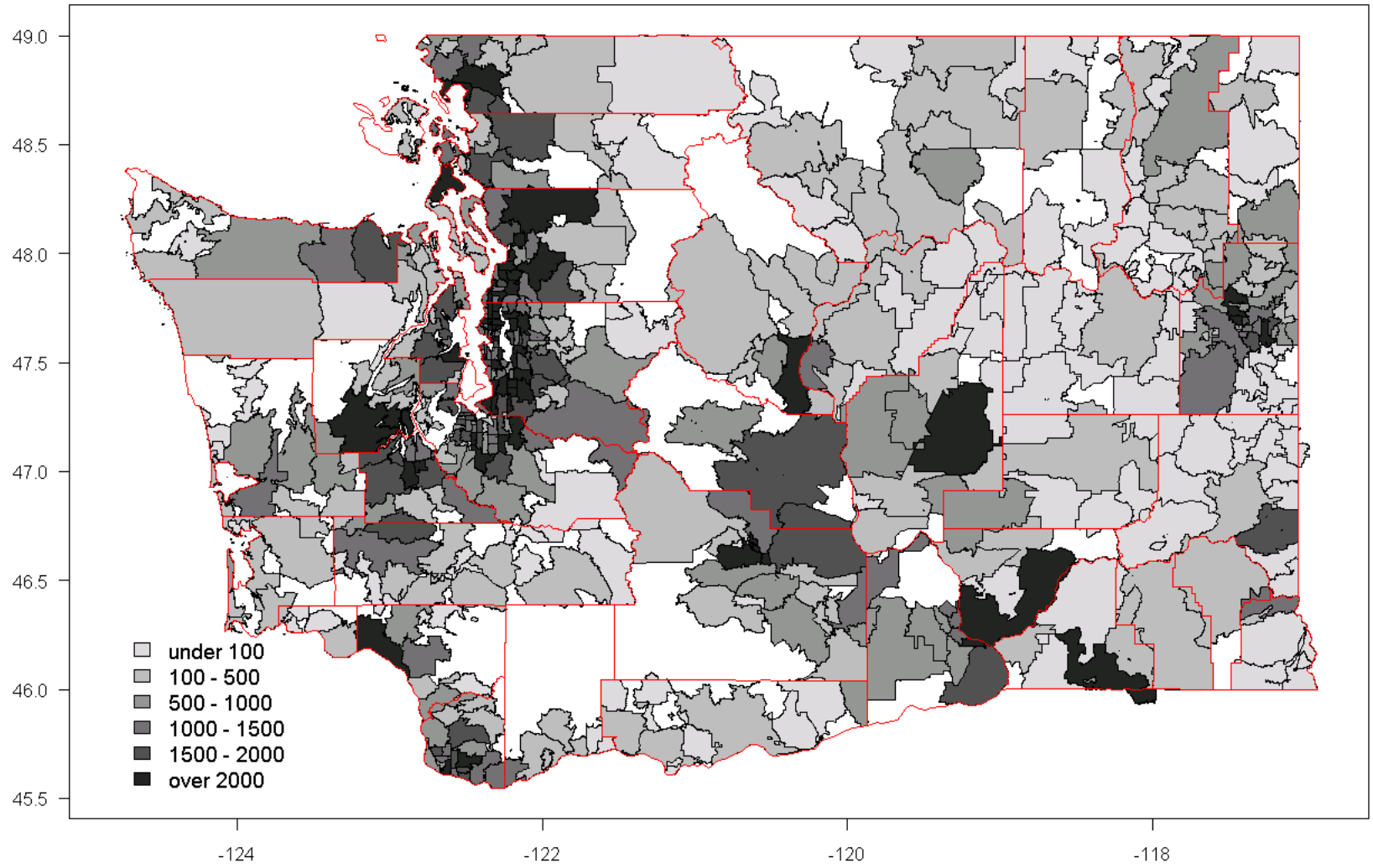


Figure 3.2: The adjusted estimates of the total diabetes counts by zip code in Washington State under the spatial model.

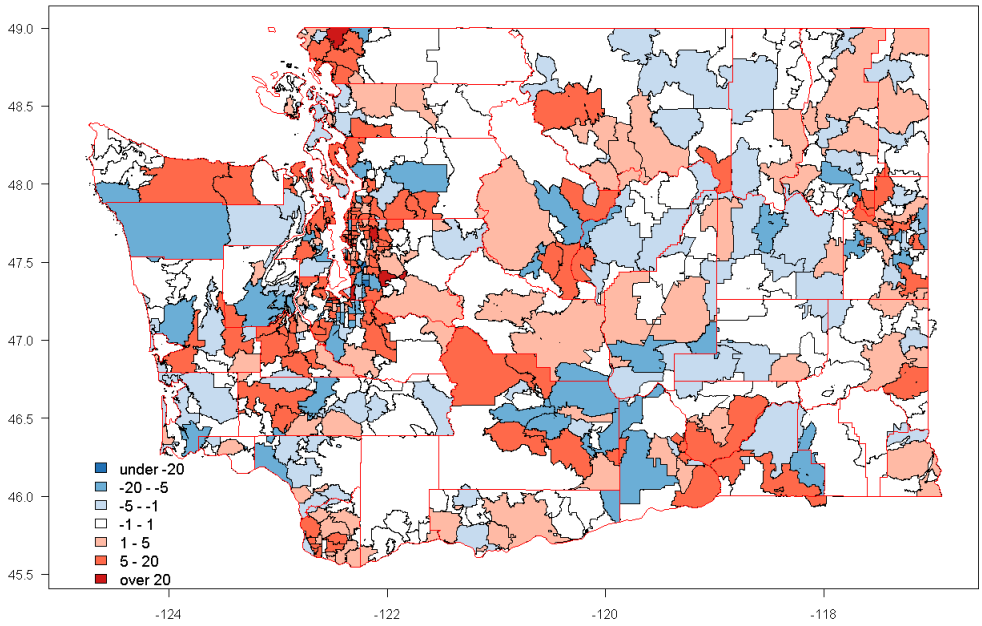


Figure 3.3: Map of the difference in the square root transformed estimated total diabetes counts between the adjusted spatial model and the adjusted conventional model.

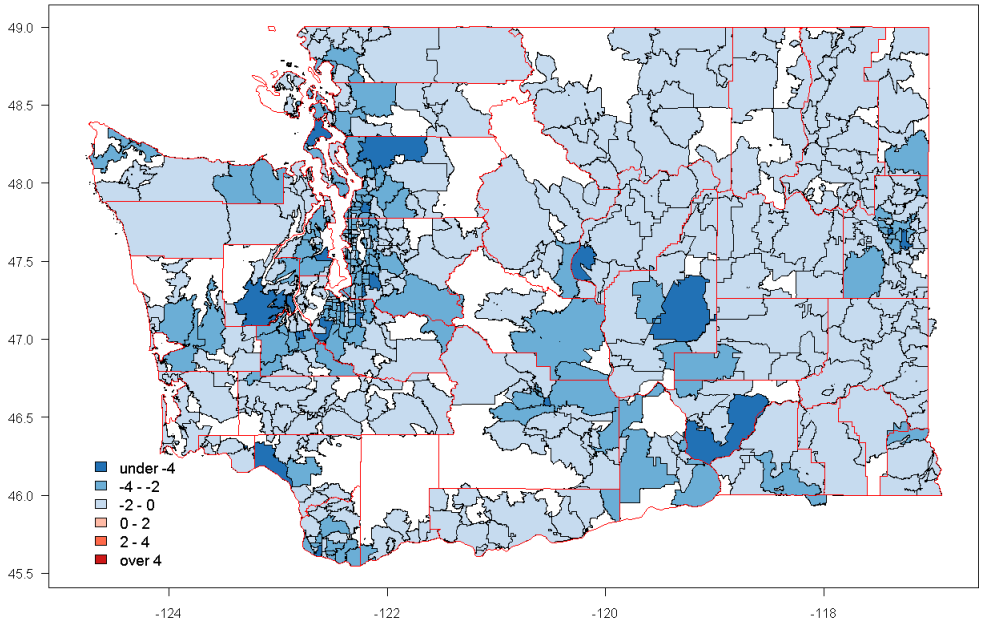


Figure 3.4: Map of the difference in the square root transformed total diabetes counts between adjusted and unadjusted spatial model.

Chapter 4

PENALIZED SPLINE MODELS FOR THE ANALYSIS OF SPATIO-TEMPORAL COUNT DATA

4.1 Introduction

In recent years, penalized spline models (P-splines) have become very popular in spatial-temporal analysis. Penalized splines were introduced by Eilers and Marx (1996), with the main idea being that the smoothing functions can be approximated by a set of B-spline functions (de Boor, 1978). To prevent overfitting, a penalty is imposed on the second or third order differences of the adjacent B-spline coefficients. Estimation of the B-spline coefficients is obtained by minimizing the penalized least squares function or by maximizing the penalized likelihood. The smoothing parameter used in the penalty term is typically chosen either by Akaike information criterion (AIC) or cross-validation. An alternative approach is to take a mixed model approach and directly estimate the smoothing parameter from the data. For multidimensional data, Currie et al. (2006) developed an algorithm for fast computation and low memory storage when applying P-splines in generalized linear models, which they called Generalized Linear Array Models (GLAMs). The algorithm takes advantage of the array structure in the data, and uses a series of nested matrix operations during the computation of the penalized Fisher scoring function. Many authors have followed this approach, in spatial-temporal analysis, Ugarte et al. (2010) and Lee and Durbán (2011) included a three-way interaction constructed using B-splines which can be decomposed into components similar to those in analysis of variance (ANOVA) spline models (Gu, 2002).

The Bayesian version of the P-spline models was introduced by Fahrmeir and Lang (2001). The Bayesian analogy of imposing the penalty term on the adjacent B-spline coefficients in Eilers and Marx (1996) is to use a random walk prior on these coefficients. Fahrmeir and Lang (2001) used a generalized additive model with the linear predictor represented

by a set of unknown smooth functions of the covariates, including covariates representing spatial and temporal random effects. The ICAR prior from Besag et al. (1991) as described in Section 2.2 was used for the spatial covariate. Work based on similar ideas in a spatial-temporal setting include Lang and Brezger (2004), Fahrmeir et al. (2004) and Kneib and Fahrmeir (2006).

The penalized spline model we develop for spatio-temporal count data uses a tensor product of B-splines as the basis functions for spatial smoothing. We include an interaction between space and time with different types of space-time structures on the regression coefficients. Compared to the methods mentioned above using B-splines for spatial-temporal analysis, our model has the following advantages. First and foremost, the set of B-spline coefficients in our model actually has some meaning – they are the indicators of the spatial or temporal breakouts for the infectious disease data we analyze. Therefore, our model provides a tool to gain insight into the movement of the epidemic centers and the breakout time points of the infectious disease, which are usually of primary interest to public health authorities. Second, the interaction term in our model has far less parameters than the three-way interactions in Ugarte et al. (2010) and Lee and Durbán (2011), which makes the computation much easier.

This chapter is organized as follows: in Section 4.2 we give a brief review of the spline models for the univariate case. In Section 4.3, we describe the spline models used in spatial smoothing, with a comparison between the radial basis functions and the tensor product of B-spline functions. In Section 4.4, we give details of the penalized spatial-temporal models we develop and then demonstrate its performance through a simulation study. The application of our model to the China Hand-Foot-Mouth disease data is presented in Section 4.5. Finally, in Section 4.6 we present conclusions and future work.

4.2 Spline Models

Let $f(x)$ denote a function mapping from $\mathbb{R}$ to $\mathbb{R}$. We assume that function $f(x)$ can be presented by a series of polynomial spline functions

$$f(x) = \sum_{j=1}^{J} \beta_j h_j(x),$$

where $h_j(x)$ is the j-th basis function of x and β_j is the coefficient associated with basis function h_j. Higher degree polynomials usually provide a smoother spline fit, as they have a higher degree of continuous derivatives. There are a variety of basis functions to choose from, with the two most popular being the truncated power functions and the B-spline functions. Below we briefly introduce these basis functions in the univariate case.

4.2.1 Truncated Power Splines

With a truncated power spline of degree p we have

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \sum_{k=1}^{K} b_k (x - \kappa_k)_+^p, \tag{4.1}$$

where $(x - \kappa_k)_+$ takes the positive part of $x - \kappa_k$ and is 0 for x less than κ_k. The κ_ks, $k = 1, \ldots, K$ are a set of K knots. Truncated power splines of degree p therefore have $p + K + 1$ basis functions.

It is easy to see that the spline function $f(x)$ can be represented by a linear model, with the design matrix constructed from the basis functions. Therefore, estimation of the basis coefficients can be made with the usual linear model methods that are easily implemented. This partially explains why spline models are popular.

In Figure 4.1, we present plots of the cubic truncated power splines (i.e., $p = 3$), with two selected knots at $\kappa_1 = 0.33$ and $\kappa_2 = 0.67$.

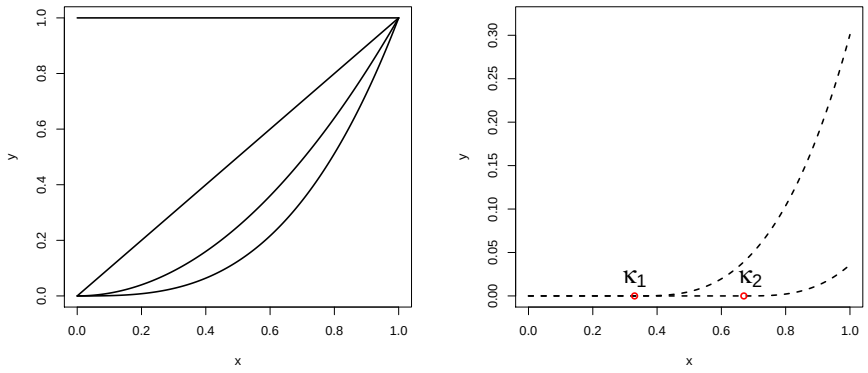
4.2.2 B-splines

A B-spline basis function of degree p is constructed recursively from lower degree B-spline basis functions (de Boor, 1978):

$$B_k^p(x) = \frac{x - \kappa_k}{\kappa_{k+p} - \kappa_k} B_k^{p-1}(x) + \frac{\kappa_{k+p+1} - x}{\kappa_{k+p+1} - \kappa_{k+1}} B_{k+1}^{p-1}(x).$$

A B-spline basis with degree $p = 0$ is just a set of piecewise constant basis functions

$$B_k^0(x) = \begin{cases} 1 & \text{if } \kappa_k \le x \le \kappa_{k+1} \\ 0 & \text{otherwise} \end{cases}$$



(a) (b)

Figure 4.1: Basis functions of cubic truncated power splines with two knots at $\kappa_1 = 0.33$ and $\kappa_2 = 0.67$. The left panel shows the bases $1, x, x^2$ and x^3. The right panel shows the bases $(x - \kappa_1)^3_+$ and $(x - \kappa_2)^3_+$.

A B-spline model with degree p is then

$$f(x) = \sum_{j=1}^{K+p+1} \beta_j B_j^p(x) \tag{4.2}$$

It is easy to see that B-splines can also be represented by a linear model similar to the truncated power splines. However, compared to the truncated splines, B-splines have the advantage of easy computation as B-splines are only positive on a domain spanned by $p + 2$ knots and are zero elsewhere. There is an equivalence between the B-splines and the truncated power splines: if one uses the same degree and same number of knots with the same locations, then the fit would be identical from the B-spline and the truncated power spline.

In Figure 4.2 we present the basis functions of a cubic B-spline, which has the form

$$f(x) = \sum_{j=1}^{K+4} \beta_j B_j^3(x) \tag{4.3}$$

We choose $K = 4$ equally-spaced inner knots indicated by the red circles between -0.5 and 0.5.

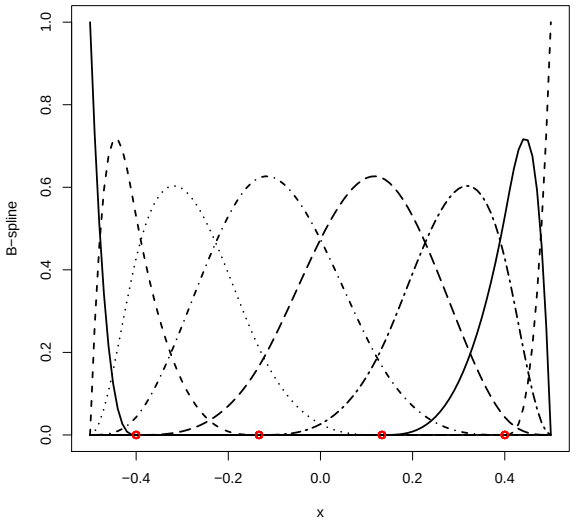


Figure 4.2: Basis functions of a cubic B-spline, with $k = 4$ inner knots indicated by red circles.

4.3 Spatial Smoothing

For spatial smoothing we need a spline model with bivariate basis functions. Two types of functions are considered in this section due to their widespread popularity. The first type is to use bivariate radial basis functions; the second is to use the tensor product of two univariate splines. We start with constructing bivariate splines using radial basis functions, list some commonly used radial basis functions, and describe the algorithm for choosing the knot locations. We then introduce bivariate splines using the tensor product of two univariate B-splines. We conclude by comparing these two approaches based on a simulated data set.

Throughout this section, we use $\boldsymbol{x}$ to denote a bivariate variable, $\boldsymbol{x} = (x_1, x_2)$. In spatial smoothing, $\boldsymbol{x}$ usually refers to the geographic location of the observation, with x_1 and x_2 as the first and second coordinate of the location (i.e., the latitude and longitude). The bivariate spline model is now $f(\boldsymbol{x}) = f(x_1, x_2)$, which is still a function on $\mathbb{R}$.

4.3.1 Radial Basis Functions

The bivariate smoothing using radial basis functions is closely connected to geostatistical kriging, as illustrated in detail in Kammann and Wand (2003). The general form of $f(\boldsymbol{x})$ is

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \beta_k C(\|(x_1, x_2), (\kappa_{x_1}^k, \kappa_{x_2}^k)\|), \tag{4.4}$$

where $C(\|(x_1, x_2), (\kappa_{x_1}^k, \kappa_{x_2}^k)\|)$ is the radial basis function, K is the number of knots and $(\kappa_{x_1}^k, \kappa_{x_2}^k), k = 1, \ldots, K$ are the selected knots. The function $\|(x_1, x_2), (\kappa_{x_1}^k, \kappa_{x_2}^k)\|$ denotes the Euclidian distance between the point (x_1, x_2) and the k-th knot $(\kappa_{x_1}^k, \kappa_{x_2}^k)$. Therefore, the correlation between point $\boldsymbol{x}$ and the knot $\boldsymbol{\kappa}$ depends on the distance, and not on the direction.

The basis function C in (4.4) is adopted from geostatistical kriging. In geostatistical kriging, we often assume the underlying stochastic process S is isotropic, meaning the correlation between locations only depends on the distance between them. Therefore, the correlation function of the process S is $C(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) = C(h_{ij})$, where h_{ij} is the distance between location $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The most commonly used empirical model for correlation is the Matern class of correlation functions

$$C(h) = \frac{1}{\Gamma(\nu)} \left(\frac{h}{2\rho}\right)^{\nu} 2K_\nu(h/\rho) \qquad \nu > 0, \rho > 0$$

where K_ν is the modified Bessel function of the second kind of order $\nu > 0$. Parameter ρ is called the range parameter and indicates the distance beyond which there is essentially no spatial correlation. The Matern model includes several special cases, for example:

- Exponential model ($\nu = 1/2$):

$$C(h) = \exp(-h/\rho), \qquad h > 0$$

- Gaussian model ($\nu \to \infty$):

$$C(h) = \exp(-h^2/\rho^2), \qquad h > 0$$

- Matern at $\nu = 3/2$:

$$C(h) = (1 + h/\rho)\exp(-h/\rho), \qquad h > 0$$

In geostatistical kriging, the function $C(h)$ is usually decided based upon some empirical analysis such as the (semi)variogram (Diggle, 2007). Once a function is selected, the range parameter ρ is estimated via some statistical approaches such as maximum likelihood. However, in the spline models, there is no general guide as to which basis function should be used under any specific circumstance. Also, there is no evidence that one particular form outperforms the other. Estimation of the range parameter ρ is generally difficult as there is often not sufficient information in the data about this parameter. In Kammann and Wand (2003), the authors choose Matern model at $\nu = 3/2$ and fix ρ as the maximum distance between the locations for scale invariance and numerical stability. Fixing the range parameter *a priori* is a rather ad hoc way and may require sensitivity analysis. In our comparison analysis, we adopt the exponential model, i.e. the model with $\nu = 1/2$ and use the same treatment for the parameter ρ as in Kammann and Wand (2003).

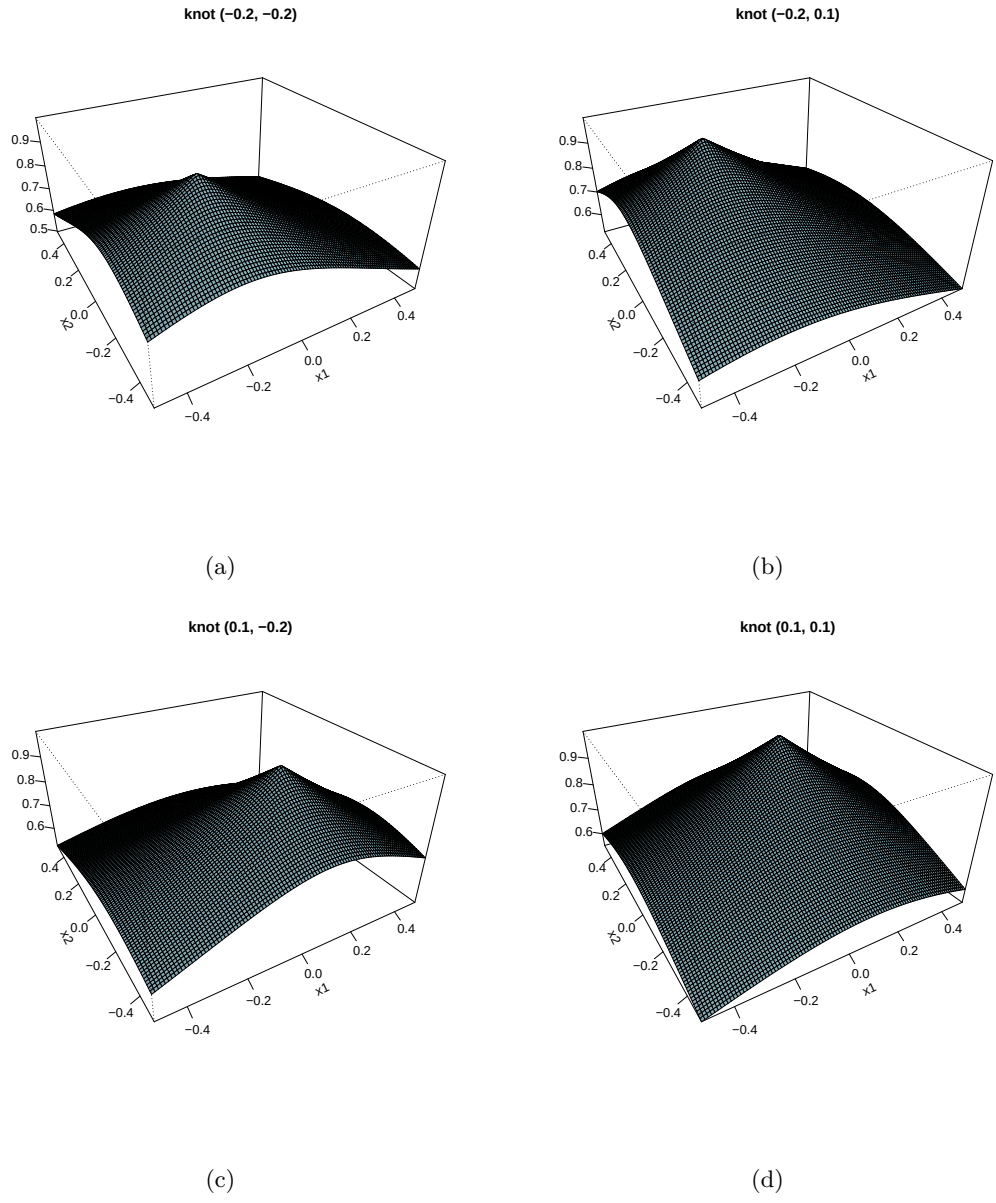
In Figure 4.3 we show the exponential radial basis functions corresponding to four selected knots.

Number of Knots and Knot Location

Too many knots lead to overfitting while too few knots lead to underfitting. The number of knots and the knot locations can be chosen in a automatic fashion, see for example, Friedman and Silverman (1989), Friedman (1991) and Kooperberg and Stone (1992).

When the observed data is of moderate size, all data locations can be used as knots, which leads to a full-rank spline. In cases where the data is of large size, using all of the locations can be computationally expensive. The remedy is to use a low-rank spline, as proposed by Nychka and Saltzman (1998). The idea of a low-rank spline is to choose a subset of the data $(\boldsymbol{x}'_1, \cdots, \boldsymbol{x}'_K)$ where K is much smaller than the number of observations, and use the selected subset as the knots. Low-rank spline models provide fast and stable computation, and also good approximation to the full rank case.

Once the number of knots is decided we need to choose the knot positions. In the one-dimensional case, the knots are usually equally spaced though they can depend on the density of the x_i points. In higher dimensions such as two-dimensional geo-referenced



(a) (b)

(c) (d)

Figure 4.3: Radial basis functions using the exponential model. The study region is a $[-0.5, 0.5] \times [-0.5, 0.5]$ square and the range parameter ρ is taken as 1.41.

data, knots are usually chosen in "clusters" using space-filling algorithms (Kaufman and Rousseeuw, 1990). The algorithm is based on selecting K representative points, called *medoids*, from all observed points. The space-filling algorithm for finding the K knots

proceeds as follows:

1. BUILD step:

   (a) Find the initial object $\boldsymbol{x}^*_{(1)}$ that has the smallest average Euclidian distance to all other point locations.

   (b) To decide which of the rest points to include in the final set of knots, first choose a non-selected point $\boldsymbol{x}_i$. For point $\boldsymbol{x}_i$, calculate

   $$\sum_{\boldsymbol{x}_j \neq \boldsymbol{x}^*_{(1)}} \left[\max(D_j - d(\boldsymbol{x}_i, \boldsymbol{x}_j), 0)\right], \tag{4.5}$$

   where D_j is the distance between point $\boldsymbol{x}_j$ and the closest point that has been included in the knot set, and $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

   (c) Choose the point $\boldsymbol{x}_i$ that maximizes (4.5).

   (d) Repeat (b) and (c) until K points $S^* = (\boldsymbol{x}^*_{(1)}, \boldsymbol{x}^*_{(2)}, \cdots, \boldsymbol{x}^*_{(K)})$ are selected.

2. SWAP step:

   (a) Consider a pair of points $(\boldsymbol{x}^*_i, \boldsymbol{x}^{-*}_j)$ where $\boldsymbol{x}^*_i$ is a point that was selected for knots (i.e. $\boldsymbol{x}^*_i \in S^*$) and $\boldsymbol{x}^{-*}_j$ is a point that was not selected (i.e. $\boldsymbol{x}^*_i \notin S^*$). We need to calculate the effect of replacing $\boldsymbol{x}^*_i$ with $\boldsymbol{x}^{-*}_j$ in S^*. Let $\boldsymbol{x}^{-*}_h$ be a non-selected point $h \neq j$, define the contribution C_{ijh} to swap $\boldsymbol{x}^*_i$ and $\boldsymbol{x}^{-*}_j$:

   $$C_{ijh} = \begin{cases} 0 & \text{if } \min(d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_i), (\boldsymbol{x}^{-*}_h, \boldsymbol{x}^{-*}_j)) > \\ & d_{s\in S^*, s\neq i}(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_s), \\ d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^{-*}_j) - d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_i) & \text{if } d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_i) \leq d_{s\in S^*, s\neq i}(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_s) \text{ and} \\ & d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^{-*}_j) < E_h, \\ E_h - d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_i) & \text{if } d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_i) \leq d_{s\in S^*, s\neq i}(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_s) \text{ and} \\ & d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^{-*}_j) \geq E_h, \\ d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^{-*}_j) - d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_i) & \text{if } d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_i) > \min(d_{s\in S^*, s\neq i}(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_s)) \text{ and} \\ & d(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^{-*}_j) > d_{s\in S^*, s\neq i}(\boldsymbol{x}^{-*}_h, \boldsymbol{x}^*_s), \end{cases}$$

   where E_h is the distance between $\boldsymbol{x}^{-*}_j$ and its second closest point in S^*.

(b) Choose the pair $(\boldsymbol{x}_i^*, \boldsymbol{x}_j^{-*})$ which minimizes $\sum_h C_{ijh}$. If the resulting minimum is negative, swap the pair. Otherwise, go back to (a).

In the SWAP step, (a) and (b) is considered for all potential swaps. The resulting K medoids are the desired knots.

The k-medoid method for selecting representative points usually works well when the data set is of moderate size. When the data set is large, for example, several thousand points or more, the algorithm described above can be slow. To deal with large datasets, the algorithm is modified as follows:

1. Draw a trial set of sample points from the data and apply the k-medoid method described above to yield K knots.

2. Assign each of the point that was not in the sample to the nearest knots.

3. Calculate the average distance between each of the data points and the selected knots.

4. Repeat 1–3 for several times (usually 5 or more) and the knot set with the lowest average distance is selected.

4.3.2 Tensor Product of Cubic B-splines

Let B_{k_1} and B_{k_2} be the B-spline basis created for coordinates x_1 and x_2, respectively. The bivariate smoothing function $f(\boldsymbol{x})$ is constructed using the tensor product of the B-spline basis:

$$\begin{aligned} f(x_1, x_2) &= \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \gamma_{k_1 k_2} B_{k_1}(x_1) B_{k_2}(x_2) \\ &= \sum_{k=1}^{K} \gamma_k B_k(x_1, x_2) \end{aligned} \tag{4.6}$$

In matrix format, the function can be represented as the Kronecker product of two matrices:

$$f(x_1, x_2) = \boldsymbol{B}_{k_1}(x_1) \otimes \boldsymbol{B}_{k_2}(x_2)$$

The bases from a cubic tensor product B-spline using 4 knots for each coordinate (and therefore 64 parameters in total) shown in Figure 4.4.

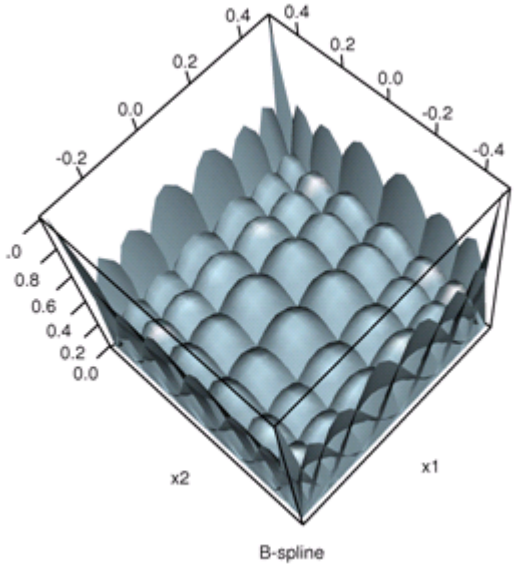


Figure 4.4: Illustration of tensor product cubic B-spline basis functions, with K = inner knots for each dimension.

4.3.3 Comparison of the Bivariate Basis Functions

In this section, we use a simulated Poisson dataset to compare the performance of the spline models with radial basis functions and tensor products of cubic B-spline functions. The details of the simulation study are as follows:

1. The underlying log relative surface is assumed to be $f(x_1, x_2) = -3\exp(-2x_1^2) + 2\exp(-3x_2^2)$. The plot of the true underlying surface is shown in Figure 4.5 (a).

2. The sample size is chosen to be $n = 150$. The locations of the sample are chosen to be uniformly distributed within a $[-0.5, 0.5] \times [-0.5, 0.5]$ square region.

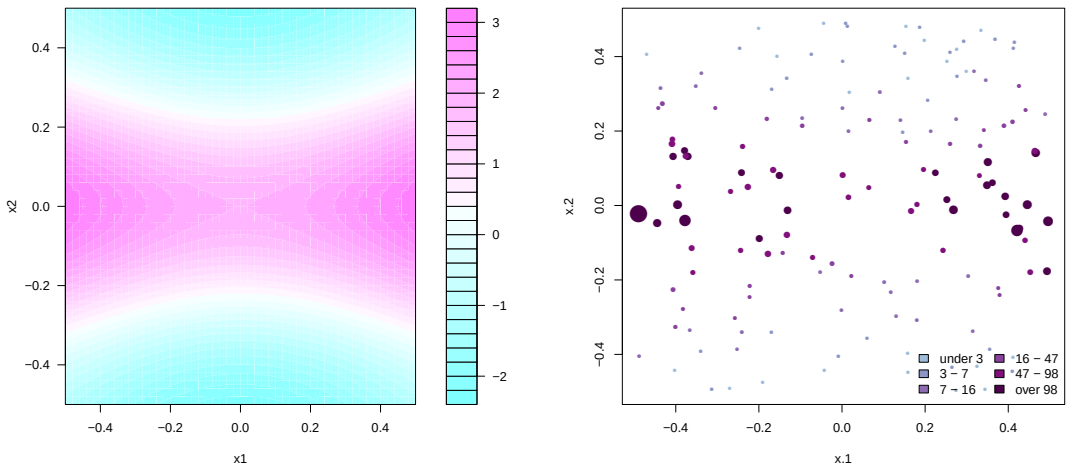
3. We choose the expected number of cases E to be between 10 and 50 and simulate the observed number of cases Y for each of the n locations from a Poisson distribution. The range of simulated observations Y is from 0 to 426.

4. Then we fit the following two models to the simulated data. The simulated observa-

tions is shown in Figure 4.5.

- Model 1 uses an exponential radial function with ρ fixed as the maximum distance between the sampled locations. We use the space-filling algorithm to select 36 knot locations. The resulting knots are shown in Figure 4.6 (a).
- Model 2 uses the tensor product of cubic B-spline basis functions. We choose 2 inner knots for each univariate cubic B-spline which gives 36 total bases. A contour of the resulting basis functions is shown in the right panel of Figure 4.6 (b).

5. Both models are fit in a Bayesian framework with a diffuse prior on the coefficients of the basis functions $\boldsymbol{b}$, i.e., $\boldsymbol{b} \sim \mathrm{N}(0, \tau_b)$ with $\tau_b \sim \mathrm{Gamma}(1, 0.005)$.

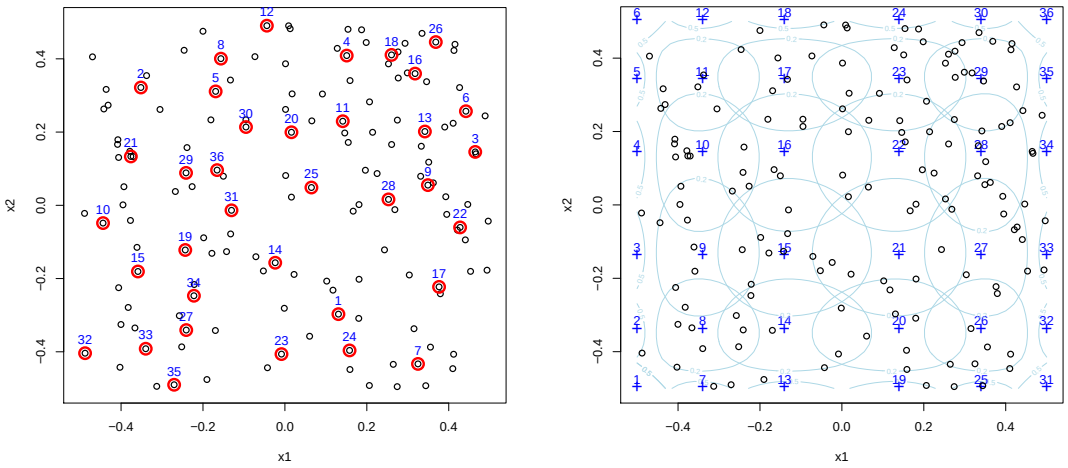


(a) Simulated log risk surface

(b) Simulated Poisson observations

Figure 4.5: Simulated surface and Poisson observations.

The simulated data are shown in Figure 4.5 (b). The fitted surfaces using exponential radial basis functions and tensor product cubic B-spline functions are shown in Figure 4.7. It is clear that the fitted surface using tensor product cubic B-splines resembles the true surface quite (except at the edges), while the fitted surface with the exponential radial basis



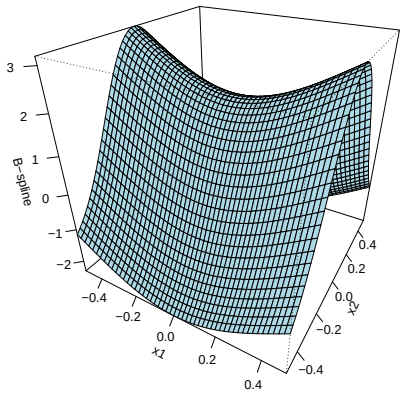
(a) Radial basis functions

(b) Tensor product of cubic B-splines

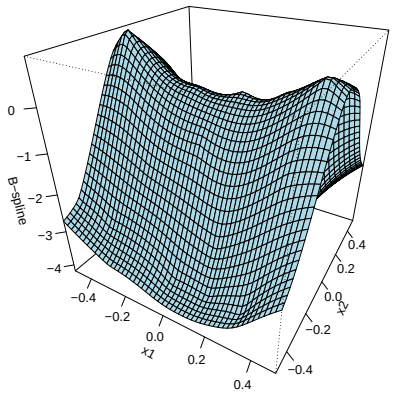
Figure 4.6: Radial basis functions and tensor product B-spline basis functions. Data locations are shown in black circles. In panel (a), the selected knots for the radial basis functions are shown as red circles. In panel (b), the contour lines of tensor product cubic-B spline are shown as blue lines. The peak of each basis functions is indicated with a blue "+".

functions shows some undesirable bumps. In Figure 4.8, we compare the estimated log relative risks for each area for both the exponential and tensor product of cubic B-spline models. The estimated log relative risks from the tensor product cubic B-spline model is again much closer to the truth than that from the exponential radial basis spline models.

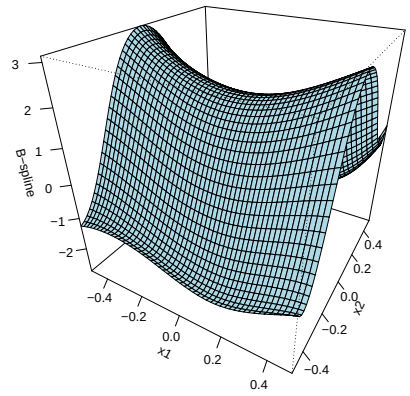
The simulation we have conducted to investigate spatial smoothing shows a superior performance for the tensor product B-spline model than for the radial basis spline model. Therefore, we choose to use the tensor product of cubic B-splines as the basis functions in the spatial-temporal interaction model described in the next section.



(a) Simulated surface



(b) Fitted surface with exponential radial basis



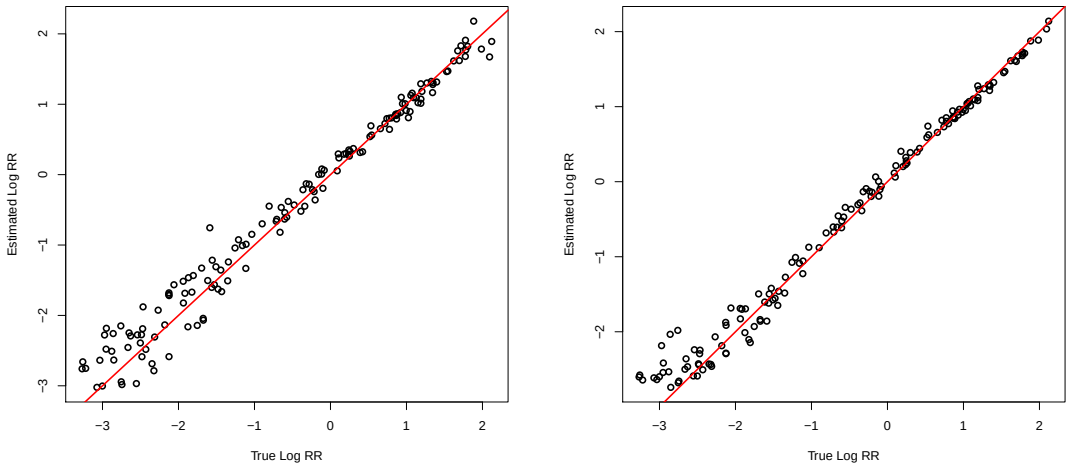
(c) Fitted surface with tensor product of cubic B-splines

Figure 4.7: Comparison of the simulated and fitted surfaces.

4.4 Spatial and Temporal Interaction Model

A full penalized spline models with spatio-temporal interaction may be described as follows:

$$\begin{aligned} y_{it}|\mu_{it} &\sim \text{Poisson}(E_i\mu_{it}), \qquad i = 1,\ldots,I, \quad t = 1,\ldots,T \\ \log(\mu_{it}) &= \alpha + f_1(t) + f_2(\boldsymbol{x}_i) + \sum_{k=1}^{K} b_{kt}B_{ik}(\boldsymbol{x}_i), \end{aligned} \tag{4.7}$$



(a) Exponential radial basis

(b) Tensor product basis

Figure 4.8: Result comparison, with independent normal prior for the regression coefficients.

where $f_1(t)$ is the temporal smoothing component and $f_2(\boldsymbol{x}_i)$ is the spatial smoothing component with $\boldsymbol{x}_i$ indicating the location of area i. The interaction term is $\sum_{k=1}^{K} b_{kt}B_{ik}(\boldsymbol{x}_i)$ with $B_{ik}(\boldsymbol{x}_i)$ the tensor product cubic B-spline basis functions introduced earlier. The structure of the interaction between space and time is imposed through the prior on the basis coefficients b_{st} as described in a disease mapping contest in Section 2.2. Depending on the specific application, there are four types of interactions that we consider:

1. Independent normal prior on b_{kt}

$$\pi(\boldsymbol{b}|\tau_b) \propto \exp\left(-\frac{\tau_b}{2}\sum_{k=1}^{K}\sum_{t=1}^{T} b_{kt}^2 \right).$$

2. Second-order random walk (RW2) prior

$$\pi(\boldsymbol{b}|\tau_b) \propto \exp\left(-\frac{\tau_b}{2}\sum_{k=1}^{K}\sum_{t=3}^{T} (b_{kt} - 2b_{k,t-1} + b_{k,t-2})^2 \right).$$

3. Spatial intrinsic conditional autoregressive (ICAR) prior

$$\pi(\boldsymbol{b}|\tau_b) \propto \exp\left(-\frac{\tau_b}{2}\sum_{t=1}^{T}\sum_{k\sim k'} (b_{kt} - b_{k't})^2 \right).$$

4. The Kronecker product of the RW2 and ICAR priors

$$\pi(\boldsymbol{b}|\tau_b) \propto \exp\left(-\frac{\tau_b}{2} \sum_{t=3}^{T} \sum_{k\sim k'} \Big((b_{kt} - 2b_{k,t-1} + b_{k,t-2}) - (b_{k't} - 2b_{k',t-1} + b_{k',t-2}) \Big)^2 \right).$$

The interaction term $\boldsymbol{b}$ has a fairly complicated structure, especially for Types 2, 3 and 4. However, all four types can be represented in Gaussian Markov Random Field (GMRF) models as described in Section 2.2:

$$\pi(\boldsymbol{b}|\tau_b) \quad \propto \quad \tau_b^{\text{rank}(\boldsymbol{Q})/2} \exp\left(-\frac{1}{2} \boldsymbol{b}^T \boldsymbol{Q} \boldsymbol{b} \right).$$

Here $\boldsymbol{Q} = \tau_b \boldsymbol{K}_{\mathcal{B}}$ is the precision matrix for the vector $\boldsymbol{b}$ which has the elements

$$\boldsymbol{b} = (b_{11}, \ldots, b_{K1}, b_{12}, \ldots, b_{K2}, \cdots, b_{1T}, \ldots, b_{KT})^T,$$

We let $\boldsymbol{K}_{\mathcal{S}}$ and $\boldsymbol{K}_{\mathcal{T}}$ denote the structure matrix for space and time respectively, the structure matrix $\boldsymbol{K}_{\mathcal{B}}$ can be expressed as the Kronecker product of the two matrices $\boldsymbol{K}_{\mathcal{T}} \otimes \boldsymbol{K}_{\mathcal{S}}$. Depending on the structure matrix we choose for $\boldsymbol{K}_{\mathcal{S}}$ and $\boldsymbol{K}_{\mathcal{T}}$, the interaction term can take different forms:

1. For the Type 1 prior the structure matrix $\boldsymbol{K}_{\mathcal{B}}$ is constructed as the Kronecker product of $\boldsymbol{K}_{\mathcal{T}}$ and $\boldsymbol{K}_{\mathcal{S}}$ given by

$$\boldsymbol{K}_{\mathcal{T}} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}_{T\times T}, \qquad \boldsymbol{K}_{\mathcal{S}} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}_{K\times K},$$

   where the blank space corresponding to zeros. This prior assumes that there is no spatial or temporal trend in the spline coefficients.

2. For the Type 2 prior we assume a RW(2)prior for the temporal trend but no spatial trend in the spline coefficients. Therefore, the structure matrix $\boldsymbol{K}_{\mathcal{T}}$ takes the form (2.13). The structure matrix $\boldsymbol{K}_{\mathcal{S}}$ remains as the identity matrix that was used in the Type 1 prior. Note that $\boldsymbol{K}_{\mathcal{T}}$ has rank $T-2$, and therefore the rank of the structure matrix $\boldsymbol{K}_{\mathcal{B}}$ is $K(T-2)$.

3. For the Type 3 prior the structure matrix $\boldsymbol{K}_{\mathcal{S}}$ has the adjacency matrix defined in (2.8) for the ICAR model. The temporal structure matrix $\boldsymbol{K}_{\mathcal{T}}$ is an identity matrix. The rank of the structure matrix $\boldsymbol{K}_{\mathcal{B}}$ is $T(K-1)$.

4. For the Type 4 prior the structure matrix $\boldsymbol{K}_{\mathcal{T}}$ for the temporal pattern is as with the Type 2 prior and the structure matrix $\boldsymbol{K}_{\mathcal{S}}$ for the spatial pattern is as with the Type 3 prior. The Kronecker product of these two matrices yields the structure matrix $\boldsymbol{K}_{\mathcal{B}}$, which has rank $(T-2)(K-1)$. This type of prior is of particular interest in an infectious disease context since the estimation of the basis coefficient b_{kt} at basis k and time t depends on those from the two previous time periods $b_{k,t-1}$ and $b_{k,t-2}$, as well as those from the neighboring knots $b_{k',t}, k' \sim k$. In addition, b_{kt} depends on those in the neighboring areas knots in the two previous time points $b_{k',t-1}$ and $b_{k',t-2}, k' \sim k$. So estimation of the coefficients borrows information in both space and time.

4.5 A Simulation Study

4.5.1 Simulated Data

We conduct a simulation study to investigate the performance of the four types of interaction model proposed in Section 4.4. We first simulate data using the following procedures:

1. Let the study region be a $[-0.5, 0.5] \times [-0.5, 0.5]$ square. We randomly select 150 locations (i.e., $I = 150$) at which we observe the counts.

2. We select 3 equal-distance inner knots for each coordinate and create the tensor product cubic B-splines. The contour plot of the basis functions is shown in Figure 4.9. In this plot, blue lines are the contour lines for the tensor product cubic B-spline bases. The red circles are 150 randomly selected locations at which the responses are observed. We use a total $K = 36$ bases in the simulation study.

3. We create the true underlying risk surface by assigning values to the set of basis coefficients b_{kt}. We assume 24 discrete time points for observations (i.e., $T = 24$). The intercept α is set to be -1.5. The values of the coefficients are chosen such that

higher values are given to those in the bottom left corner at early stages of the study period, those in the top left corner at middle stages, and those in the top right corner at later stages. The true b_{kt}, along with their corresponding locations, are shown in Figure 4.10.

4. We choose the expected number of cases E_i to be between 5 and 50 and assume they are constant over the study period. We then simulate the observed counts Y_{it} at the selected locations, based on the model

$$\log(\mu_{it}) = \alpha + \sum_{k=1}^{K} b_{kt} B_{ik}, \quad k = 1, \ldots, 36, \quad T = 1, \ldots, 24 \tag{4.8}$$

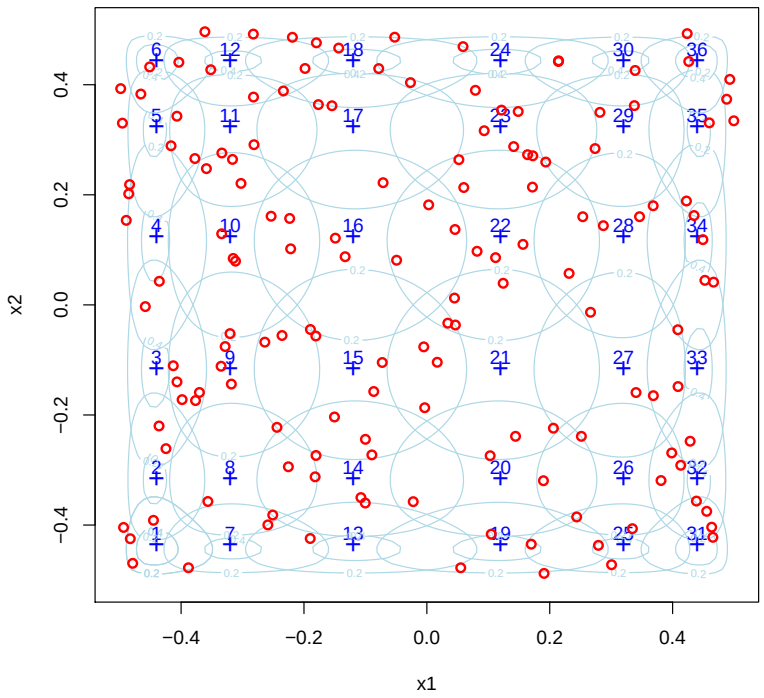


Figure 4.9: Contour plot of the tensor product cubic B-spline basis function: blue lines are the contour lines for the tensor product cubic B-spline bases. The red circles are 150 randomly selected locations at which the responses are observed.

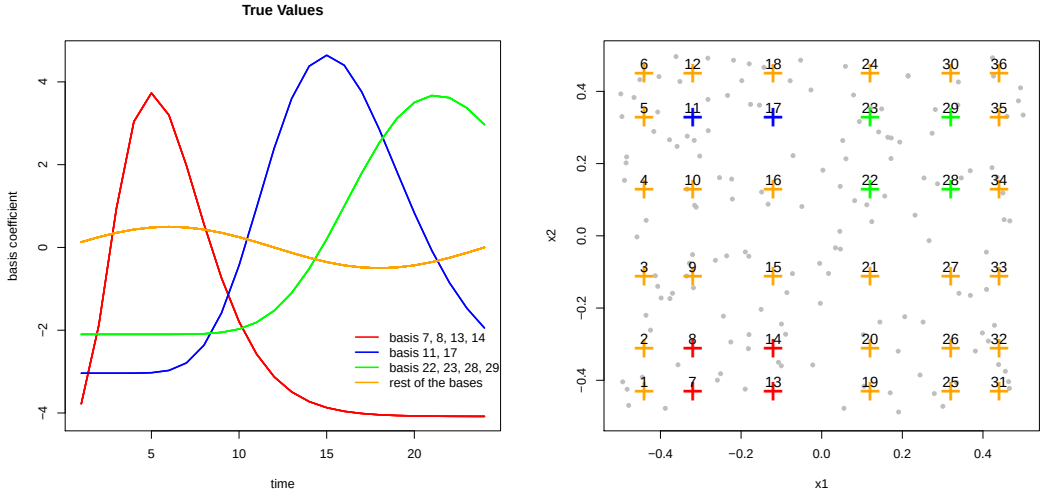


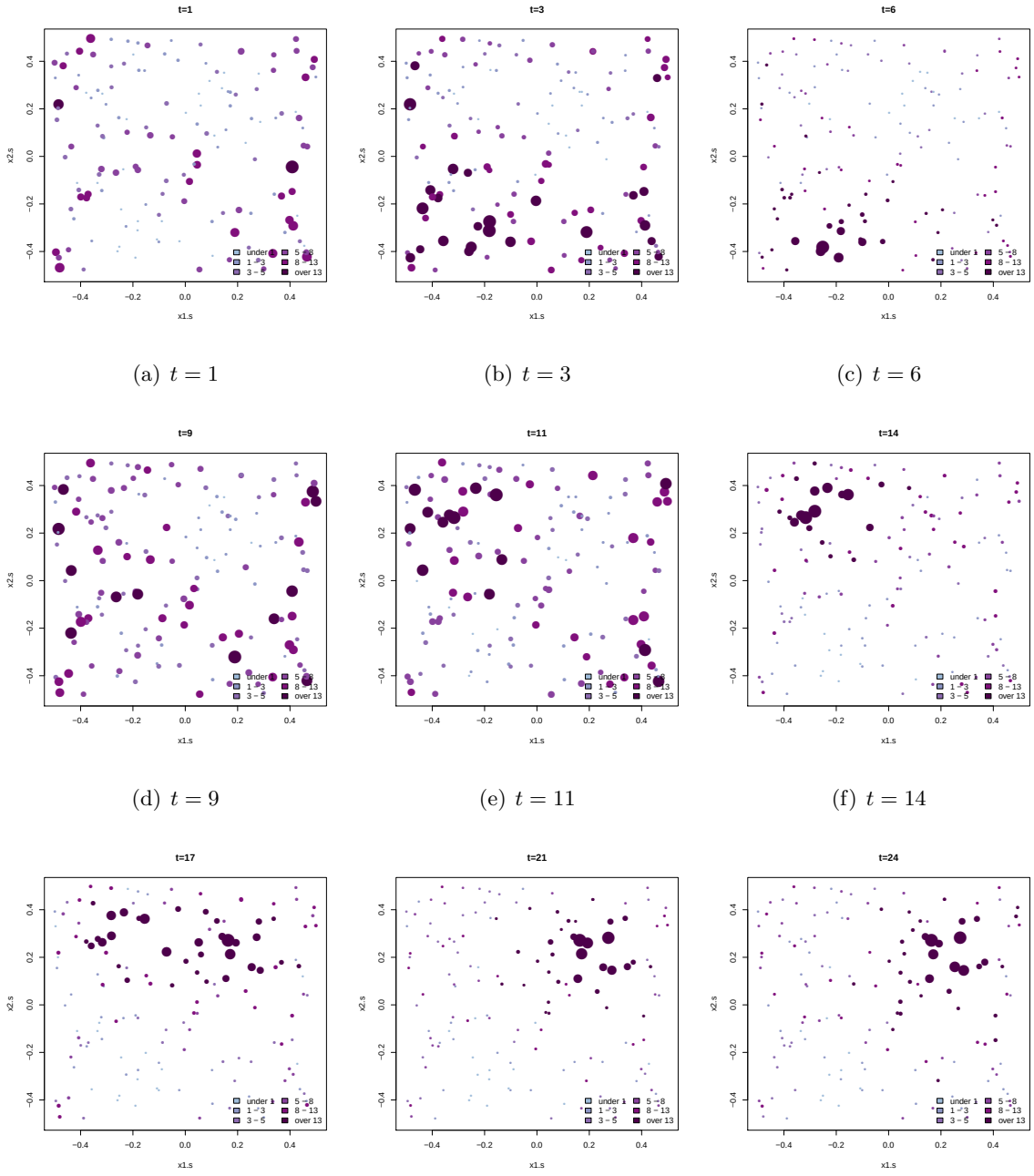
Figure 4.10: Simulated coefficients for the tensor product basis functions using cubic B-splines. True values of the basis functions b_{kt} are shown in the left plot, with their corresponding locations shown in the right plot. Grey dots are the locations of observed counts.

The simulation is set up to mimic the movements of epidemic centers in infectious diseases. Bubble plots of the simulated number of cases at selected time points are shown in Figure 4.11. From these plots we see that at time $t = 3$, local epidemic breakout starts in the bottom left region. As time goes by, the center moves to the top left corner around $t = 11$. Subsequently, the overall epidemic starts to die down, with the epidemic center moving to the top right corner at the end of the study period.

4.5.2 Analysis of the Simulated Data

For the analysis of the simulated data, we fit the interaction models with the four types of priors on the basis coefficients b_{kt} we proposed in Section 4.4. We use a Gamma(1, 0.005) distribution as the prior for the precision parameter τ_b. For each of the interaction models, we run the MCMC for $75,000$ iterations after an initial $25,000$ iterations of burn-in. Visual inspection shows good mixing of the MCMC chains. Example trace plots from the Type 4 interaction model are given in Figure 4.12.

Comparison of the parameter estimates using the four types of priors are shown in

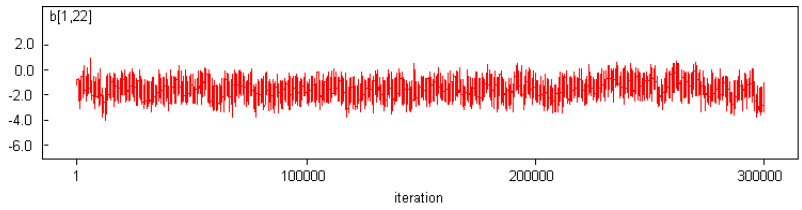


(a) $t = 1$ (b) $t = 3$ (c) $t = 6$

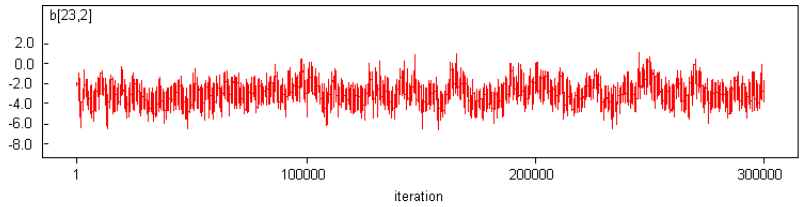
(d) $t = 9$ (e) $t = 11$ (f) $t = 14$

(g) $t = 17$ (h) $t = 21$ (i) $t = 24$

Figure 4.11: Bubble plots of the simulated number of cases at selected time points.



(a) $b_{1,22}$



(b) $b_{23,2}$

Figure 4.12: Selected trace plots of the basis coefficients b_{kt}.

Table 4.1. For inference, we use the posterior medians for the intercept α and the precision parameter τ_b. In all four models, the estimated overall level α is very close to the true value of -1.5. The estimated precision parameter τ_b are very different across the four models. However, the precision parameters from the four interaction models are not directly comparable because they have different meanings. For example, in the Type 3 interaction model, the inverse of τ_b^{-1} is the variability of the basis coefficients b_{kt}, conditioned on those at the neighbor basis $b_{k't}, k' \sim k$. By contrast, in the Type 2 interaction model, τ_b^{-1} is the variability of basis coefficients b_{kt}, conditioned on those at the neighboring time points $b_{kt'}, t' \sim t$.

The time series of the estimated basis coefficients are shown in Figure 4.10. We superimpose the true values and color code them according to their numbers. In the Type 1 interaction model, these lines are very jagged and are quite different from the truth. In the Type 2 model, the lines are much smoother over time due to the RW(2) prior we use. In the Type 3 model, the estimated basis coefficients $\boldsymbol{b}$ from the adjacent bases are pulled together due to the ICAR prior on the coefficients. In the Type 4 model, we see the estimated basis coefficients are a combination of those from the Type 2 and Type 3 models, which is as we

would expect.

Table 4.1: Comparison of the parameter estimates using four type of priors. For inference, we use the posterior median for intercept α and the precision parameter τ_b.

parameter	Type 1	Type 2	Type 3	Type 4
α	-1.52	-1.48	-1.46	-1.48
τ_b	0.36	7.67	0.20	5.30

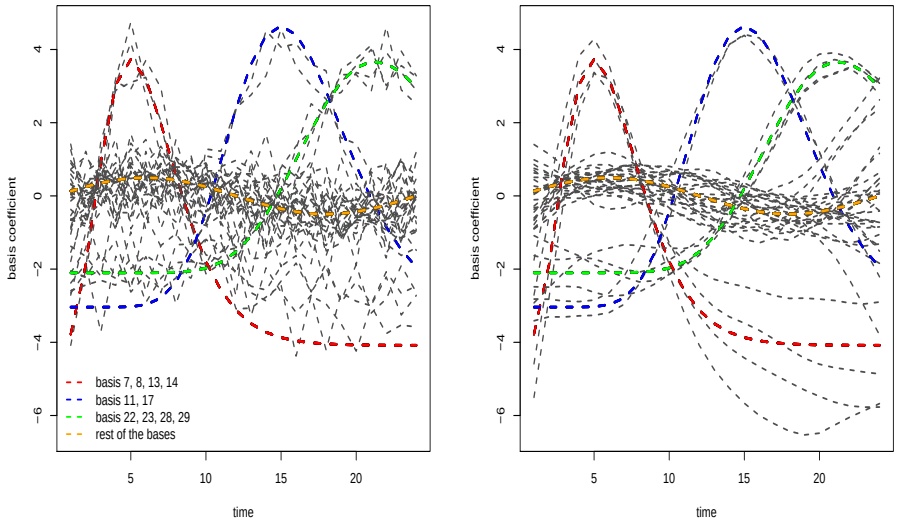
We are also interested in the estimated log relative risk at each of the observed locations. We select four areas with very small expected number ($E = 5$) and four with large expected number ($E = 48$). In Figure 4.14 (a), we give the true log relative risk for these four areas. In the following panel (b), we plot the log SMR for the same areas. It is clear that areas with smaller expected numbers tend to have larger variability in the estimated log SMRs. In panels (c) to (f) we give the results from the four types of interaction models. From these plots we see that estimated log relative risks using our models are clearly more accurate than the raw log SMRs. Between the four types of interactions, the estimated log relative risks from the Type 2 and Type 4 interaction models are the closest to the truth.

We also compare the Mean Squared Errors (MSE), defined as:

$$\text{MSE}_y = \tfrac{1}{KT}\textstyle\sum_k \sum_t (\widehat{Y}_{kt} - Y_{kt})^2 \tag{4.9}$$

$$\text{MSE}_b = \tfrac{1}{KT}\textstyle\sum_k \sum_t (\widehat{b}_{kt} - b_{kt})^2, \tag{4.10}$$

where Y_{kt} and b_{kt} are the true values used in the simulation, and $\hat{Y}_{kt}$ and $\hat{b}_{kt}$ are the fitted values. The results are shown in Table 4.2. As a comparison, we also include results using a generalized linear model (GLM) with no penalty on the regression coefficients. We fit T different models to the data, one for each time point. Clearly the MSE of the basis coefficients $\boldsymbol{b}$ is much larger using the GLM than for the interaction models. This demonstrates that effect of the priors we impose as the penalty on the basis coefficients. When comparing the MSE of the observed counts, we notice that the GLM has the smallest value. At first glance this may indicate GLM a better model. However, the GLM model is



(a) Type 1

(b) Type 2

(c) Type 3

(d) Type 4

Figure 4.13: Estimated basis coefficients with four different priors: colored lines are the true values used in the simulation.

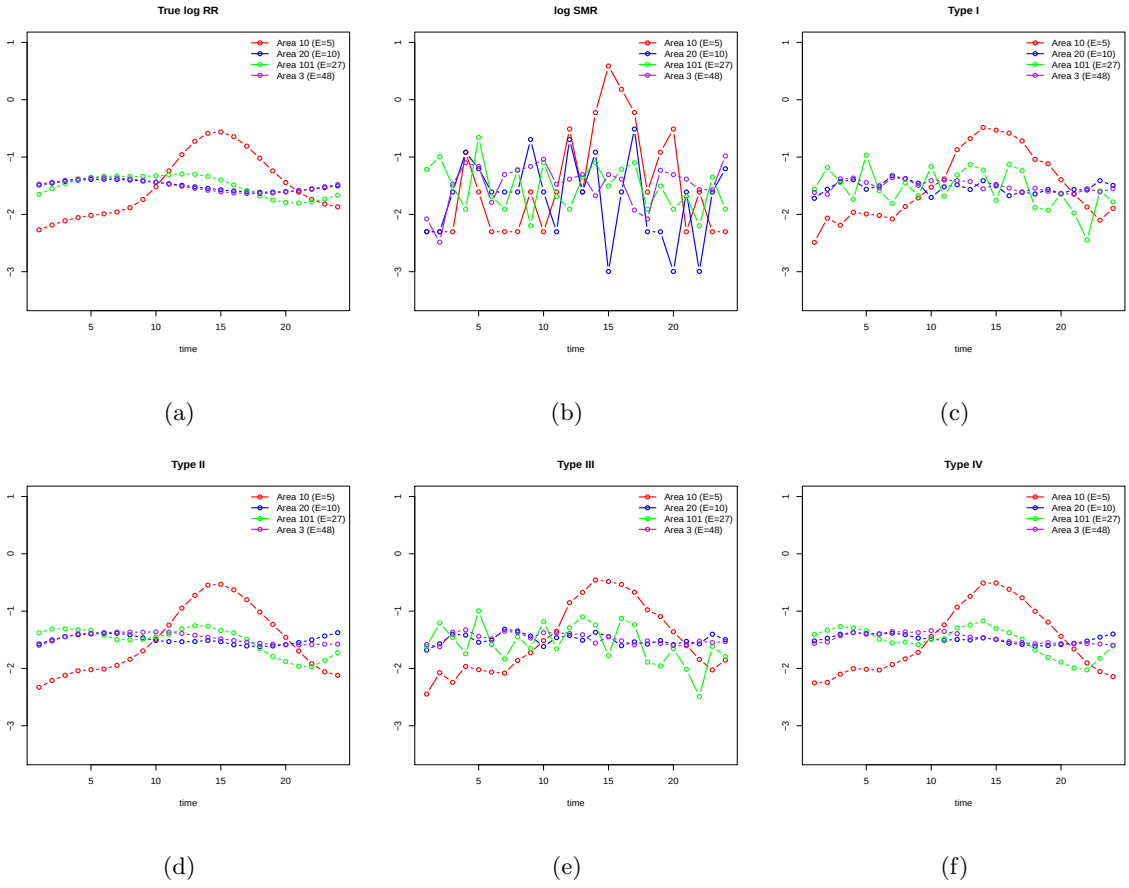


Figure 4.14: Comparison of the log relative risk of selected areas.

a fully saturated model and this is why it gives the smallest MSE for the observed counts y. An overfitted model usually performs poorly when used for prediction.

Table 4.2: MSE comparison for the simulated data.

b	GLM	Type 1	Type 2	Type 3	Type 4
MSE	0.74	0.11	0.094	0.084	0.092
Y	GLM	Type 1	Type 2	Type 3	Type 4
MSE	7.16	7.65	8.9	8.13	9.09

4.6 Prediction

One important objective of the analysis of HFMD surveillance data in China is to predict disease risk and disease counts over time periods (e.g. weeks or months) into the future, either for a single geographical area, or several geographical areas combined. In this section, we compare the performance of prediction using the GLM and the interaction models we developed in Section 4.4.

A measure that will be used to compare different models is the mean squared error of prediction for time point t based on partial data g:

$$\text{MSEP}_t^{(g)} = E[(Y_t^{(g)} - y_t)^2] = \text{var}[\widehat{Y}_t^{(g)}] + \text{bias}(\widehat{Y}_t^{(g)})^2 \tag{4.11}$$

where $Y_t^{(g)}$ is the prediction and y_t is the truth. Hence, we are looking for a prediction which has both low variance and low bias. These two aims are usually in conflict since complex/simple models will have low/high bias but high/low variance. The $\text{MSEP}_t^{(g)}$ is estimated by

$$\widehat{\text{MSEP}}_t^{(g)} = \text{var}[Y_t^{(g)}] + (E[Y_t^{(g)}] - y_t)^2. \tag{4.12}$$

The above measures may also be extended to average across areas via

$$\text{MSEP}_t^{(g)} = \frac{1}{I}\sum_{i=1}^{I} E[(Y_{it}^{(g)} - y_{it})^2] = \frac{1}{I}\sum_{i=1}^{I}\left(\text{var}[\widehat{Y}_{it}^{(g)}] + \text{bias}(\widehat{Y}_{it}^{(g)})^2\right) \tag{4.13}$$

and

$$\widehat{\text{MSEP}}_t^{(g)} = \frac{1}{I}\sum_{i=1}^{I}\left(\text{var}[Y_{it}^{(g)}] + (E[Y_{it}^{(g)}] - y_{it})^2\right). \tag{4.14}$$

For the simulated data, we take the first 21 weeks as the observation and predict the counts Y_i at t from weeks 22 to 24 in all areas using the GLM and the interaction models we proposed. We use (4.14) to compare the prediction performance using the observed y_{it} as the truth. The results are summarized in Table 4.3. The prediction is greatly improved when we include temporal components in the Type 2 and Type 4 interaction models. The full interaction model gives the smallest MSEP and is hence considered the best prediction model. The comparison of the predicted counts with the observed counts is shown in Figure 4.15.

Table 4.3: MSPE for the simulated data with $g = 21$.

prediction	Type 1	Type 2	Type 3	Type 4
MSPE	841.4	240.0	857.6	195.9

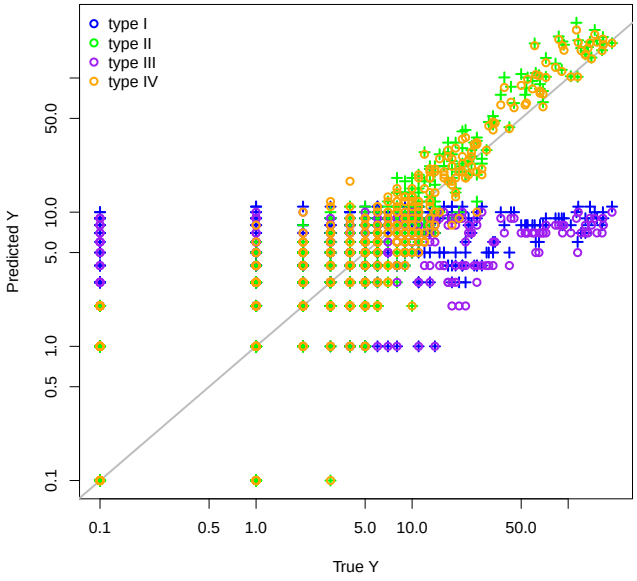


Figure 4.15: Comparison of the prediction at t from 22 to 24 using four interaction models.

4.7 Application to HFMD Data in the Central North of China, 2009-2010

The study area we choose for analysis is located in the central north region of China and is shown in Figure 4.16. The area consists of five provinces and direct-controlled municipalities, of Beijing, Tianjin, Hebei, Henan, Shandong and Shanxi. The total population in the region is estimated to be $318,022,505$ in 2009. Within the study region, the weekly number of HFMD counts are collected in 59 prefecture (shown in the left panel of Figure 4.17) between 2009 and 2010, with a total of $418,949$ reported HFMD cases in 2009 and $478,238$ in 2010. Based on the population composition in each prefecture, we calculate the expected number

of counts which is assumed to be constant across the study time period. The expected number of cases range from 18 to 364, with the map of expected number shown in the right panel of Figure 4.17.

In Figures 4.19 and 4.20 we show maps of the log SMR at selected weeks, with a common scale for easy comparison. The overall temporal trend is quite clear – in both years, the epidemic kicks off around March, reaches its peak in June and July, and dies off towards the end of the year. As for the spatial movement of the epidemic center interpretation is much more difficult. There is little clear pattern except to say that there is a general south-to-north movement. The interpretation of the log SMR maps is difficult partially due to the limited information using aggregated data, and partially due to the sampling variability. In addition, the area-level summary lacks the continuity that is desired for visual maps. Now we proceed to the analysis with the proposed penalized spline model.

Let Y_{it} be the observed counts in prefecture i ($i = 1, \ldots, 59$) and week t ($t = 1, \ldots, 104$), we assume the number of reported HFMD cases follow a Poisson distribution with expected number of cases E_i. On the log scale, we model the relative risk μ_{it} as

$$\log(\mu_{it}) = \alpha + \beta_z z_i + \gamma_t + \phi_t + v_i + \sum_{k=1}^{K} b_{kt} B_{ik}(\boldsymbol{x}_i),$$

where $\boldsymbol{x}_i = (x_{1i}, x_{2i})$ is the centroid location of prefecture i, after rescaling the study region to a $[0,1] \times [0,1]$ square with the north/south ratio kept unchanged. We include the calculated population density per square kilometers z_i, which is normalized such that the mean is 0 and the standard deviation is 1. We select 4 inner knots for each coordinate to create the marginal cubic B-splines, and then construct the tensor product basis function B with b_{kt} as the coefficient for the k-th basis at time t. Due to the lack of data at some areas, we have to manually remove some of the basis functions which leaves $K = 22$ bases in the final analysis. The marginal cubic B-spline functions and the location of the two-dimentional basis functions are shown in Figure 4.18.

For the basis coefficients b_{kt}, we include the four types of interaction priors introduced earlier. For the temporal trend we use a RW(2) prior γ_t and an independent and identically distributed normal prior for ϕ_t. We also include v_i to capture the extra Poisson variability in the data. Initially we included a spatial component with ICAR prior in the model but we

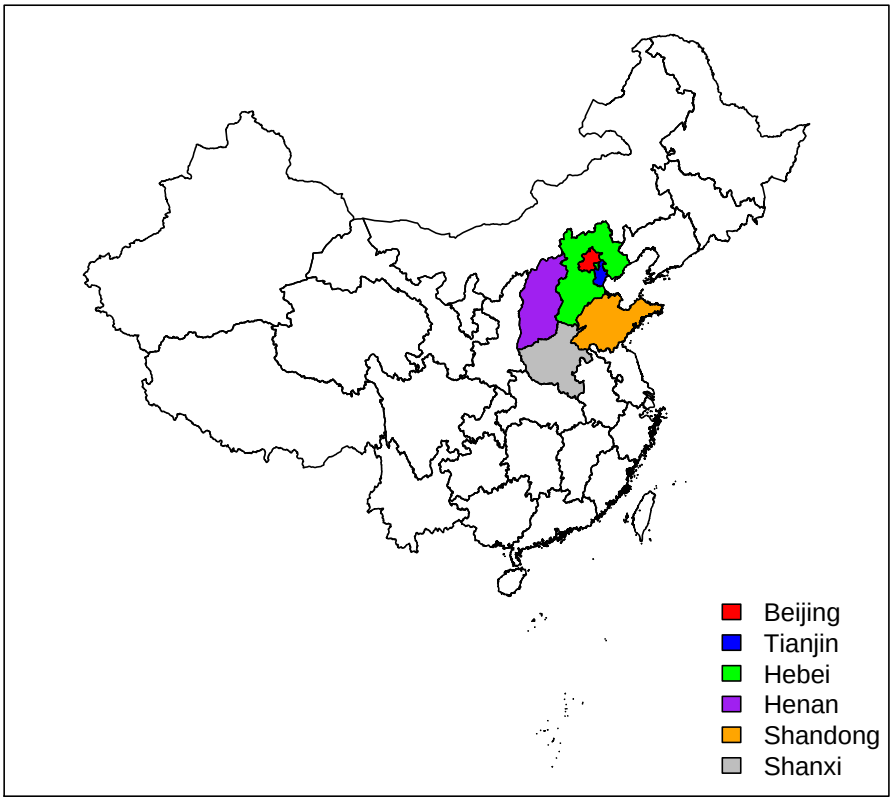


Figure 4.16: Location of Central North region in China.

found it was not needed, i.e. there was very little marginal spatial variability. The model is implemented using MCMC, with the parameter estimates summarized in Table 4.4. In all four models, we find the population density has a significant positive relationship with the log relative risks. Estimates of the structured and unstructured temporal components γ_t and ϕ_t are very similar in the four models, and therefore we only show the results from the full interaction model (i.e. the Type 4 model) in Figure 4.21. The overall temporal pattern

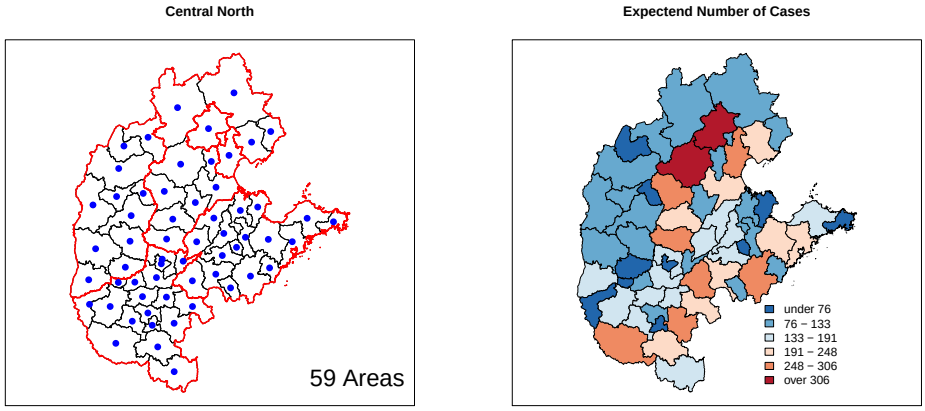
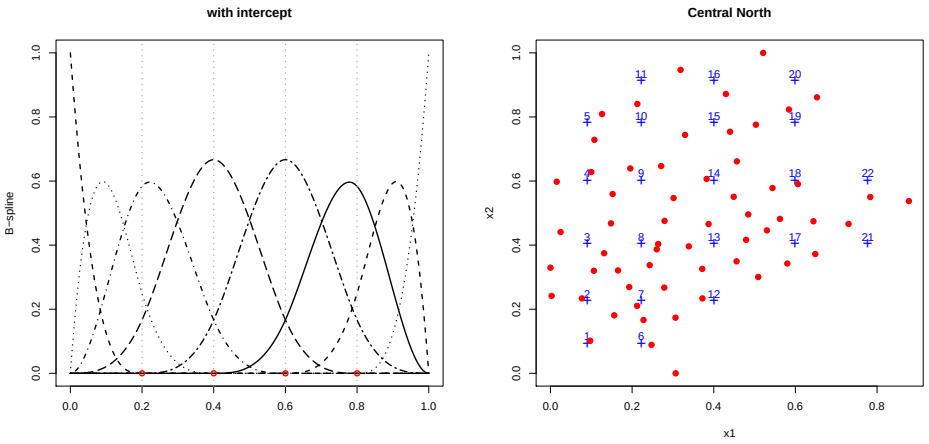


Figure 4.17: Map of the study region with centroids of the prefecture in blue dots (plot on the left), and the weekly expected number of cases (plot on the right).



(a) Marginal cubic B-spline basis

(b) Final tensor product basis used in analysis

Figure 4.18: Location of the bases.

that is common to all areas is captured in the structured temporal random effect $\boldsymbol{\gamma}$. In both 2009 and 2010, the epidemic starts around April, with a higher peak in 2010. The epidemic almost dies out during the winder time. The unstructured temporal random effect $\boldsymbol{\phi}$ reflects

Table 4.4: Parameter estimates from four types of interaction models, using weekly CN HFMD data 2009–2010.

Parameter	Type I	Type II	Type III	Type IV
α	-1.03 (0.079)	-0.96 (0.076)	-1.0(0.074)	-1.0 (0.066)
β_z	0.27 (0.043)	0.25 (0.021)	0.36 (0.032)	0.36 (0.031)
σ_γ	0.15	0.17	0.16	0.17
σ_ϕ	0.19	0.056	0.053	0.042
σ_v	0.55	0.50	0.53	0.52
σ_b	4.01	0.54	12.29	1.15

the residual temporal variability after accounting for the main RW(2) temporal effect, and it does not show any clear pattern as expected has has a very narrow range between -0.005 and 0.005. The estimated unstructured spatial random effect $\boldsymbol{v}$ ranges from -1.1 to 1.6, and its map is shown in Figure 4.22. No clusters of high or low risks areas are seen in the map.

In Figure 4.23, we plot the estimated basis coefficients from all four models, with each dotted line representing the estimated time series for one basis coefficient. We use the same scale for easy comparison. The estimated basis coefficients $\boldsymbol{b}$ are very similar for the Type 1 and Type 3 interaction models, which implies that there is not much structured spatial variability in the data. Recall that we initially fit the model with a structured spatial random effect but the random effects were estimated as being very small. The results here agree with this finding. Time series of the estimated $\boldsymbol{b}$ from the Type 2 and Type 4 models are very similar, but both are much smoother than those from the Type 1 and Type 3 models.

Finally in Figure 4.24 and Figure 4.25, we present maps of the term

$$\gamma_t + \sum_{k=1}^{K} b_{kt} B_{ik}(\boldsymbol{x}_i).$$

These maps give the broad-scale space-time dynamics of the HFMD. Compared to the raw log SMR maps, the movement of the epidemic centers is much clearer as the regions with

elevated relative risks are easier to identify. In addition, an animation with these maps in the time order shows two local epidemics that start in the north-west and south-west areas around week 22, and gradually move towards each other. This movement is not seen in the log SMR maps and deserves further investigation.

We also preform some prediction analyses on the China HFMD data. We take the first 101 weeks of total HFMD data in the Central North region between 2009 and 2010, and predict total HFMD counts for the next 3 weeks. The performance of the prediction is still measured with (4.14). The baseline model we use for comparison contains only the main effects

$$\log(\mu_{it}) = \alpha + \beta_z z_i + \gamma_t + \phi_t + v_i.$$

The MSPE is estimated to be 1018.7 in the baseline model. We then apply the four types of interaction models to the data and the MSPE results is summarized in Table 4.5. The Type 2 interaction model gives the smallest MSPE and is considered to be the best prediction model. The MSPE using the Type 3 and Type 4 interaction models are smaller than that using the main effect model, but is larger than those using the Type 1 and Type 2 interaction models. One possible explanation is that the bias brought by the additional spatial structure imposed on the basis coefficients outweighs the variance reduction.

Table 4.5: MSPE for the Central North HFMD data in China, with $g = 101$.

prediction	Main effect	Type 1	Type 2	Type 3	Type 4
MSPE	1018.7	1011.4	708.2	1013.7	1014.8

4.8 Discussion

In this chapter we have developed penalized spline models for spatial-temporal infectious disease data. The structure of the spatial, temporal and spatial-temporal interaction can have four possible forms and these structures are imposed through the prior distributions. We conduct a simulation study where the simulated Poisson observations exhibit temporal

and spatial trends. The analysis results show that our proposed models can accurately capture these trends. We then use our proposed models to analyze HFMD counts in Central North China between 2009 and 2010 and showed that can reveal the temporal and spatial trends in the data. In addition, we are able to provide maps that can help indicate the spatial movement of the epidemic centers. There is still plenty of work that needs to be done for these models and we now discuss different aspects.

The first aspect can be viewed as a general question for non-/semi- parametric models. Although they offer great flexibility. they also add substantial problems to model selection and inference, for example, see Gu and Wahba (1991), Hastie and Tibshirani (1990) or Wood (2000). We have the same problems for the models we develop. How do we decide the best model to use? In some settings we can use cross-validation, but this method might be difficult for Bayesian penalized splines due to computational issues. Other methods for model selection include the Bayesian Information Criterion (BIC), see for example, Knorr-Held (2000). Another issue with the Bayesian penalized spline model is the choice of prior distribution. In Fong et al. (2010), the authors suggested specifying the variance component priors by equating with the effective degrees of freedom of the spline models. This approach is worth investigating for specifying the priors for the precision parameter τ_b. A different approach is based on specifying a range for the more interpretable marginal distribution of the random effect and use the marginal to drive specification of prior parameters, see Wakefield (2007). A crude idea for borrowing this idea to specify the precision parameter in our proposed models is as follows: we may believe that the marginal distribution of the residual relative risks has 95% range $[0.1, 10]$. Translating into our model, this means that we believe that on the log scale, the 95% range of $\sum_k Z_{ik} b_{kt}$ is $[\log(0.1), \log(10)]$. Because we know the range of the basis functions Z_{ik}, we can backsolve to find the range for b_{kt} and then specify the prior distribution based on this range. This idea will be further pursued in our future study.

From a practical point of view, one of the problem of implementing the model is the heavy computational burden using MCMC. The Bayesian penalized spline models we develop are fairly complicated with many latent random effects and we need very long runs of MCMC to achieve convergence. As an alternative, we can use the INLA approach introduced in Section

2.4, which can provide results in seconds for moderate sizes of data (such as the simulated data we used in this chapter). An illustration of the MCMC and INLA comparison, we take the simulated data set and run all four types of models using both methods. The estimated basis coefficients b_{kt} is very similar in all cases (see the plots in Figure 4.26). However, in terms of computation time, MCMC runs take hours or days while INLA takes only 69 seconds! *R* code to implement the INLA approach using the simulated data can be found in Appendix A.

A natural extension of the current work is to provide prediction of the next epidemics based on surveillance data from previous years. This endeavor is of particular interest to health authorities. The prediction can be carried out at difference spatial scales, for example, at the prefecture, regional or national level. Supplemented with limited data from the current year, we can investigate how early in the current year it is possible to reliably predict the magnitude of the epidemic at each of the three levels of geographic aggregation. Clearly, the more historical data we have, the more likely we are to come up with a better predictive model (in terms of narrower predictive intervals). On the other hand, with more data we may wish to specify a more parametric model that can provide more precise predictions than the spline models we propose. We have touched on the prediction topic in this chapter, however, we need more thorough analysis to investigate the prediction as a function of the different epidemic stages. Such analysis requires performing a series of prediction analysis based on different partial data and is computationally intensive. We hope that with the aid of fast computation such as INLA, this analysis can be carried out relatively efficiently.

Finally, there are still more aspects of the China HFMD data to investigate. We only implemented the model with data from a subregion in China and we need to analyze the data at the national level. Also, we should take into account meteorological and climate data, and also school openings. In China, the continuous population movement from the countryside to the cities may contribute greatly to infectious disease transmission. Therefore, including some measurement of transportation between different areas may help us explain the breakouts of the diseases at particular locations and therefore gain insights into effective public health interventions.

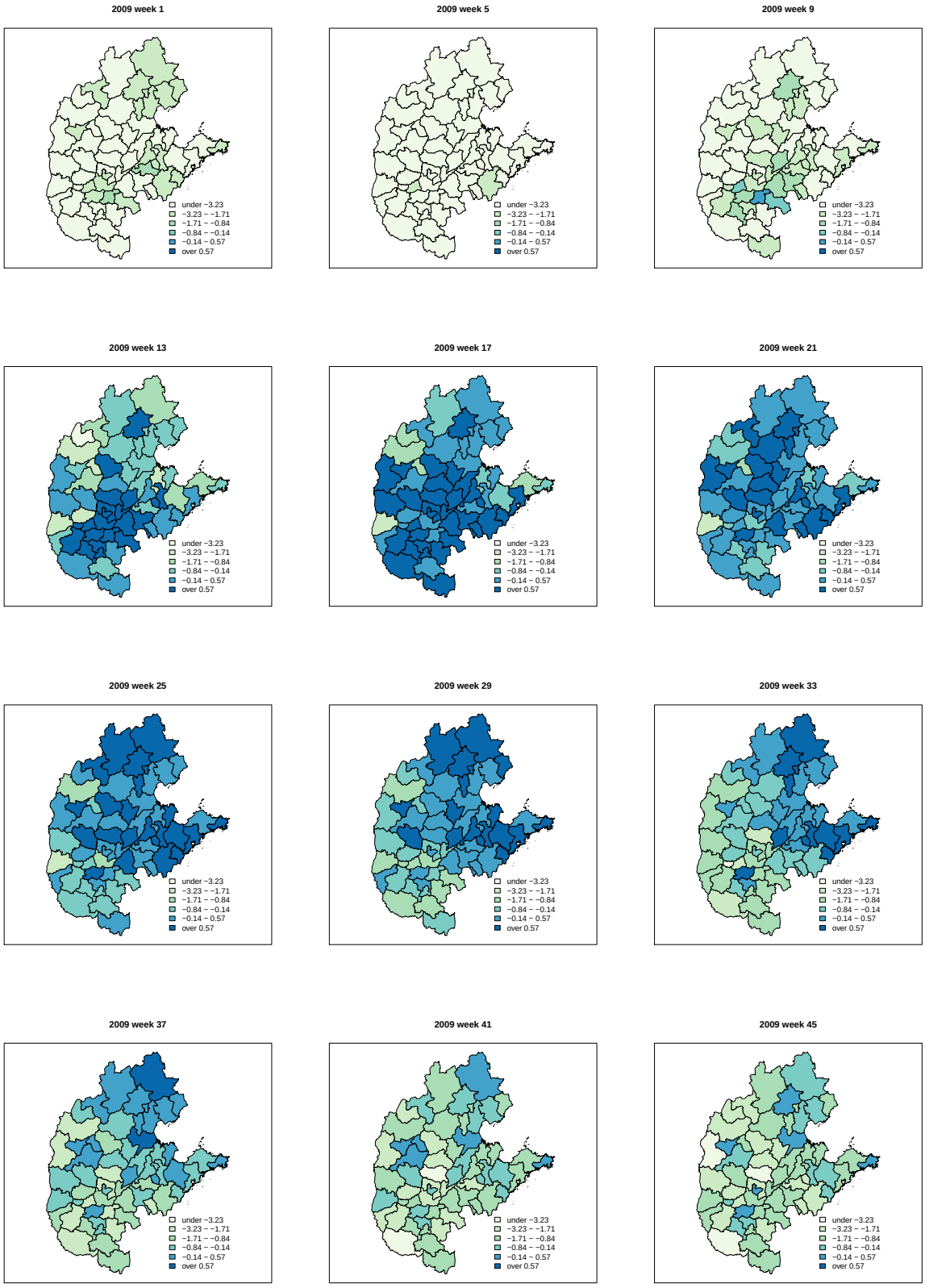


Figure 4.19: Selected weekly log SMR of Central North prefectures in 2009.

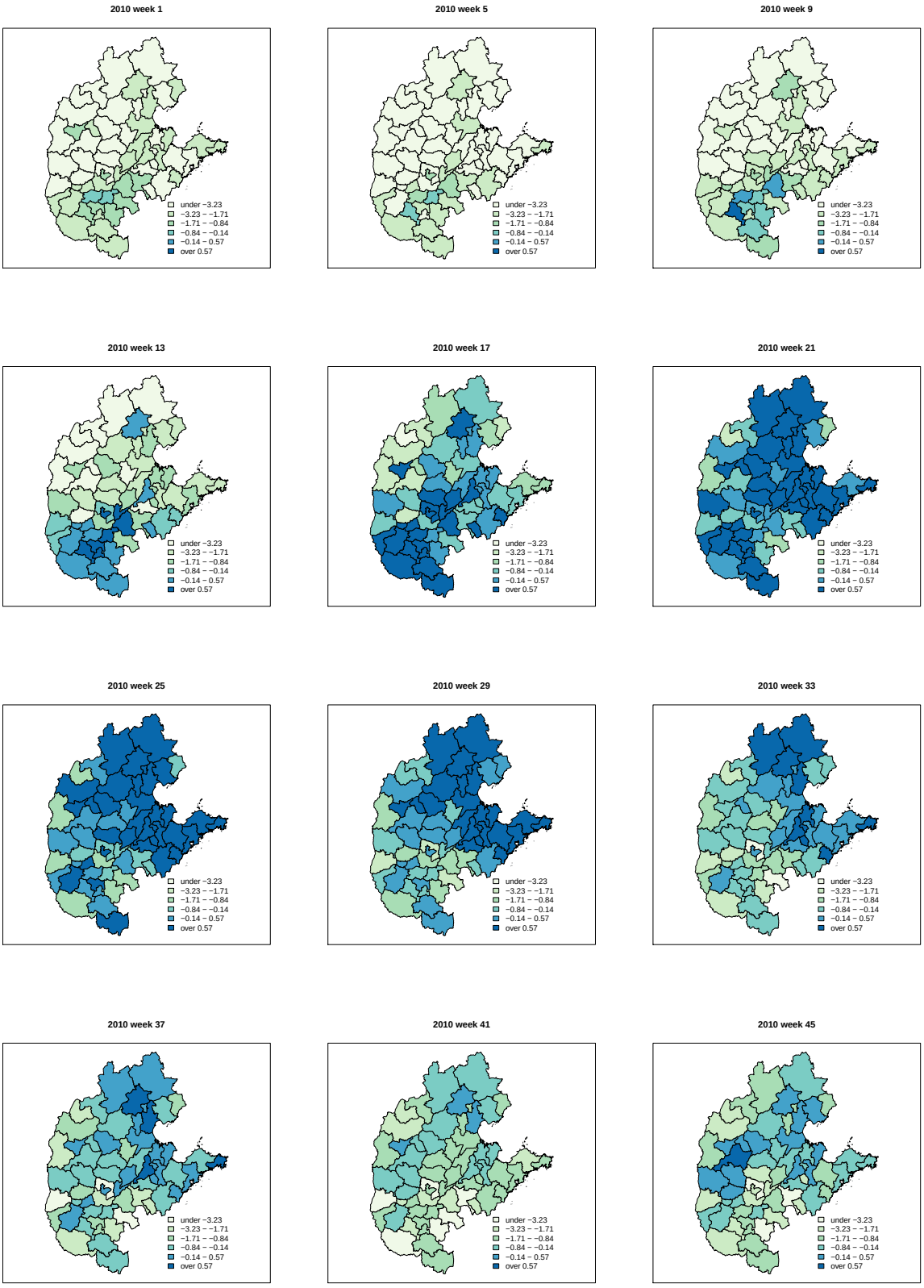
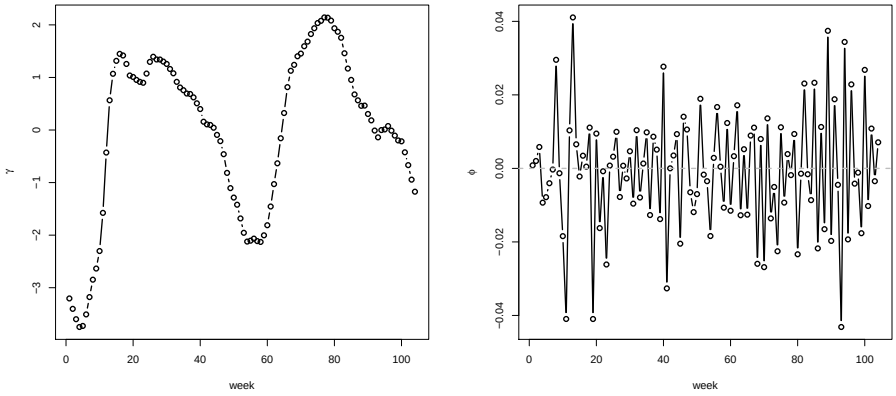


Figure 4.20: Selected weekly log SMR of Central North prefectures in 2010.



(a) Structured time $\boldsymbol{\gamma}$ (b) Unstructured time $\boldsymbol{\phi}$

Figure 4.21: Estimated temporal component $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ from the Type 4 interaction model with the China Central North HFMD data.

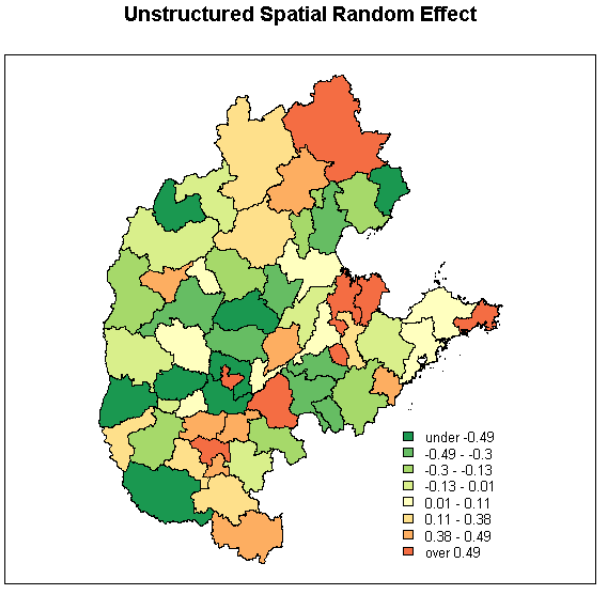
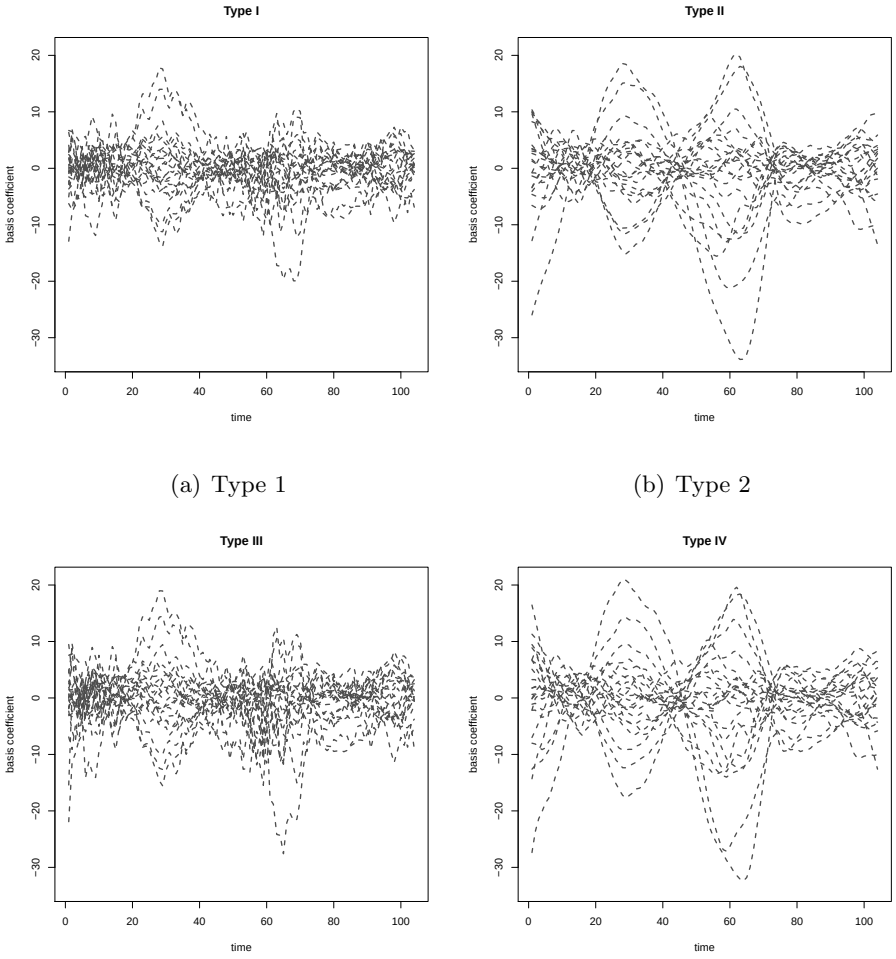


Figure 4.22: Estimated unstructured spatial component $\boldsymbol{v}$ from the Type 4 interaction model with the China Central North HFMD data.



(a) Type 1

(b) Type 2

(c) Type 3

(d) Type 4

Figure 4.23: Estimated basis coefficients with four different priors with the China Central North HFMD data.

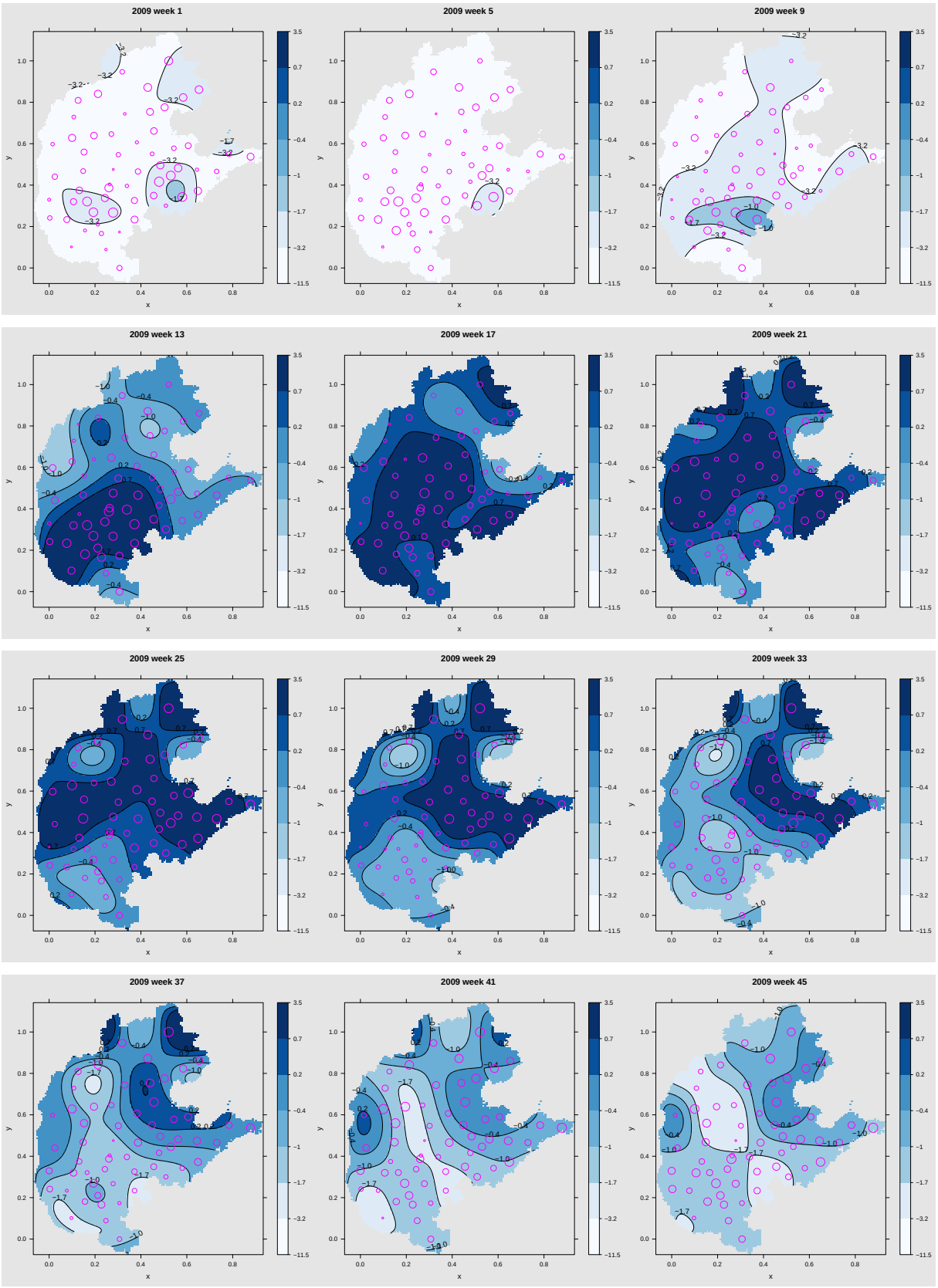


Figure 4.24: Estimated broad-scale space-time dynamics of 2009 Central North HFMD in China, at selected weeks.

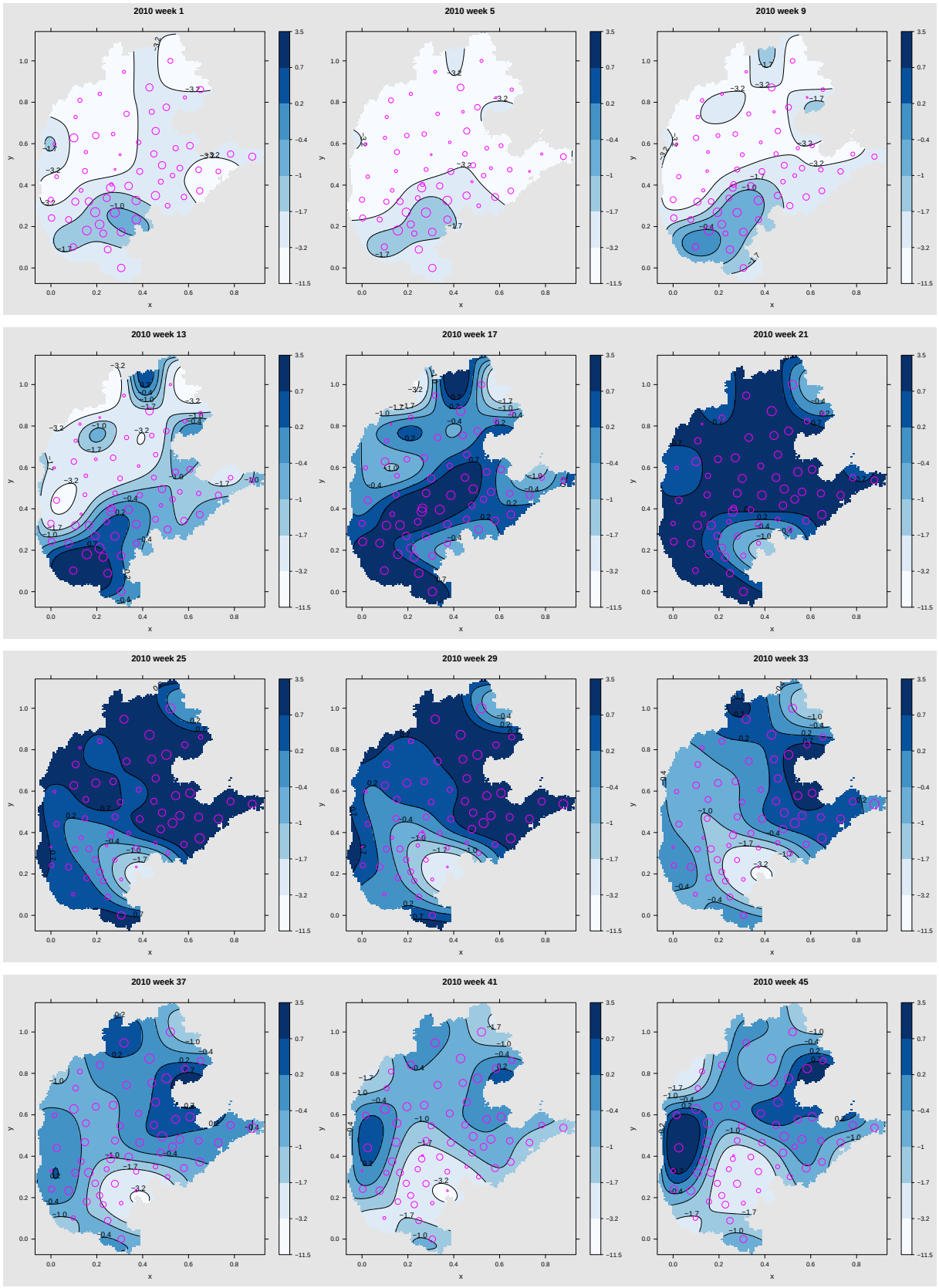


Figure 4.25: Estimated broad-scale space-time dynamics of 2009 Central North HFMD in China, at selected weeks.

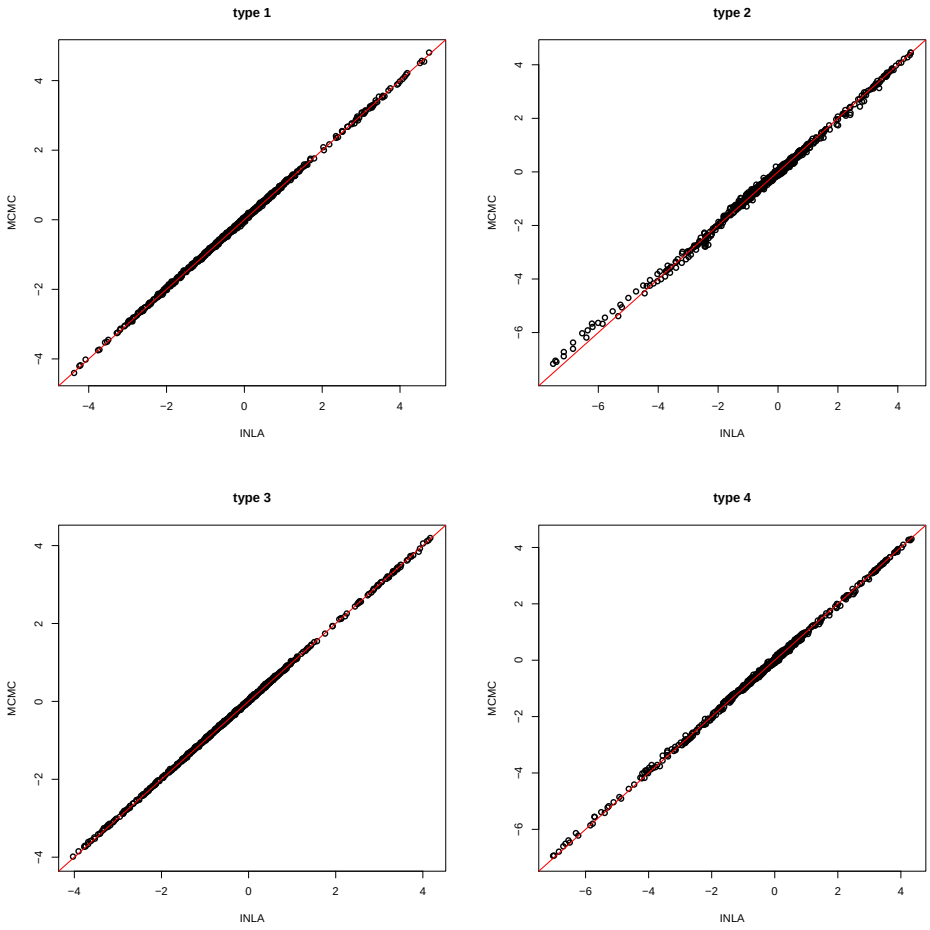


Figure 4.26: Comparison of the fitted basis coefficients b_{kt} between MCMC and INLA.

Chapter 5

SPACE-TIME MODELS FOR AGGREGATED INFECTIOUS DISEASE DATA WITH DIFFERENT STRAINS

5.1 Motivating Data

In 2009, a total of 101,092 Hand-Foot-Mouth disease (HFMD) cases were reported in Henan province. Among these cases, $36,370$ were females (36%) and $64,722$ were males. The number of clinically diagnosed severe cases was $3,921$. In addition, 45.7% of the severe cases and 4.7% of the mild cases received lab tests that recorded the virus strain information. Table 5.1 summarizes the results of these tests.

A total of 6,318 cases were lab-tested, of which 57.2% were reported to have the EV71 strain, 13% had the CoxA16 strain and the rest had other Enteroviruses. Among the clinically diagnosed severe cases, 83% were reported to have the EV71 strain and 0.6% had the CoxA16 strain. Among clinically diagnosed mild cases, 46.8% were reported to have the EV71 strain and 17.9% had the CoxA16 strain.

In Figure 5.1 we present maps of the Henan 2009 HFMD data at the prefecture level. The four maps show the numbers of total cases, severe cases, EV71 cases and CoxA16 cases. One interesting aspect of these maps is the resemblance between maps of the severe cases and the EV71 cases. This may indicate that even though both EV71 and CoxA16 strains are associated with HFMD, the percentage of severe cases is higher among patients with the EV71 strain than for those with the other enteroviruses. In Figure 5.2, we show time series plots of the weekly number of Henan 2009 HFMD total cases, severe cases, EV71 cases, and CoxA16 cases, all disaggregated by gender. Again, we see the similarity of the temporal trend between the severe cases and EV71 cases, but less similarity between the severe cases and the CoxA16 cases. Interpretation is not straightforward, however, because the EV71 and Cox A16 cases are based on the number of lab-tested samples which differ over time (see Figure 5.4).

An analysis by strain type can help us understand more about the etiology of HFMD and the relationship between the strain types and the severity of the disease. With the surveillance data and the lab test results of selected patients, the scientific questions in which we are interested include:

1. What are the strain-specific probabilities of contracting HFMD?

2. What are the strain-specific probabilities of being a clinically severe case?

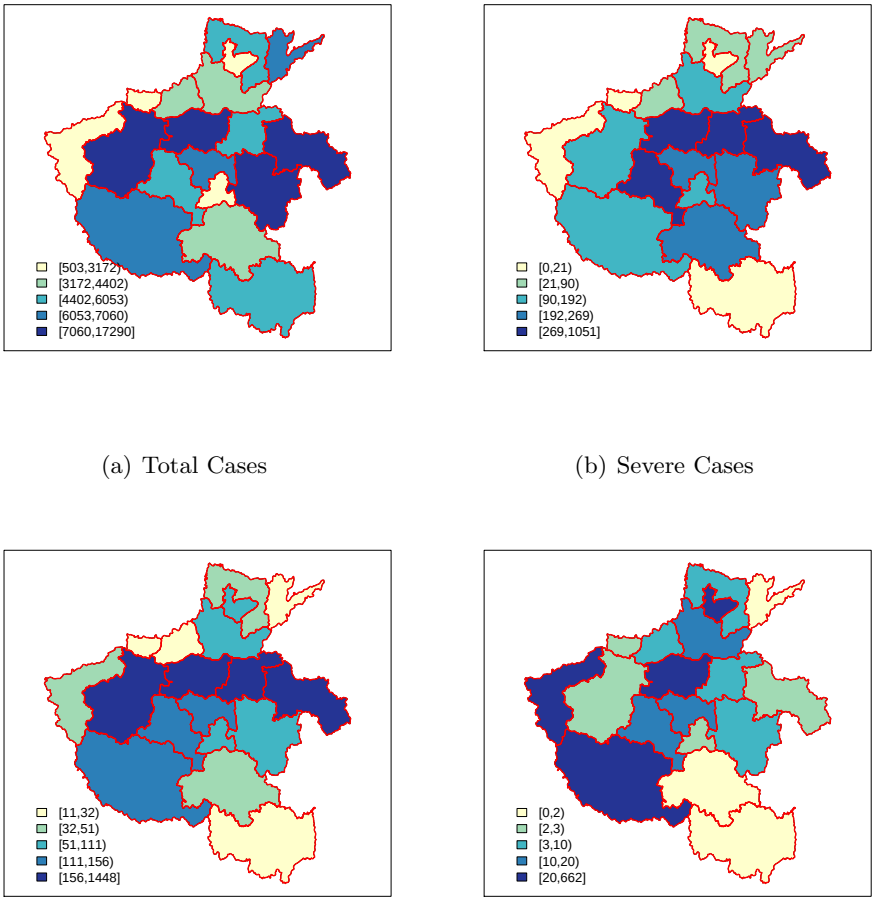
3. Is there any interaction between the strain-specific probabilities?

4. How do these probabilities evolve over time and space?

Table 5.1: Summary statistics of Henan 2009 HFMD data.

		Severe	Mild	Total
Lab-tested	EV71	1,496	2,116	3,612
	CoxA16	11	808	819
	Other	286	1,601	1,887
Not tested		2,128	92,646	94,774
Total		3,921	97,171	101,092

The rest of this chapter is organized as follows: in Section 5.2 we describe the naive estimates of the probabilities of interest and their shortcomings. In Section 5.3 we describe our proposed model for estimating these probabilities. We also provide the algorithms for MCMC computation required for making inference using our proposed model. We then turn our attention to an initial examination of the spatial and temporal patterns in the probabilities of interest, with the temporal pattern being investigated and summarized in Section 5.4 and the spatial pattern in Section 5.5. Finally, we describe future work in Section 5.6.



(a) Total Cases

(b) Severe Cases

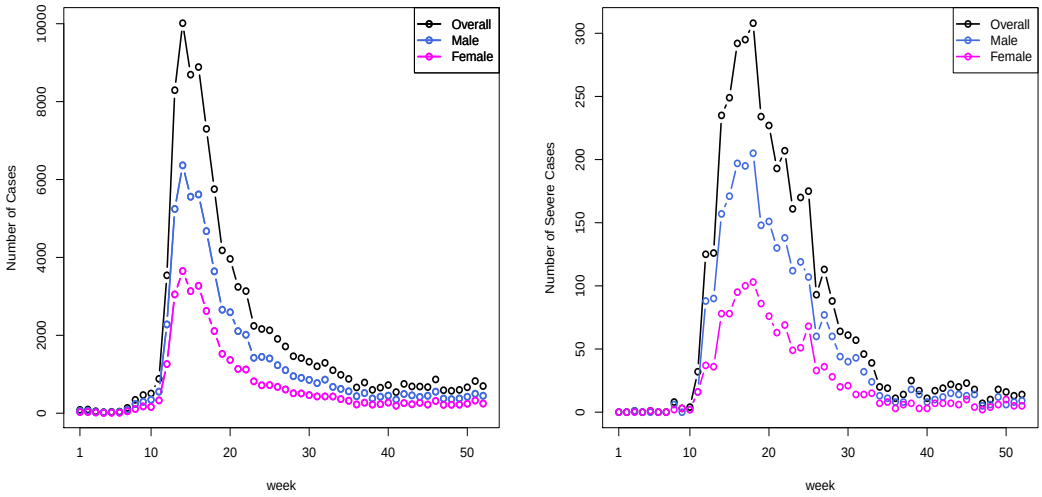
(c) EV71 Cases

(d) CoxA16 Cases

Figure 5.1: Maps of Henan 2009 HFMD data at the prefecture level.

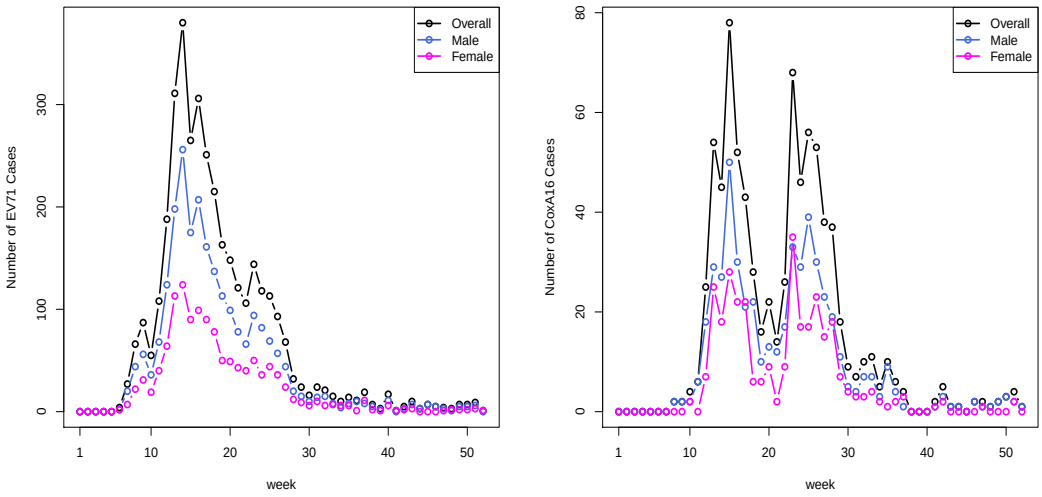
5.2 Naive Estimation

For simplicity, we assume there are only two co-circulating strain categories (1 and 2, for example EV71 and others) and two disease severity categories (mild and severe). For a generic area and time period, let Y_+ be the total number of HFMD cases, and Y_+^s and Y_+^m be the total clinically diagnosed severe and mild cases. These three quantities can be obtained from the surveillance system. We let k^s and k^m indicate the sample sizes



(a) HFMD Cases

(b) Severe Cases



(c) EV71 Cases

(d) CoxA16 Cases

Figure 5.2: Time series of the Henan 2009 HFMD data by week.

for clinically diagnosed severe and mild cases. Let Z_1^s and Z_2^s denote the lab-confirmed number of *severe* cases from strain 1 and strain 2 respectively. Similarly, we let Z_1^m and Z_2^m denote the lab-confirmed number of *mild* cases from strain 1 and strain 2 respectively. The variables k^s, k^m, Z_1^s, Z_2^s, Z_1^m and Z_2^m are obtained from the lab test results. The relationship

between the surveillance and the lab test data can be summarized in Figure 5.3, with the blue boxes showing the surveillance data and the salmon boxes showing the lab test data. The following deterministic relationships are obvious from the diagram: $Y_+ = Y_+^s + Y_+^m$, $k^s = Z_1^s + Z_2^s$ and $k^m = Z_1^m + Z_2^m$. The parameters of interest here are: $q = Pr(\text{strain 1}|\text{case})$, $p_1^s = Pr(\text{severe}|\text{strain 1})$, and $p_2^s = Pr(\text{severe}|\text{strain 2})$.

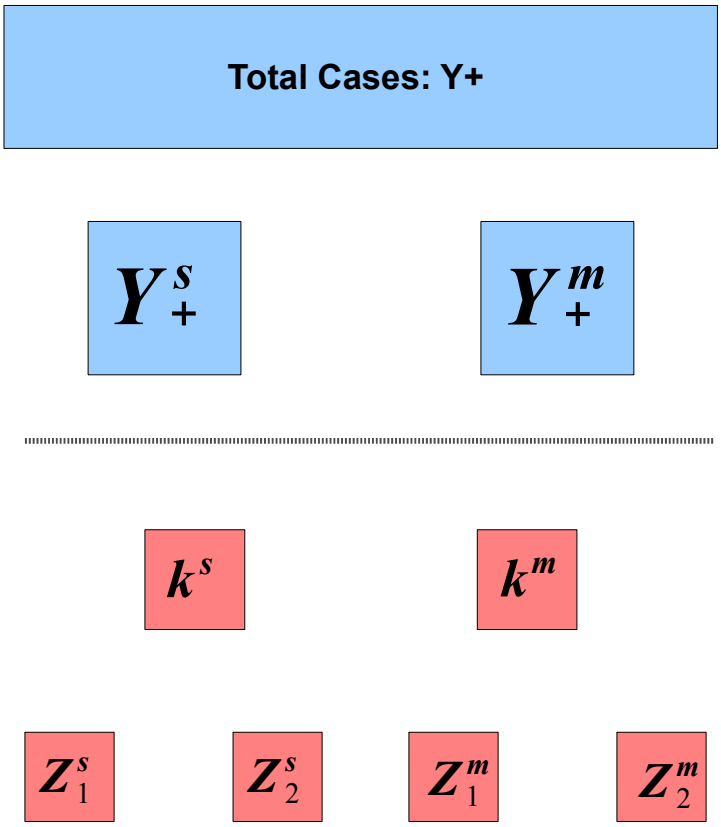


Figure 5.3: Observed data for a generic area, with blue boxes showing the surveillance data and salmon boxes showing the lab test data.

A naive approach to estimating the parameters of interest is as follows:

$$\begin{aligned}\hat{q} =& Pr(\text{strain 1}|\text{case}) = Pr(\text{strain 1}|\text{severe})Pr(\text{severe}) + Pr(\text{strain 1}|\text{mild})Pr(\text{mild}) \\ =& (z_1^s/k_s) \times (y_+^s/y_+) + (z_1^m/k_m) \times (y_+^m/y_+), \end{aligned} \tag{5.1}$$

$$\begin{aligned}\hat{p}_1^s =& Pr(\text{severe}|\text{strain 1}) = \frac{Pr(\text{strain 1}|\text{severe})Pr(\text{severe})}{Pr(\text{strain 1})} \\ =& \frac{Pr(\text{strain 1}|\text{severe})Pr(\text{severe})}{Pr(\text{strain 1}|\text{severe})Pr(\text{severe}) + Pr(\text{strain 1}|\text{mild})Pr(\text{mild})} \\ =& \frac{(z_1^s/k_s) \times (y_+^s/y_+)}{(z_1^s/k_s) \times (y_+^s/y_+) + (z_1^m/k_m) \times (y_+^m/y_+)} \\ =& \frac{(z_1^s/k_s) \times (y_+^s/y_+)}{\hat{q}}, \end{aligned} \tag{5.2}$$

$$\begin{aligned}\hat{p}_2^s =& Pr(\text{severe}|\text{strain 2}) = \frac{Pr(\text{strain 2}|\text{severe})Pr(\text{severe})}{Pr(\text{strain 2})} \\ =& \frac{Pr(\text{strain 2}|\text{severe})Pr(\text{severe})}{Pr(\text{strain 2}|\text{severe})Pr(\text{severe}) + Pr(\text{strain 2}|\text{mild})Pr(\text{mild})} \\ =& \frac{(z_2^s/k_s) \times (y_+^s/y_+)}{(z_2^s/k_s) \times (y_+^s/y_+) + (z_2^m/k_m) \times (y_+^m/y_+)} \\ =& \frac{(z_2^s/k_s) \times (y_+^s/y_+)}{\hat{q}} \end{aligned} \tag{5.3}$$

To get the estimated total counts of strain 1, and strain-specific severe cases we have:

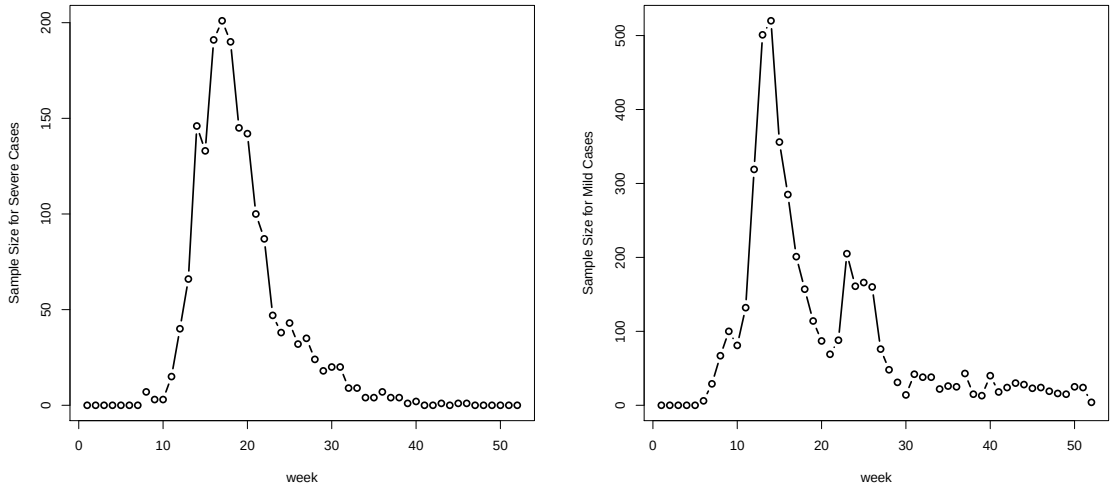
$$\hat{y}_1 = y_+ \times \hat{q}, \tag{5.4}$$

$$\hat{y}_1^s = \hat{y}_1 \times \hat{p}_1^s, \tag{5.5}$$

$$\hat{y}_2^s = \hat{y}_2 \times \hat{p}_2^s. \tag{5.6}$$

It is easy to see that when the lab test data is not available (i.e., k_s or k_m is 0), the naive estimates in (5.1), (5.2) and (5.3) are no longer defined. This is is relevant for the China HFMD data. In Figure 5.4 we present the sample sizes of the lab tested cases for the clinically diagnosed severe and mild patients. The sample sizes vary greatly over the year, with zero samples in the first and last several weeks for the severe cases. In addition,

even when lab test data are available, if the lab test sample size is small which may occur for example, if we cross tabulate the data by area and time, the naive estimates are highly unreliable.



(a) Severe Cases

(b) Mild Cases

Figure 5.4: Sample sizes of the lab tested cases, (a) K^s and (b) k^m, for the Henan 2009 HFMD data by week.

5.3 A Strain-Specific Model: For A Generic Area and Time Period

The model we develop for a generic area introduces some additional variables, and is shown in Figure 5.5. The solid arrow indicates conditional independence while the dashed arrow indicates a deterministic relationship. The rectangular boxes are used for observed data while the circles are used for random variables. We let Y_1 and Y_2 denote the number of HFMD cases with strain 1 and strain 2. Let Y_+ denote the total number of HFMD cases where $Y_+ = Y_1 + Y_2$. We assume Y_1 follows a Binomial(Y_+, q) distribution, with q being the probability of having strain 1 given being a HFMD case, i.e. $q = Pr(\text{strain 1}|\text{case})$. Let Y_1^s denote the number of severe cases from strain 1, and we assume Y_1^s follows a Binomial(Y_1, p_1^s) distribution where p_1^s is the probability of being a severe case given having

strain 1, i.e. $p_1^s = Pr(\text{severe}|\text{strain 1})$. Similarly, we define Y_2^s as the number of severe cases from strain 2, and we assume Y_2^s follows a Binomial(Y_2, p_2^s) distribution where p_2^s is the probability of being a severe case given having strain 2, i.e. $p_2^s = Pr(\text{severe}|\text{strain 2})$. The rest of the variables in the diagram remain as described in Section 5.2. For the lab test data, we assume that Z_1^s follows a Hypergeometric(k^s, Y_+^s, Y_1^s) distribution and Z_1^m follows a Hypergeometric(k^m, Y_+^m, Y_1^m) distribution.

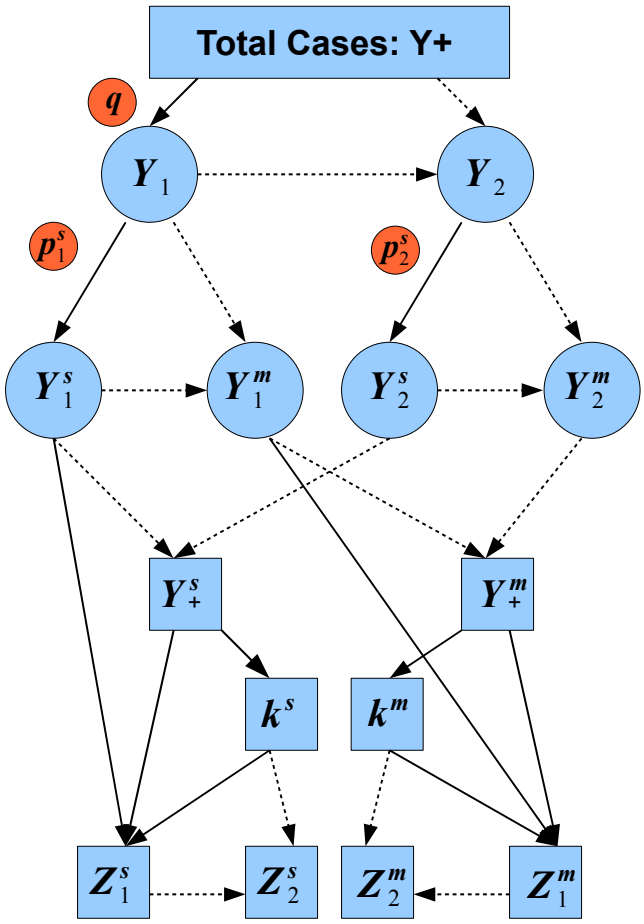


Figure 5.5: Diagram showing the strain-specific conditional independencies we assume to define a hierarchial model, hierarchial model we propose for a generic area and time period.

Compared to Figure 5.3, the missing pieces of the strain-specific variables Y_1, Y_2, Y_1^s, Y_1^m, Y_2^s and Y_2^m are imputed by our proposed model. Through these variables, our proposed model connects the observed surveillance data and the lab test data and provides a coherent and unified way to estimate all variables simultaneously.

In Appendix B, we provide the detailed derivation of the likelihood and the conditional distributions in our proposed model for a generic area and time period, when using simple noninformative priors. Inference for the proposed model is made using Bayesian MCMC, since the model is highly intractable. The Metropolis-Hastings algorithm is needed as some of the conditional posterior distributions are not in standard form. The MCMC procedure now includes sampling from the discrete variables Y_1 and Y_1^s which have constraints on their ranges. Details of the MCMC algorithms are described in Appendix C.

5.4 A Strain-Specific Model: With Temporal Component

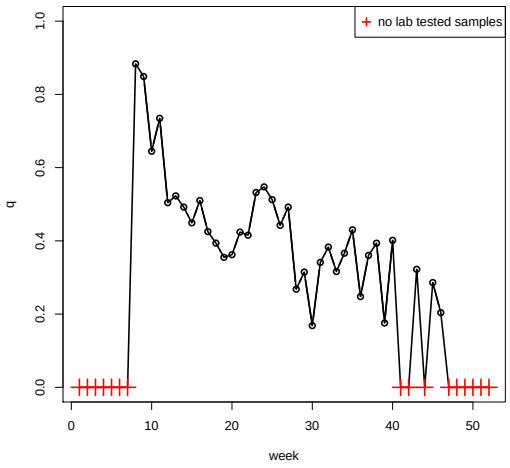
As mentioned earlier, we are interested in how the probabilities q, p_1^s and p_2^s evolve over time. In Figure 5.6 we present the naive estimates of q, p_1^s and p_2^s with the weekly HFMD data in Henan providence in 2009. We use a red cross to indicate when there are no lab test samples. The variability in the naive estimate is large and the temporal pattern is difficult to identify. This suggests that the naive estimates are not satisfactory and we need to investigate more complicated models that can "smooth" out the large variation we see in the naive estimates.

We start by fitting a simple model with independent normal random effect on the linear predictor scale

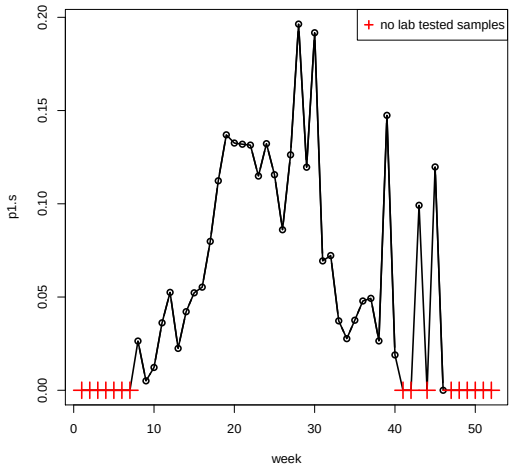
$$\text{logit}(p_{1t}^s) = \alpha_1 + \gamma_{1t}, \tag{5.7}$$

$$\text{logit}(p_{2t}^s) = \alpha_2 + \gamma_{2t}, \tag{5.8}$$

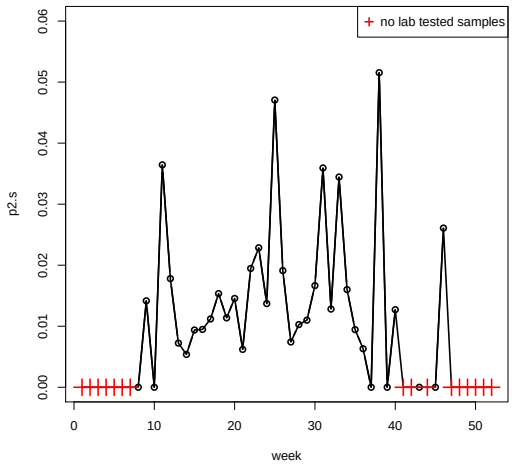
where $\gamma_{1t} \sim_{iid} \text{N}(0, \tau_{\gamma_1}^{-1})$ and $\gamma_{2t} \sim_{iid} \text{N}(0, \tau_{\gamma_2}^{-1})$. We use flat priors for α_1 and α_2, and a Gamma(1, 0.26) prior for the precision parameters τ_{γ_1} and τ_{γ_2} as suggested in Fong et al. (2010). The parameter q_t is assigned a noninformative Beta$(1, 1)$ prior. We run this model for $850,000$ iterations with the first $50,000$ discarded as burn-in. Convergence of the MCMC iterations is assessed using the trace plots. Some selected trace plots are shown in Figure 5.7,



(a) q



(b) p_1^s



(c) p_2^s

Figure 5.6: Time series plot of the naive estimates of q, p_1^s and p_2^s, with Henan HFMD data aggregated over space.

where we show every fifth MCMC sample. We use the posterior median for inference, which is estimated as -2.84 for α_1 and -4.32 for α_2. The estimated latent variables γ_{1t} and γ_{2t} are shown in Figure 5.8. The time series plot of the estimated γ_{1t} shows elevated values

between weeks 15 and 30, followed by a small bump between weeks 40 and 45. The time series plot of γ_{2t}, however, does not reveal any clear pattern. We also present the scatterplot of the latent variables γ_{1t} and γ_{2t}, in an attempt to see if they are associated. However, no clear pattern is observed in the plot.

The estimated probabilities q, p_1^s and p_2^s are shown in Figure 5.9, along with the naive estimates for comparison. The overall shrinkage obtained by using the normal random effect model is obvious in these plots. Also, notice that in the normal random effect model, when there is no lab-tested sample, the estimated probability is close to the overall average for all three parameters q, p_1^s and p_2^s. The temporal pattern is more clear from the estimates using the normal random effect model. However, the estimates still lack the smoothness we desire.

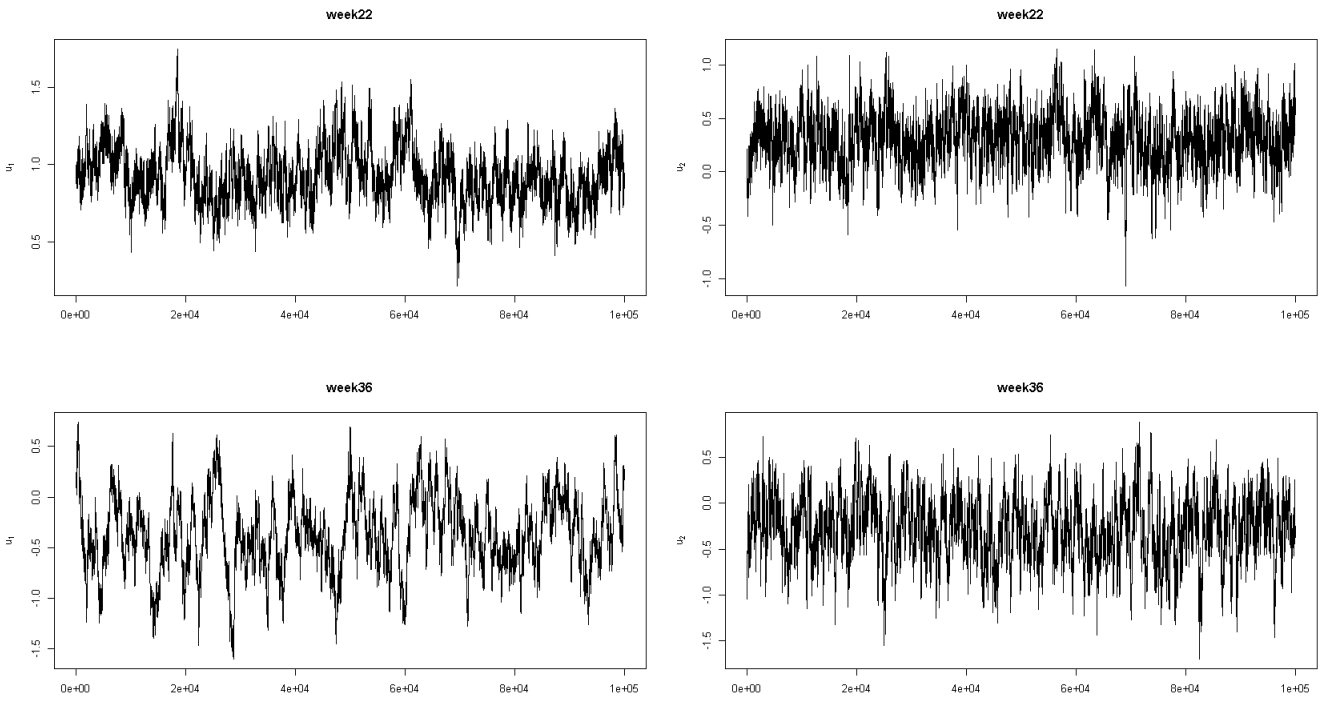
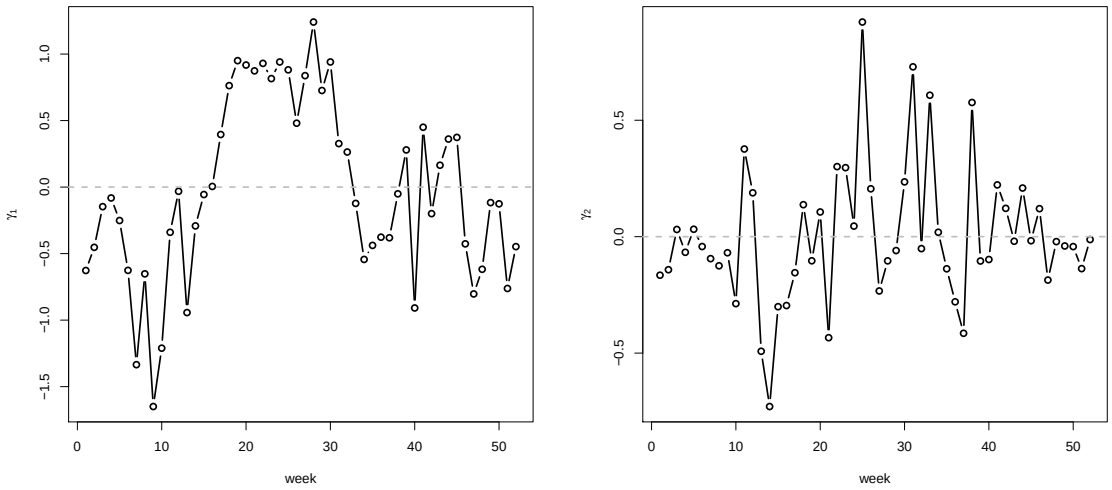
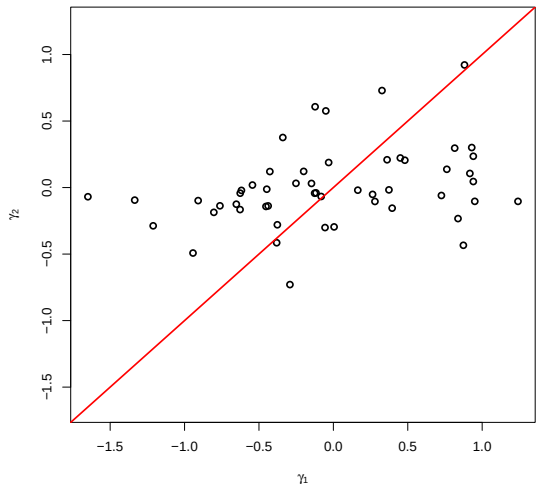


Figure 5.7: Selected trace plots using the normal random effect model.

To achieve a smoothed temporal fit, we adopt a spline model with a cubic truncated



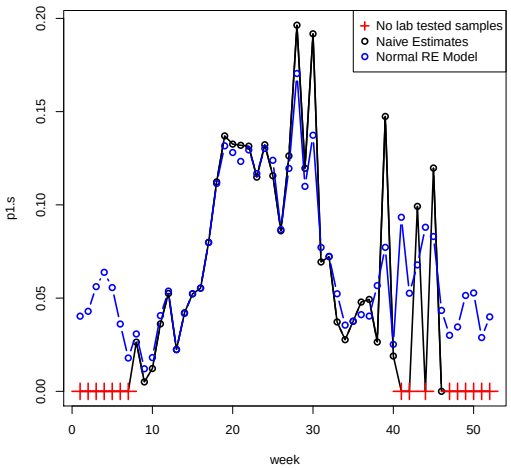
(a) $\hat{\gamma}_1$ (b) $\hat{\gamma}_2$



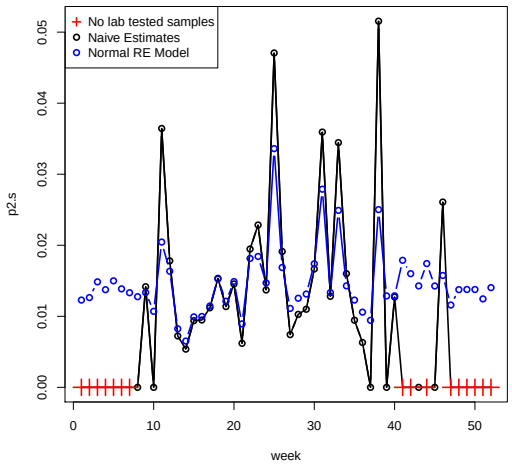
(c) $\hat{\gamma}_1$ vs. $\hat{\gamma}_2$

Figure 5.8: Estimated latent variables γ_{1t} and γ_{2t} using the normal random effect model.

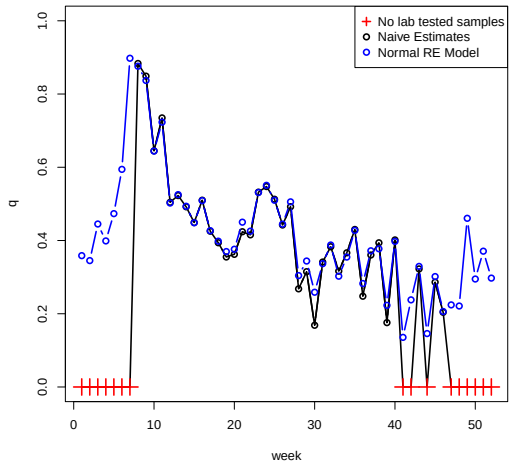
power splines as described in Section 4.2.1. The truncated power spline model is applied on



(a) $\hat{p}_1^s$



(b) $\hat{p}_2^s$



(c) $\hat{q}$

Figure 5.9: Estimated probabilities p_1^s, p_2^s and q using the normal random effect model.

the logit link function scales for p_1^s and p_2^s

$$\text{logit}(p_{1t}^s) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{k=1}^{2} b_k (t - \kappa_k)_+^3, \tag{5.9}$$

$$\text{logit}(p_{2t}^s) = \xi_0 + \xi_1 t + \xi_2 t^2 + \xi_3 t^3 + \sum_{k=1}^{2} b_k (t - \kappa_k)_+^3, \tag{5.10}$$

with the knots chosen at weeks 18 and 35. We choose noninformative normal priors for all regression coefficients in the model. We run this model for $850,000$ iterations with the first $50,000$ discarded as burn-in. The estimated probabilities q, p_1^s and p_2^s are shown in Figure 5.10. Again we include the naive estimates in the plot for comparison. The plot for fit p_1^s shows one peak around week 25. The small increase at the end of the year may be an artifact of the lack of data in the last few weeks. The estimated p_2^s shows a steady increase over the entire year. However, the scale of the estimated p_2^s is very small and therefore the increase is small. Around week 46 we see a small peak in the estimated p_2^s, however, notice that between weeks 41 and 52, we only have lab tests data in 3 of the weeks. Therefore, the observed data in the 3 weeks can highly influence the estimates for these later weeks and suggests a misleading peak.

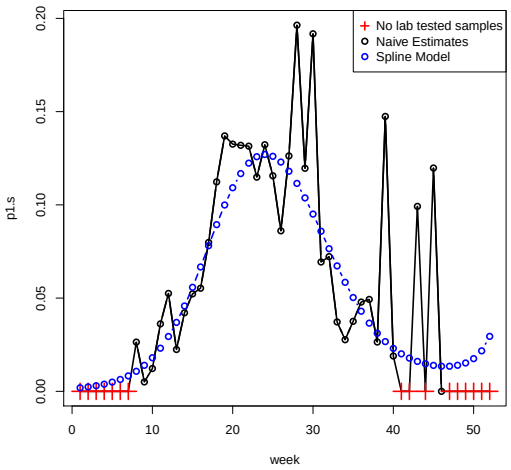
Compared to the normal random effect model, the spline models obviously provide much smoother estimates of the temporal patterns. However, in both the temporal models we have investigated, we assume independence between different strains based on the exploratory analysis. Jointly modeling the different strains is discussed in Section 5.6.

5.5 A Strain-Specific Model: With Spatial Component

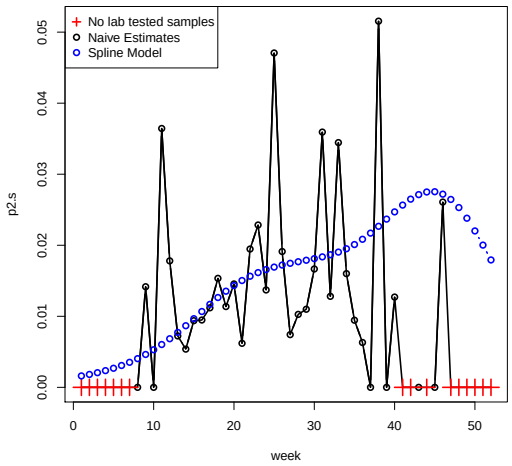
Besides the temporal pattern, we are also interested in the spatial pattern of the parameters q, p_1^s and p_2^s. In Table 5.2 we present summaries of the surveillance data and the lab test data at the prefecture level in Henan providence. The corresponding locations of the prefectures are shown in Figure5.11 (a). Notice that in prefecture 11, there are no observed severe cases and lab-tested severe cases (numbers indicated in red in the table). Therefore, the naive estimates are no longer available for this area.

For visualization purpose, we set the naive estimates to be 0 when there is no lab-tested samples and present the maps of the naive estimates of the three probabilities of interest in Figures5.11 (b) and (c). It is difficult to discern spatial pattern from these maps, probably due to the small number of areas that we are investigating.

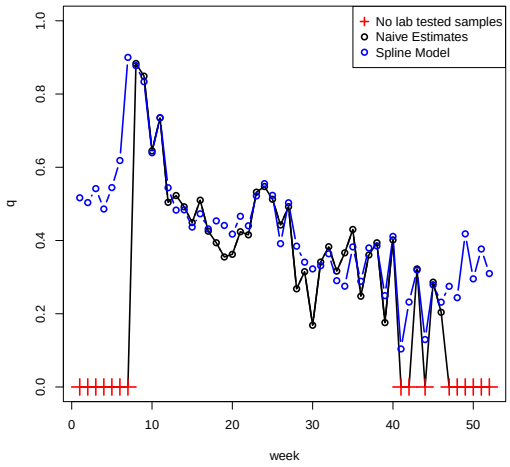
We begin by investigating a spatial smoothing model on the probability p. Let i index prefecture, $i = 1, \ldots, 18$. The spatial smoothing is imposed through a prior on the



(a) $\hat{p}_1^s$



(b) $\hat{p}_2^s$



(c) $\hat{q}$

Figure 5.10: Estimated probabilities p_1^s, p_2^s and q using truncated power spline model.

probability p_i

$$\text{logit}(p_i) = \alpha + u_i,$$

where $\boldsymbol{u} \sim \text{ICAR}(\tau_u^{-1})$ as introduced in Section 2.2. We use noninformative Beta(1,1) priors for both p_1^s and p_2^s. This model is run for $850,000$ iterations with the first $50,000$

discarded as burn-in. The results are shown in Figure 5.12. For visual comparison, we use the same scale in the maps as was used for the naive estimates. The overall pattern is very similar between these two maps. However, notice now that the probability of p in area 11 is estimated to be 0.47, which is pulled up towards the average of its neighbors, as we would expect from spatial smoothing.

We can also perform similar spatial smoothing on the parameter p_1^s and p_2^s, but with the small number of prefectures the benefit of this approach may not be large. A more sensible approach is to investigate the spatial smoothing at a smaller geographical scale, for example, at the township scale.

5.6 Discussion

In this chapter, we have proposed a Bayesian hierarchical model that combine both the surveillance data and the lab-tested data and have made simultaneous inference about the strain-specific parameters of interest. In addition, we have performed some exploratory analysis to investigate the spatial and temporal patterns in the strain-specific probabilities of interest, and have compared the temporal trends between the strain-specific probabilities.

We emphasize that the results we have shown from the selected models serve as an exploratory analysis only, and are not considered as models for making firm conclusions. In order to make valid conclusions we should perform model checking and also examine model selection, which is certainly included in future model development.

In terms of model development of our proposed model, we should consider jointly modeling the three probabilities q, p_1^s and p_2^s over time and/or over space. Taking time as an illustration, two methods may be considered for this purpose, the first being to use a multivariate normal distributions on the logit link function scale:

$$\text{logit}\begin{pmatrix} p_{1t}^s \\ p_{2t}^s \\ q_t \end{pmatrix} \sim \text{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ is the mean vector and $\boldsymbol{\Sigma}$ is the 3×3 covariance matrix. The mean function $\boldsymbol{\mu}$ can be modeled in a parametric fashion, for example, with a frequency domain

Table 5.2: Henan 2009 HFMD data at the prefecture level.

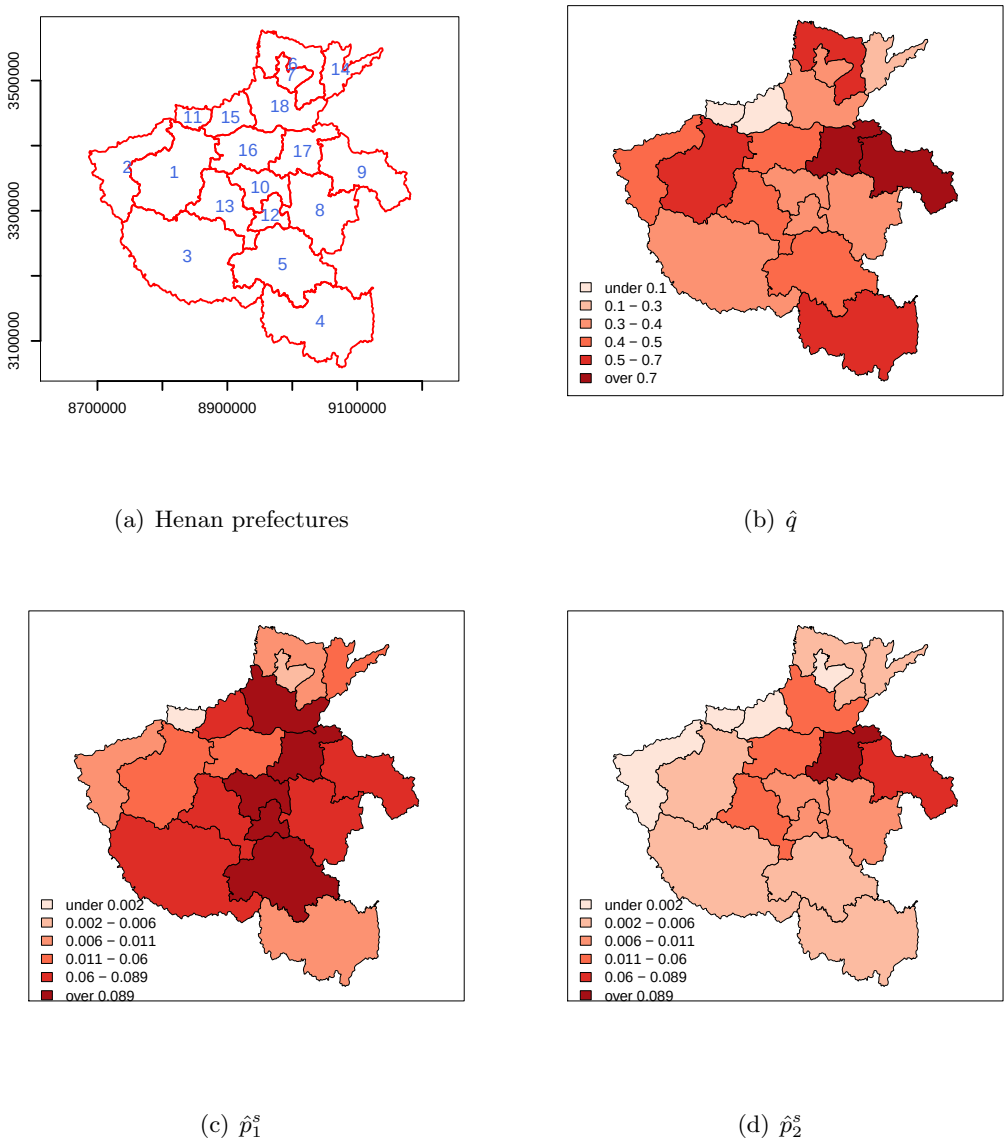
Prefecture	Y_+^s	Y_+^m	Y_+	z_1^s	z_2^s	z_1^m	z_2^m
1	148	6975	7123	75	6	93	79
2	4	1544	1548	2	0	31	42
3	184	6782	6966	89	11	44	94
4	19	4441	4460	5	2	6	6
5	227	2950	3177	30	1	4	5
6	23	4385	4408	8	3	36	31
7	9	3159	3168	5	3	48	95
8	225	7380	7605	75	15	32	66
9	597	8302	8899	325	37	327	27
10	265	6541	6806	95	18	30	59
11	0	503	503	0	0	14	15
12	102	2781	2883	40	10	23	54
13	271	5767	6038	81	43	58	86
14	26	6088	6114	12	8	19	81
15	42	4337	4379	9	1	5	51
16	549	16741	17290	271	69	1177	1530
17	1051	4335	5386	325	57	140	25
18	179	4160	4339	49	13	29	63

model

$$\mu_l = \alpha_l + \beta_l t + \sum_{s=1}^{S} \left(a_{ls}\sin(\omega_s t) + b_{ls}\cos(\omega_s t)\right), \qquad l = 1, 2, 3, \quad s = 1, \ldots, T/2,$$

or in a semiparametric fashion as in the spline models we used in Section 5.5. The structure of the covariance $\mathbf{\Sigma}$ can take various forms. For example, we can use a Wishart distribution as the prior of $\mathbf{\Sigma}^{-1}$ to estimate the elements in the covariance matrix. The structure of the covariance matrix can determine how the three probabilities relate to each other.

An alternative method is the shared component model introduced in Knorr-Held and



(a) Henan prefectures

(b) $\hat{q}$

(c) $\hat{p}_1^s$

(d) $\hat{p}_2^s$

Figure 5.11: Naive estimation for data aggregated over time.

Best (2001). For example, to jointly model probabilities p_{1t}^s and p_{2t}^s we have

$$\text{logit}(p_{1t}^s) = \alpha_1 + \psi\nu_t + \phi_{1t}, \tag{5.11}$$

$$\text{logit}(p_{2t}^s) = \alpha_2 + \frac{1}{\psi}\nu_t + \phi_{2t}. \tag{5.12}$$

So the log odds ratio for being a severe case given strain 1 and 2 can be decomposed into

a shared temporal random effect between strains, and a strain-specific temporal random effect. The scaling parameter ψ determines the effect of the shared component on each strain.

Clearly, similar ideas can be applied to joint modeling the spatial pattern. In the spatial-temporal analysis, we would include the shared spatial and temporal components, and the strain-specific spatial and temporal components. This is a model we plan to investigate in future work.

For the future work, we should also examine the effects of covariates on the probabilities of having any particular strain given a case, and the strain-specific probability of being a severe case. These covariates should include the usual confounders in epidemiological studies such as age, gender and race, as well as area-level ecological covariates. Another extension of our proposed model is to include multiple strains and multiple clinical severity categories. The mathematical derivation for such extension is straightforward, with careful bookkeeping required on the range of counts for each strain and severity category. For the future work, we will also investigate the prediction of the total number of disease counts, and the strain-specific number of cases. We will examine the accuracy of the prediction, and how far into the future we can provide prediction with predetermined accuracy.

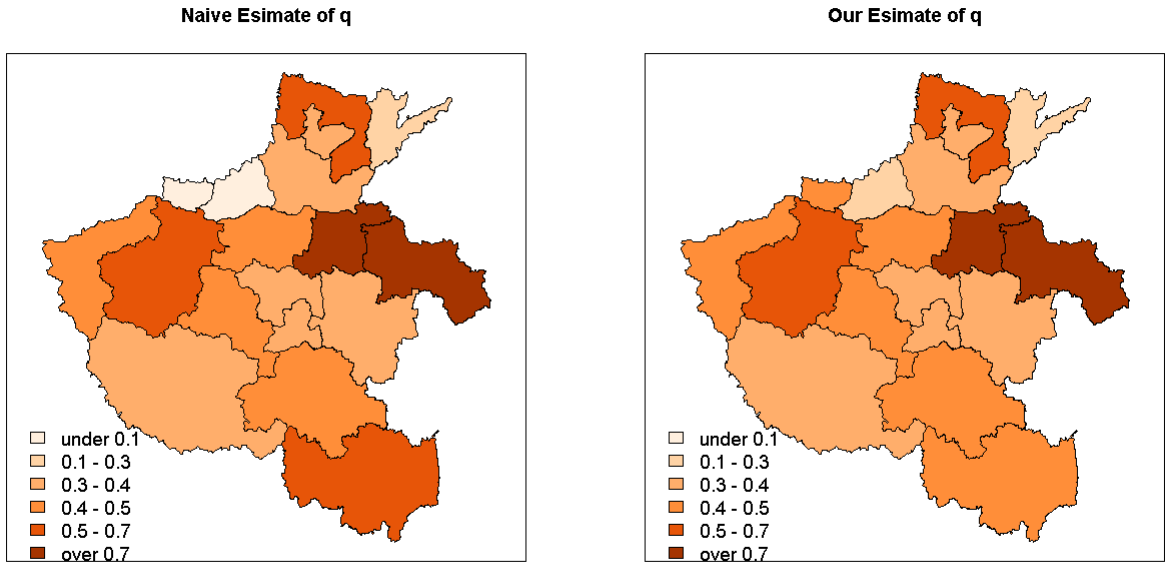


Figure 5.12: Comparison of the estimated q using the naive estimation and the strain-specific model.

Chapter 6

CONCLUSIONS AND FURTHER WORK

There has been a recent increase in interest in the spatial-temporal modeling of public health data. Our work has been motivated by two real-world data sets, the Washington BRFSS data from 2006 and China HFMD data from 2009 to 2010. In this dissertation, three different aspects of spatial temporal modeling have been considered.

Chapter 3 addressed the sampling issues in Bayesian hierarchical models in small area estimation. A pragmatic approach to incorporate the sampling weights in Bayesian hierarchial models was proposed. A simulation study showed that when estimating the proportion or the total counts in the outcome of interest, the bias resulting from non-response can be greatly reduced using the proposed method while the variance can be greatly reduced using Bayesian hierarchical models. The method was then applied to the 2006 Washington BRFSS data.

Chapter 4 developed penalized spline models to investigate the spatial and temporal patterns in infectious diseases. The structure of the spatial, temporal and spatial-temporal interaction was allowed to have one of four possible forms and these structures were imposed through the prior distributions. The simulation study we conducted showed that the proposed models can identify the movement of epidemic centers which is of primary interest to epidemiologists. The method was then applied to the China HFMD surveillance data from 2009 and 2010.

In the case of the China HFMD data, in addition to the surveillance data we also have lab test results that record the specific virus strain present in the patient. However, the lab tests are available for only a very small number of patients, therefore inference made for each strain type is highly unreliable when using the lab test data alone. In Chapter 5 we proposed a method for combining the surveillance data with the lab test data into a single, coherent Bayesian hierarchical model. We demonstrated how to extend our proposed

model to include the spatial and temporal components and presented some initial results as exploratory data analysis using space-time models.

Each individual aspect dealt with in this dissertation has potential for future work. With respect to Chapter 3, future work will investigate the performance of our proposed method on correcting selection bias, which occurs when the sampling scheme is highly correlated with the outcome variable. For example, in the context of the diabetes data, let x be a surrogate covariate that is highly correlated with diabetes status in a patient. If the sampling is conducted based on variable x, then model-based inference will be biased if we do not include the individual level covariate x in the analysis. However, such individual level information may be lost when we aggregate data over space. In this case, selection bias occurs when we attempt to make inference about the diabetes prevalence in each area. Some initial simulation studies have shown that our proposed model has the potential to correct such selection bias. In the future, we plan to conduct extensive simulation studies to evaluate the ability of our proposed method to correct such selection bias.

In Chapter 4, we carried out fitting based on the risk surface evaluated at the centroid of area i. A more appropriate method is to average the risk surface over the area at time t

$$\int_{\boldsymbol{x}\in A_i} \exp[f_2(\boldsymbol{x}) + f_3(\boldsymbol{x}, t)]d\boldsymbol{x},$$

where $f_2(\boldsymbol{x})$ denotes the function for spatial smoothing and $f_3(\boldsymbol{x}, t)$ the function for the space-time interaction. This integral may be evaluated on a grid laid out over the study area. The fitted risk surface from such an integral is aggregate consistent, meaning that area level estimates are consistent with point level estimates. Aggregation inconsistency is known as the ecological bias and is described in Wakefield and Salway (2001) and Wakefield (2003).

With reference to the material of Chapter 5, we recognize the connection between our proposed strain-specific model and a hybrid model that combines the ecological and case-control data, introduced in Haneuse and Wakefield (2008). In Table 6.1, the ecological and case-control data in the hybrid model is summarized via two 2×2 tables. Let X denote a binary exposure variable and Y a binary outcome variable. In the ecological data we observe the aggregate response $N_1 = N_{10} + N_{11}$, as well as the marginal exposure data M_0

and M_1. However, the internal cells are not directly observed. In the case-control sample, n_0 controls are randomly selected from the N_0 total non-cases, and n_1 cases are randomly selected from the N_1 total cases. In the hybrid model, the internal cells N_{10} and N_{11} in the table showing the ecological data, and n_{01} and n_{11} in the table showing the case-control data, are treated as random variables and their values can be imputed either by maximum likelihood or using a Bayesian approach.

In the China HFMD data, the surveillance and lab test data can also be summarized into two 2×2 tables shown in Table 6.2. We observe marginal counts Y_+^s and Y_+^m and would like to fill in the internal cells, using the information from the table of the lab test data. Compared to Table 6.1, the internal cells with missing data are different. In the China HFMD, the internal cells in the table for lab test data are all filled. However, the marginal counts for the surveillance data are not observed, and therefore we have to impute both the marginal Y_1 and the internal Y_1^s. The connection between the two set-ups is interesting on its own and deserves further study.

Table 6.1: Ecological and case-control data with a binary exposure X (values in square brackets are unobserved).

(a)

	Ecological		
X	$Y=0$	$Y=1$	
0		$[N_{10}]$	M_0
1		$[N_{11}]$	M_1
	N_0	N_1	N

(b)

	Case–control		
X	$Y=0$	$Y=1$	
0	n_{00}	n_{10}	
1	n_{01}	n_{11}	
	n_0	n_1	n

In this dissertation, we have discussed three different issues when modeling public health data with space and time components. For each of the issues, we have proposed an approach to solving existing problems. However, our proposed approaches are in fact related to each other and can be combined into a single framework. For example, in the strain-specific model we combined surveillance data with lab test data. When such lab test data is the result of a complex survey, we can apply the method we developed to incorporate the

Table 6.2: Surveillance and lab test data (values in square brackets are unobserved).

(a) Surveillance data

	Severity category		
Strain	Severe	Mild	
1	$[Y_1^s]$		$[Y_1]$
2			
	Y_+^s	Y_+^m	Y_+

(b) Lab test data

	Severity category		
strain	severe	mild	
1	Z_1^s	Z_1^m	
2	Z_2^s	Z_2^m	
	k_s	k_m	

sampling weights in the strain-specific model. The penalized spline model we proposed to investigate the spatial, temporal and spatial-temporal interaction pattern can also be added as another layer in the strain-specific model. We believe that our approaches show great promise, especially given the increasing availability of public health data.

BIBLIOGRAPHY

Assunçaõ, R. M., I. A. Reis, and C. D. Oliveira (2001, Aug). Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model. *Statistics in Medicine 20*, 2319–2335.

Bernardinelli, L., D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine 14*, 2433–2443.

Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43*, 1–20.

Clayton, D. (1996). Generalized linear mixed models. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 275–301. Chapman & Hall, London.

Cressie, N. and N. Chan (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association 84*, 393–401.

Currie, I., M. Durbán, and P. Eilers (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 68*, 259–280.

Datta, G. and M. Ghosh (1991). Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics 19*, 1748–1770.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer.

Diggle, P. (2007). *Model-based Geostatistics*. Springer.

Eilers, P. and B. Marx (1996). Flexible smoothing using b-splines and penalized likelihood. *Statistical Science 11*, 89–121.

Fahrmeir, L., T. Kneib, and Lang (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica 14*, 715–745.

Fahrmeir, L. and S. Lang (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society Series C (Applied Statistics) 50*, 201–220.

Farrell, P. (2000). Bayesian inference for small area proportions. *The Indian Journal of Statistics, Series B 62*, 402–416.

Farrell, P., B. MacGibbon, and T. Tomberlin (1997). Empirical Bayes estimation of small area proportions in multistage designs. *Statistical Sinica 7*, 1065–1083.

Fay, R. and R. Herriot (1979). Estimates of income for small places: an application of James–Stein procedure to census data. *Journal of the American Statistical Association 74*, 269–277.

Fong, Y., H. Rue, and J. Wakefield (2010). Bayesian inference for generalized linear mixed models. *Biostatistics 11*(3), 397–412.

French, J. L. and M. P. Wand (2004). Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics 5*(2), 177–191.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics 19*, 1–141.

Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics 31*, 3–39.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science 22*, 153–164.

Ghosh, M., K. Natarajan, T. Stroud, and B. Carlin (1998). Generalized linear models for small-area estimation. *Journal of American Statistical Association 93*(441), 55–93.

Graubard, B. and E. Korn (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science 17*, 73–96.

Gu, C. (2002). *Smoothing spline ANOVA models*. Springer.

Gu, C. and G. Wahba (1991). Discussion: Multivariate adaptive regression splines. *The Annals of Statistics 19*(1), pp. 115–123.

Haneuse, S. and J. Wakefield (2008). The combination of ecological and case-control data. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 70*, 73–93.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Held, L., M. Höhle, and M. Hofmann (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling 5*, 187–199.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

Kammann, E. E. and M. P. Wand (2003). Geoadditive models. *Applied Statistics 52*, 1–18.

Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics 11*, 55–77.

Kneib, T. and L. Fahrmeir (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics 62*, 109–118.

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

Knorr-Held, L. and J. Besag (1998). Modelling risk from a disease in time and space. *Statistics in Medicine 17*(18), 2045–2060.

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

Knorr-Held, L. and S. Richardson (2003). A Hierarchical model for space-time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 52*(2), 169 – 183.

Knorr-Held, L. and H. Rue (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics 29*(4), 597–614.

Kooperberg, C. and C. J. Stone (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics 1*(4), pp. 301–328.

Korn, E. and B. Graubard (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology 23*, 192–201.

Lang, S. and A. Brezger (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics 13*, 183–212.

Lawson, A. (2009). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. Chapman & Hall/CRC.

Lee, D. and M. Durbán (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling 11*(1), 49–69.

MacNab, Y. C. and C. B. Dean (2002). Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine 21*, 347–358.

MacNab, Y. C. and P. Gustafson (2007). Regression B-spline smoothing in Bayesian disease mapping: with an application to patient safety surveillance. *Statistics in Medicine 26*(24), 4455–4474.

Malec, D., J. Sedransk, C. L. Moriarity, and F. B. LeClere (1997). Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association 92*(439), 815–826.

Mohadjer, L., J. Montaquila, J. Waksberg, B. Bell, P. James, I. Flores-Cervantes, and M. Montes (1996). Nhanes iii weighting and estimation methodology. Technical report, National Center for Health Statistics.

Mugglin, A., N. Cressie, and I. Gemmell (2002). Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine 21*(18), 2703–2721.

Nandram, B. and J. Sedransk (1993). Bayesian predictive inference for a finite population proportion: two-stage cluster sampling. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 55*, 399–408.

Nychka, D. and N. Saltzman (1998). Design of air quality monitoring networks. *Lect. Notes Statist. 132*, 51–76.

Paul, M., L. Held, and A. M. Toschke (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine 27*(29), 6250–6267.

Pfeffermann, D. (2002). Small area estimation — new developments and directions. *International Statistical Review 70*(1), 125–143.

Plummer, M. (2009). JAGS Version 1.0.3 Manual. Technical report.

Raghunathan, T., D. Xie, N. Schenker, V. Parsons, W. Davis, K. Dodd, and E. Feuer (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association 102*, 474–486.

Rao, J. (1999). Some recent advances in model-based small area estimation. *Survey Methodology 25*(2), 12–001.

Rao, J. (2003). *Small Area Estimation*. New York: John Wiley and Sons.

Rao, J. and C. Wu (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 72*, 533–544.

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

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability 7*(1), 110–120.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 63*(2), 325–338.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall, London.

Rue, H. and S. Martino (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), 1–35.

Ruppert, D. and M. Wand (2000). Spatially-adaptive penalites for spline fitting. *Australian and New Zealand Journal of Statistics 42*, 205–223.

Spiegelhalter, D., A. Thomas, and N. Best (1998). *WinBUGS User Manual* (1.1.1 ed.). Cambridge, UK.

Stroud, T. (1994). Bayesian inference from categorical survey data. *Canadian Journal of Statistics 22*, 33–45.

Tong, C. Y. W. and J. M. Bible (2009). Global epidemiology of enterovirus 71. *Future Virology 4*(5), 501–510.

Ugarte, M. D., T. Goicoa, and A. F. Militino (2010, May-Jun). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics 21*(3-4), 270–289.

Wakefield, J. (2003). Sensitivity analyses for ecological regression. *Biometrics 59*(1), 9–17.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics 8,2*, 158–186.

Wakefield, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology 38*, 330–336.

Wakefield, J., S. Haneuse, A. Dobra, and E. Teeple (2011). Bayes computation for ecological inference. *Statistics in Medicine 30*(12), 1381–1396.

Wakefield, J. and R. Salway (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 164*(1), 119–137.

Waller, L. A., B. P. Carlin, H. Xia, and A. E. Gelfand (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association 92*, 607–617.

Wand, M. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics 15*, 443–462.

Wang, Y., Z. Feng, Y. Yang, S. Self, Y. Gao, I. Longini, J. Wakefield, J. Zhang, L. Wang, X. Chen, L. Yao, J. Stanaway, Z. Wang, and W. Yang (2011). Hand, foot, and mouth disease in China: patterns of spread and transmissibility. *Epidemiology 23*(2), 781–792.

Wood (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 62*(2), 413–428.

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

Appendix A

SELECTED R CODE FOR FITTING PENALIZED SPLINE MODEL IN INLA

```
## Y is the matrix for the observations
## with I rows for areas and T column for time points
n.time <- ncol(Y)
n.area <- nrow(Y)
# Z is the design-matrix for the spline basis
n.basis <- ncol(Z)

##-----------------------------------------##
### for space-time interaction models
head(Y)
y <- c(t(Y)) # yi1 .... yit
E.lvec <- rep(E, each=n.time)
n <- length(y) # n.area*n.time
m <- n.basis*n.time

# create the stacked design matrix
X.mat <- NULL
I.tmp <- diag(1, nrow=n.time, ncol=n.time)
for (i in 1:n.area){
test <- I.tmp %x% matrix(Z[i,], nrow=1)
X.mat <- rbind(X.mat, test)
}
dim(X.mat) # should be (IT)*(KT)
```

```
##---- type I interaction
intercept <- c(1,  rep(NA, n.basis*n.time))
idx       <- c(NA, 1:(n.basis*n.time))
data.inla <- list(y=y, idx=idx, intercept=intercept)

formula <- y ~ -1 + f(intercept, model="iid") +
f(idx, model = "iid",param=c(1, 0.005))

r.1 <- inla(formula,data=data.inla,family="poisson",
control.predictor=list(A=cBind(rep(1, nrow(X.mat)), X.mat),
compute=TRUE),verbose=T, E=E.lvec)
summary(r.1)

# extract intercept term
r.1$summary.linear.predictor[(n+1), ]
# extract the spline coefficients
b.hat.1.inla <- r.1$summary.linear.predictor[(n+2):(n+1+m), ]
b.hat.mat.1.inla <- matrix(b.hat.1.inla, nrow=n.basis, ncol=n.time, byrow=F)



##---- type II interaction
intercept <- c(1,  rep(NA, n.basis*n.time))
idx       <- c(NA, rep(1:n.basis, each=n.time))
data.inla <- list(y=y, idx=idx, intercept=intercept)

formula.2 <- y ~ -1 + f(intercept, model="iid") +
f(idx, model = "rw2", constr=F, param=c(1, 0.005))
r.2 <- inla(formula.2, data=data.inla,family="poisson",
```

```
control.predictor = list(A=cBind(rep(1, nrow(X.mat)), X.mat),
compute=TRUE),verbose=T, E=E.lvec)
summary(r.2)

# extract intercept term
r.2$summary.linear.predictor[(n+1), ]

# extract the spline coefficients
b.hat.2.inla <- r.2$summary.linear.predictor[(n+2):(n+1+m), "mean"]
b.hat.mat.2.inla <- matrix(b.hat.2.inla, nrow=n.basis, ncol=n.time, byrow=F)
```

Appendix B

DERIVATION OF THE POSTERIOR DISTRIBUTIONS OF THE STRAIN-SPECIFIC MODEL FOR A GENERIC AREA AND TIME PERIOD

B.1 Notation

For simplicity, assume there are only two co-circulating strain categories (1 and 2) and two disease severity categories (mild and severe):

- Y_1 and Y_2 are number of HFMD disease cases caused by strain 1 and strain 2. $Y_+ = Y_1 + Y_2$ is totally number of HFMD cases. We assume $Y_1 \sim \text{Binomial}(Y_+, q)$, where q is the probability of having strain 1 given being a HFMD case, i.e. $q = Pr(\text{strain 1}|\text{case})$.

- Y_1^s is the number of severe cases from strain 1. $Y_1^s \sim \text{Binomial}(Y_1, p_1^s)$, where p_1^s is the probability of being a severe case given having strain 1, i.e. $p_1^s = Pr(\text{severe}|\text{strain 1})$. Similarly for Y_2^s.

- $Y_+^s = Y_1^s + Y_2^s$ is the total number of severe cases. $Y_+^m = Y_1^m + Y_2^m$ is the total number of mild cases. Note the following relationship: $Y_+ = Y_1 + Y_2 = Y_+^s + Y_+^m$.

- $\boldsymbol{k} = (k^s, k^m)$ where k^s and k^m are the lab tested sample sizes for severe and mild cases respectively.

- The lab-confirmed cases are represented by the letter Z and $\boldsymbol{Z}_1 = (Z_1^s, Z_1^m)$ where Z_1^s and Z_1^m are the lab-confirmed number of severe and mild cases caused by strain 1. Similarly for $\boldsymbol{Z}_2 = (Z_2^s, Z_2^m)$. Because we know the sample size $\boldsymbol{k} = (k^s, k^m)$, $\boldsymbol{Z}_2 = \boldsymbol{k} - \boldsymbol{Z}_1$.

- The parameter of interest is $\boldsymbol{\theta} = (q, p_1^s, p_2^s)$.

B.2 Hierarchical Probability Model

B.2.1 The Likelihood

Following the notation we use for the strain-specific model for a generic area and time period, we have the likelihood for each component as follows:

1. $Y_{1t}|q_t \sim \text{Binomial}(Y_{+t}, q_t)$, q is the parameter of interest and
$$Pr(Y_1|Y_+, q) = \binom{Y_+}{Y_1} q^{Y_1}(1-q)^{(Y_+ - Y_1)}.$$

2. $Y_1^s|p_1^s \sim \text{Binomial}(Y_1, p_1^s)$ and $Y_2^s|p_2^s \sim \text{Binomial}(Y_2, p_2^s)$, p_1^s and p_2^s are the parameters of interest and
$$Pr(Y_1^s|Y_1, p_1^s) = \binom{Y_1}{Y_1^s} {p_1^s}^{Y_1^s}(1-p_1^s)^{(Y_1 - Y_1^s)},$$
$$Pr(Y_2^s|Y_2, p_2^s) = \binom{Y_2}{Y_2^s} {p_2^s}^{Y_2^s}(1-p_2^s)^{(Y_2 - Y_2^s)}.$$

3. $(Z_1^s|k^s, Y_+^s, Y_1^s) \sim \text{Hypergeometric}(k^s, Y_+^s, Y_1^s)$ and $(Z_1^m|k^m, Y_+^m, Y_1^m) \sim \text{Hypergeometric}(k^m, Y_+^m, Y_1^m)$:
$$Pr(Z_1^s|k^s, Y_+^s, Y_1^s) = \frac{\binom{Y_1^s}{Z_1^s}\binom{Y_+^s - Y_1^s}{k^s - Z_1^s}}{\binom{Y_+^s}{k^s}}, \tag{B.1}$$
$$Pr(Z_1^m|k^m, Y_1^m, Y_+^m) = \frac{\binom{Y_1^m}{Z_1^m}\binom{Y_+^m - Y_1^m}{k^m - Z_1^m}}{\binom{Y_+^m}{k^m}}. \tag{B.2}$$

   Because $Y_1^m = Y_1 - Y_1^s$ and $Y_+^m = Y_+ - Y_+^s$, we rewrite $Z_1^m \sim \text{Hypergeometric}(k^m, Y_1^s - Y_1^s, Y_+ - Y_+^s)$
$$Pr(Z_1^m|k^m, Y_1 - Y_1^s, Y_+ - Y_+^s) = \frac{\binom{Y_1 - Y_1^s}{Z_1^m}\binom{Y_+ - Y_+^s - Y_1 + Y_1^s}{k^m - Z_1^m}}{\binom{Y_+ - Y_+^s}{k^m}}.$$

Comibining the previous elements the likelihood function is

$$\begin{aligned}
L(Y_1, \boldsymbol{\theta}|\boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}) &= p(\boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}|Y_1, \boldsymbol{\theta}, Y_+) \\
&= p(\boldsymbol{Z}_1|Y_+^s, Y_+^m, \boldsymbol{k}, Y_1, \boldsymbol{\theta}, Y_+) p(Y_+^s, Y_+^m, \boldsymbol{k}|Y_1, Y_+, \boldsymbol{\theta}) \\
&= p(Z_1^s, Z_1^m|Y_+^s, Y_+^m, \boldsymbol{k}, Y_1, \boldsymbol{\theta}, Y_+) p(Y_+^s, Y_+^m|\boldsymbol{k}, Y_1, Y_+, \boldsymbol{\theta}) p(\boldsymbol{k}|Y_1, Y_+, \boldsymbol{\theta}) \\
&= \sum_{Y_1^s} p(Z_1^s, Z_2^s, Y_1^s|Y_+^s, Y_+^m, \boldsymbol{k}, Y_1) p(Y_+^s, Y_+^m|Y_1, Y_+, \boldsymbol{\theta}) \\
&= \sum_{Y_1^s} p(Z_1^s, Z_1^m|Y_1^s, Y_+^s, Y_+^m, \boldsymbol{k}, Y_1) p(Y_1^s|Y_+^s, m^s, Y_1, \boldsymbol{\theta}) p(Y_+^s, Y_+^m|Y_1, Y_+, \boldsymbol{\theta}) \\
&= \sum_{Y_1^s} p(Z_1^s, Z_1^m|Y_1^s, Y_+^s, Y_+^m, \boldsymbol{k}, Y_1) p(Y_1^s|Y_+^s, Y_1, \boldsymbol{\theta}) p(Y_+^s, Y_+^m|Y_1, Y_+, \boldsymbol{\theta}) \\
&= \sum_{Y_1^s} p(Z_1^s, Z_1^m|Y_1^s, Y_+^s, Y_+^m, \boldsymbol{k}, Y_1) p(Y_1^s|Y_+^s, Y_1, \boldsymbol{\theta}) p(Y_+^s|Y_1, Y_+, \boldsymbol{\theta}) \\
&= \sum_{Y_1^s} p(Z_1^s, Z_1^m|Y_1^s, Y_+^s, Y_+^m, \boldsymbol{k}, Y_1) p(Y_1^s, Y_+^s|Y_1, Y_+, \boldsymbol{\theta})
\end{aligned}$$

Notice that, conditionally, Z_1^s and Z_1^m follow two independent Hypergeometric distributions,

$$p(Z_1^s, Z_1^m|Y_1^s, Y_+^s, Y_+^m, \boldsymbol{k}, Y_1, Y_+) = \frac{\binom{Y_1^s}{Z_1^s}\binom{Y_+^s - Y_1^s}{k^s - Z_1^s}}{\binom{Y_+^s}{k^s}} \frac{\binom{Y_1 - Y_1^s}{Z_1^m}\binom{Y_+ - Y_1 - Y_+^s + Y_1^s}{k^m - Z_1^m}}{\binom{Y_+^m}{k^m}},$$

where $Y_1^s \in R_1^* = \big(\max(Z_1^s, Y_1 - Y_+^m + Z_2^m), \ldots, \min(Y_+^s - k^s + Z_1^s, Y_1 - Z_1^m)\big)$ as given in (B.1) and (B.2).

To obtain $p(Y_1^s, Y_+^s|Y_1, Y_+, \boldsymbol{\theta})$ we have

$$p(Y_1^s, Y_+^s|Y_1, Y_+, \boldsymbol{\theta}) = \binom{Y_1}{Y_1^s}(p_1^s)^{Y_1^s}(1-p_1^s)^{(Y_1 - Y_1^s)}\binom{Y_+ - Y_1}{Y_+^s - Y_1^s}(p_2^s)^{Y_+^s - Y_1^s}(1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)},$$

where $Y_1^s \in (0, \min(Y_1^s, Y_+^s))$.

To obtain the range of Y_1, notice that on the one hand $Y_1 \geq Y_1^s + Z_1^m$ and on the other hand, we know that $Y_2 \geq Y_2^s + Z_2^s + Z_2^m$. Therefore:

$$\begin{aligned}
Y_+ - Y_1 &\geq Y_2^s + Z_2^s + Z_2^m \\
Y_1 &\leq Y_+ - (Y_2^s + Z_2^s + Z_2^m) \\
Y_1 &\leq Y_+ - (Y_+^s - Y_1^s + Z_2^s + Z_2^m)
\end{aligned}$$

Combining the two conditions we have

$$Y_1 \in (Y_1^s + Z_1^m, Y_+ - Y_+^s + Y_1^s - Z_2^s - Z_2^m)$$

We now consider the range of Y_1^s. From $Z_1^s \sim \text{Hypergeometric}(k^s, Y_+^s, Y_1^s)$, we have the range

$$Y_1^s \in (Z_1^s, \ldots, Y_+^s - (k^s - Z_1^s)).$$

From $Z_1^m \sim \text{Hypergeometric}(k^m, Y_+^m, Y_1^m)$, where $Y_1^m \in (Z_1^m, \ldots, Y_+^m - (k^m - Z_1^m))$ and $Y_1^s = Y_1 - Y_1^m$, we obtain

$$Y_1^s \in (Y_1 - Y_+^m + (m^m - Z_1^m), \cdots, Y_1 - Z_1^m),$$

and therefore

$$Y_1^s \in (Y_1 - Y_+^m + Z_2^m, \ldots, Y_1 - Z_1^m).$$

Combing these two constraints we have

$$Y_1^s \in \left(\max(Z_1^s, Y_1 - Y_+^m + Z_2^m), \ldots, \min(Y_+^s - (m^s - Z_1^s), Y_1 - Z_1^m)\right)$$

B.2.2 Posterior Distributions

The joint posterior distribution is

$$\begin{aligned}
p(Y_1, \boldsymbol{\theta} | \boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+) &= \frac{p(\boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k} | Y_1, \boldsymbol{\theta}, Y_+) p(Y_1, \boldsymbol{\theta} | Y_+)}{p(\boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k} | y_+)} \\
&\propto p(Y_1, \boldsymbol{\theta} | \boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+) p(Y_1, \boldsymbol{\theta} | Y_+) \\
&\propto p(Y_1, \boldsymbol{\theta} | \boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+) p(Y_1 | Y_+, q) \pi(\boldsymbol{\theta}) \\
&\propto p(Y_1, \boldsymbol{\theta} | \boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+) p(Y_1 | Y_+, q) \pi(\boldsymbol{\theta} | Y_+) \\
&\propto p(Y_1, \boldsymbol{\theta} | \boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+) p(Y_1 | Y_+, q) \pi(p_1^s) \pi(p_2^s) \pi(q)
\end{aligned}$$

Introducing the auxiliary variable Y_1^s we have

$$
\begin{aligned}
& p(Y_1, \boldsymbol{\theta} | \boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+) \\
= & \sum_{Y_1^s} p(Z_1^s, Z_2^s | Y_1^s, Y_+^s, Y_+^m, \boldsymbol{k}, Y_1) p(Y_1^s, Y_+^s | Y_1, Y_+, \boldsymbol{\theta}) p(Y_1 | Y_+, q) \pi(p_1^s) \pi(p_2^s) \pi(q) \\
= & \sum_{Y_1^s} \Big(\frac{\binom{Y_1^s}{Z_1^s}\binom{Y_+^s - Y_1^s}{k^s - Z_1^s}}{\binom{Y_+^s}{k^s}} \frac{\binom{Y_1 - Y_1^s}{Z_1^m}\binom{Y_+ - Y_1 - Y_+^s + Y_1^s}{k^m - Z_1^m}}{\binom{Y_+^m}{k^m}} \binom{Y_1}{Y_1^s} (p_1^s)^{Y_1^s} (1-p_1^s)^{(Y_1 - Y_1^s)} \times \\
& \binom{Y_+ - Y_1}{Y_+^s - Y_1^s} (p_2^s)^{Y_+^s - Y_1^s} (1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)} \Big) \binom{Y_+}{Y_1} q^{Y_1} (1-q)^{Y_+ - Y_1} \pi(p_1^s) \pi(p_2^s) \pi(q) \\
\propto & \sum_{Y_1^s} \left(\binom{Y_1^s}{Z_1^s} \binom{Y_+^s - Y_1^s}{k^s - Z_1^s} \binom{Y_+ - Y_1 - Y_+^s + Y_1^s}{k^m - Z_1^m} \binom{Y_1 - Y_1^s}{Z_1^m} \times \right. \\
& \left. \binom{Y_1}{Y_1^s} (p_1^s)^{Y_1^s} (1-p_1^s)^{(Y_1 - Y_1^s)} \binom{Y_+ - Y_1}{Y_+^s - Y_1^s} (p_2^s)^{Y_+^s - Y_1^s} (1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)} \right) \times \\
& \binom{Y_+}{Y_1} q^{Y_1} (1-q)^{Y_+ - Y_1} \pi(p_1^s) \pi(p_2^s) \pi(q)
\end{aligned}
$$

Let $\pi(q) = \text{Beta}(\alpha, \beta)$, $\pi(p_1^s) = \text{Beta}(\alpha_1, \beta_1)$ and $\pi(p_2^s) = \text{Beta}(\alpha_2, \beta_2)$, then the conditional posterior distributions required for an MCMC algorithm are

$$
\begin{aligned}
p(q | \boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+, Y_1, p_1^s, p_2^s) & \propto q^{Y_1} (1-q)^{Y_+ - Y_1} q^{\alpha - 1} (1-q)^{\beta_1} \\
& = \text{Beta}(Y_1 + \alpha, Y_+ - Y_1 + \beta)
\end{aligned}
$$

$$
\begin{aligned}
p(p_1^s | \boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+, Y_1, p_2^s, q) & \propto \left(\binom{Y_1^s}{Z_1^s} \binom{Y_+^s - Y_1^s}{k^s - Z_1^s} \binom{Y_+ - Y_1 - Y_+^s + Y_1^s}{k^m - Z_1^m} \binom{Y_1 - Y_1^s}{Z_1^m} \times \right. \\
& \left. \binom{Y_1}{Y_1^s} (p_1^s)^{Y_1^s} (1-p_1^s)^{(Y_1 - Y_1^s)} \times \right. \\
& \left. \binom{Y_+ - Y_1}{Y_+^s - Y_1^s} (p_2^s)^{Y_+^s - Y_1^s} (1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)} \right) \pi(p_1^s) \\
& \propto (p_1^s)^{Y_1^s} (1-p_1^s)^{(Y_1 - Y_1^s)} \pi(p_1^s) \\
& = \text{Beta}(Y_1^s + \alpha_1, Y_1 - Y_1^s + \beta_1)
\end{aligned}
$$

$$
\begin{aligned}
p(p_2^s|\boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+, Y_1, p_1^s, q) &\propto \left(\binom{Y_1^s}{Z_1^s}\binom{Y_+^s - Y_1^s}{k^s - Z_1^s}\binom{Y_+ - Y_1 - Y_+^s + Y_1^s}{k^m - Z_1^m}\binom{Y_1 - Y_1^s}{Z_1^m} \times \right. \\
&\quad \binom{Y_1}{Y_1^s}(p_1^s)^{Y_1^s}(1-p_1^s)^{(Y_1 - Y_1^s)} \times \\
&\quad \left. \binom{Y_+ - Y_1}{Y_+^s - Y_1^s}(p_2^s)^{Y_+^s - Y_1^s}(1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)} \right) \pi(p_2^s) \\
&\propto (p_2^s)^{Y_+^s - Y_1^s}(1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)}\pi(p_2^s) \\
&= \text{Beta}(Y_+^s - Y_1^s + \alpha_2, Y_+ - Y_1 - Y_+^s + Y_1^s + \beta_2)
\end{aligned}
$$

$$
\begin{aligned}
p(Y_1|\boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+, p_1^s, p_2^s, q) &\propto \left(\binom{Y_+ - Y_1 - Y_+^s + Y_1^s}{k^m - Z_1^m}\binom{Y_1 - Y_1^s}{Z_1^m} \times \right. \\
&\quad \binom{Y_1}{Y_1^s}(p_1^s)^{Y_1^s}(1-p_1^s)^{(Y_1 - Y_1^s)} \times \\
&\quad \left. \binom{Y_+ - Y_1}{Y_+^s - Y_1^s}(p_2^s)^{Y_+^s - Y_1^s}(1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)} \right) \\
&\quad \binom{Y_+}{Y_1} q^{Y_1}(1-q)^{Y_+ - Y_1} \\
\text{The range of } Y_1 \text{ is:} \qquad Y_1 &\in (Y_1^s + Z_1^m, Y_+ - Y_+^s + Y_1^s - Z_2^m)
\end{aligned}
$$

For the auxiliary variable Y_1^s :

$$
\begin{aligned}
p(Y_1^s|\boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+, p_1^s, p_2^s, q, Y_1) &\propto \binom{Y_1^s}{Z_1^s}\binom{Y_+^s - Y_1^s}{k^s - Z_1^s}\binom{Y_+ - Y_1 - Y_+^s + Y_1^s}{k^m - Z_1^m}\binom{Y_1 - Y_1^s}{Z_1^m} \times \\
&\quad \binom{Y_1}{Y_1^s}(p_1^s)^{Y_1^s}(1-p_1^s)^{(Y_1 - Y_1^s)} \times \\
&\quad \binom{Y_+ - Y_1}{Y_+^s - Y_1^s}(p_2^s)^{Y_+^s - Y_1^s}(1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)} \\
\text{The range of } Y_1^s \text{ is:} \qquad Y_1^s &\in \left(\max(Z_1^s, Y_1 - Y_+^m + Z_2^m), ..., \min(Y_+^s - (m^s - Z_1^s), Y_1 - Z_1^m)\right)
\end{aligned}
$$

Appendix C

MARKOV CHAIN MONTE CARLO FOR DISCRETE VARIABLES

In this section, we describe two approaches for making inference for discrete variables using the MCMC method.

C.1 A Full Enumerate Method

The full enumeration method proceeds as follows:

1. Set the initial value of $(Y_1)^0, (Y_1^s)^0, q^0, (p_1^s)^0$ and $(p_2^s)^0$. For example, we can set $(Y_1)^0 = Z_1^s + Z_1^m$ as $Y_1 \in (Z_1^s + Z_1^m, Y_+^s + Y_+^m - Z_2^s + Z_2^m)$, $q^0 = p_1^{s0} = p_2^{s0} = 0.5$ and $(Y_1^s)^0 = (Y_1)^0 \times (p_1^s)^0$.

2. Calculate the cumulative distributions of Y_1^s based on these initial values and on the unnormalized conditional distribution of Y_1^s, from Appendix B.

3. Update Y_1^s from its conditional distribution. This is done by first generating a value $u_1 \sim \text{Unif}(0,1)$ and finding

$$(Y_1^s)^1 = F^{-1}(u_1) = \inf_{x \in \mathbb{R}} \{F(x) \geq u_1\}.$$

4. Calculate the cumulative distributions of Y_1 based on $(Y_1^s)^1, q^0, p_1^{s0}$ and p_2^{s0}, and on the unnormalized conditional distribution of Y_1, from Appendix B.

5. Update Y_1 from its conditional distribution. This is done by first generating a value $u_2 \sim \text{Unif}(0,1)$ and finding

$$(Y_1)^1 = F^{-1}(u_2) = \inf_{x \in \mathbb{R}} \{F(x) \geq u_2\}.$$

6. Update q, p_1^s and p_2^s.

7. Repeat steps 2 to 6 until we obtain the desired number of iterations.

The full numeration method is computationally expensive because we need to compute the cumulative distributions of Y_1 and Y_1^s based on their ranges for each MCMC iteration. An alternative method for sampling for discrete variables is presented in Wakefield et al. (2011), and we describe briefly the implemention of this method for our proposed model in the next section.

C.2 A Metropolis-Hastings Algorithm for Sampling Discrete Variables

Consider the updating step for variable Y_1 in MCMC iterations, the posterior conditional distribution of Y_1 is proportional to

$$\begin{aligned} p(Y_1|\boldsymbol{Z}_1, Y_+^s, Y_+^m, \boldsymbol{k}, Y_+, p_1^s, p_2^s, q) \propto & \left(\binom{Y_+ - Y_1 - Y_+^s + Y_1^s}{k^m - Z_1^m} \times \right. \\ & \binom{Y_1 - Y_1^s}{Z_1^m} \binom{Y_1}{Y_1^s} (p_1^s)^{Y_1^s} (1-p_1^s)^{(Y_1 - Y_1^s)} \times \\ & \left. \binom{Y_+ - Y_1}{Y_+^s - Y_1^s} (p_2^s)^{Y_+^s - Y_1^s} (1-p_2^s)^{(Y_+ - Y_1 - Y_+^s + Y_1^s)} \right) \\ & \binom{Y_+}{Y_1} q^{Y_1} (1-q)^{Y_+ - Y_1} \end{aligned} \tag{C.1}$$

with the range of $Y_1 \in (Y_1^s + Z_1^m, Y_+ - Y_+^s + Y_1^s - Z_2^m)$.

Rather than evaluating the normalizing constant of the conditional distribution, we use a Metropolis-Hastings step. Let y_1' denote the current value of y_1, we first generate a proposal of $y_1^* = y_1 \pm d$ where d is drawn uniformly from $(1, 2, \ldots, D)$, for $D > 0$. We then check if the proposed value is valid with respect to its range. If it is, then we proceed to calculate the acceptance/rejection ratio and decide if we accept the proposed value. This is similar to the usual Metropolis-Hastings algorithm. Otherwise, the MCMC chain remains at the current value. The value of D is set to achieve an acceptance rate that is around 30% to 40% (Roberts et al., 1997).

Compared to the full enumeration method, this modified Metropolis-Hastings algorithm is much faster, especially when Y_1 has a broad range. For the analysis of the China HFMD data, we take this modified approach when we carry out the MCMC iterations.

Appendix D

DERIVATION OF THE POSTERIOR DISTRIBUTIONS OF THE STRAIN-SPECIFIC MODEL, WITH TEMPORAL TREND

D.1 Notation

For a specific area, we let t index time, $t = 1, \ldots, T$ where T is the number of equally-spaced time points. The notation we use for including the temporal component in the model is as follows:

- Y_{1t} and Y_{2t} are number of HFMD disease cases caused by strain 1 and strain 2 at time t. $Y_{+t} = Y_{1t} + Y_{2t}$ is totally number of HFMD cases at time t. We assume $Y_{1t} \sim \text{Binomial}(Y_{+t}, q_t)$, where q_t is the probability of having strain 1 given being a HFMD case at time t, i.e. $q_t = Pr(\text{strain 1}|\text{case at time t})$.

- Y_{1t}^s is the number of severe cases from strain 1 at time t. $Y_{1t}^s \sim \text{Binomial}(Y_{1t}, p_{1t}^s)$, where p_{1t}^s is the probability of being a severe case given having strain 1 at time t, i.e. $p_{1t}^s = Pr(\text{severe}|\text{strain 1 at time t})$. Similarly for Y_{2t}^s.

- $Y_{+t}^s = Y_{1t}^s + Y_{2t}^s$ is the total number of severe cases at time t. $Y_{+t}^m = Y_{1t}^m + Y_{2t}^m$ is the total number of mild cases at time t. Note the following relationship: $Y_{+t} = Y_{1t} + Y_{2t} = Y_{+t}^s + Y_{+t}^m$.

- $\boldsymbol{k}_t = (k_t^s, k_t^m)$ where k_t^s and k_t^m are the lab tested sample sizes for severe and mild cases respectively at time t.

- The lab-confirmed cases are represented by the letter Z and $\boldsymbol{Z}_{1t} = (Z_{1t}^s, Z_{1t}^m)$ where Z_{1t}^s and Z_{1t}^m are the lab-confirmed number of severe and mild cases caused by strain 1 at time t. Similarly for $\boldsymbol{Z}_{2t} = (Z_{2t}^s, Z_{2t}^m)$. Because we know the sample sizes $\boldsymbol{k}_t = (k_t^s, k_t^m)$, $\boldsymbol{Z}_2 = \boldsymbol{k}_t - \boldsymbol{Z}_1$.

- Let $\boldsymbol{\theta}_t = (q_t, p^s_{1t}, p^s_{2t}), t = 1, ..., T$.

D.2 Hierarchical probability model

D.2.1 The Likelihood

We assume the following hierarchical structure for the data-generating mechanism at time point t:

1. $Y_{1t}|q_t \sim \text{Binomial}(Y_{+t}, q_t)$:

$$Pr(Y_{1t}|Y_{+t}, q) = \binom{Y_{+t}}{Y_{1t}} {q_t}^{Y_{1t}} (1 - q_t)^{(Y_{+t} - Y_{1t})}.$$

2. $Y^s_{1t}|p^s_{1t} \sim \text{Binomial}(Y_{1t}, p^s_{1t})$ and $Y^s_{2t}|p^s_{2t} \sim \text{Binomial}(Y_{2t}, p^s_{2t})$:

$$Pr(Y^s_{1t}|Y_{1t}, p^s_{1t}) = \binom{Y_{1t}}{Y^s_{1t}} {p^s_t}^{Y^s_{1t}} (1 - p^s_{1t})^{(Y_{1t} - Y^s_{1t})}.$$

$$Pr(Y^s_{2t}|Y_{2t}, p^s_{2t}) = \binom{Y_{2t}}{Y^s_{2t}} {p^s_t}^{Y^s_{2t}} (1 - p^s_{2t})^{(Y_{2t} - Y^s_{2t})}.$$

3. $(Z^s_{1t}|k^s_t, Y^s_{1t}, Y^s_{+t}) \sim \text{Hypergeometric}(k^s_t, Y^s_{1t}, Y^s_{+t})$:

$$Pr(Z^s_{1t}|k^s_t, Y^s_{1t}, Y^s_{+t}) = \frac{\binom{Y^s_{1t}}{Z^s_{1t}} \binom{Y^s_{+t} - Y^s_{1t}}{k^s_t - Z^s_{1t}}}{\binom{Y^s_{+t}}{k^s_t}}.$$

$(Z^m_{1t}|k^m_t, Y^m_{1t}, Y^m_{+t}) \sim \text{Hypergeometric}(k^m_t, Y^m_{1t}, Y^m_{+t})$,

$$Pr(Z^m_{1t}|k^m_t, Y^m_{1t}, Y^m_{+t}) = \frac{\binom{Y^m_{1t}}{Z^m_{1t}} \binom{Y^m_{+t} - Y^m_{1t}}{k^m_t - Z^m_{1t}}}{\binom{Y^m_{+t}}{k^m_t}}.$$

Because $Y^m_{1t} = Y_{1t} - Y^s_{1t}$ and $Y^m_{+t} = Y_{+t} - Y^s_{+t}$, we rewrite $(Z^m_{1t}|k^m_t, Y^s_{1t} - Y^s_{1t}, Y_{+t} - Y^s_{+t}) \sim$ $\text{Hypergeometric}(k^m_t, Y^s_{1t} - Y^s_{1t}, Y_{+t} - Y^s_{+t})$,

$$Pr(Z^m_{1t}|k^m_t, Y_{1t} - Y^s_{1t}, Y_{+t} - Y^s_{+t}) = \frac{\binom{Y_{1t} - Y^s_{1t}}{Z^m_{1t}} \binom{Y_{+t} - Y^s_{+t} - Y_{1t} + Y^s_{1t}}{k^m_t - Z^m_{1t}}}{\binom{Y_{+t} - Y^s_{+t}}{k^m_t}}$$

Combining the previous elements we have

$$L(Y_{1t}, \boldsymbol{\theta}_t | Z^s_{1t}, Z^m_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{+t}) \;=\; p(Z^s_{1t}, Z^m_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t} | Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t)$$

Introducing the auxiliary variable Y^s_{1t} we have

$$\begin{aligned}
& L(Y_{1t}, \boldsymbol{\theta}_t | Z^s_{1t}, Z^m_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{+t}) \\
= \;& \sum_{Y^s_{1t}} p(Z^s_{1t}, Z^m_{1t}, Y^s_{1t} | \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) p(Y^s_{+t}, Y^m_{+t} | Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) \\
= \;& \sum_{Y^s_{1t}} p(Z^s_{1t}, Z^m_{1t} | Y^s_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) p(Y^s_{1t} | \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) \times \\
& p(Y^s_{+t}, Y^m_{+t} | Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) \\
= \;& \sum_{Y^s_{1t}} p(Z^s_{1t}, Z^m_{1t} | Y^s_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{1t}) p(Y^s_{1t} | Y^s_{+t}, Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) p(Y^s_{+t}, Y^m_{+t} | Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) \\
= \;& \sum_{Y^s_{1t}} p(Z^s_{1t}, Z^m_{1t} | Y^s_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{1t}) p(Y^s_{1t} | Y^s_{+t}, Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) p(Y^s_{+t} | Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) \\
= \;& \sum_{Y^s_{1t}} p(Z^s_{1t}, Z^m_{1t} | Y^s_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{1t}) p(Y^s_{1t}, Y^s_{+t} | Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t)
\end{aligned}$$

where,

$$p(Z^s_{1t}, Z^m_{1t} | Y^s_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t}, Y_{1t}) = \frac{\binom{Y^m_{1t}}{Z^m_{1t}} \binom{Y^m_{+t} - Y^m_{1t}}{k^m_t - Z^m_{1t}}}{\binom{Y^m_{+t}}{k^m_t}} \frac{\binom{Y_{1t} - Y^s_{1t}}{Z^m_{1t}} \binom{Y_{+t} - Y^s_{+t} - Y_{1t} + Y^s_{1t}}{k^m_t - Z^m_{1t}}}{\binom{Y_{+t} - Y^s_{+t}}{k^m_t}},$$

where $Y^s_{1t} \in \big(\max(Z^s_{1t}, Y_{1t} - Y^m_{+t} + Z^m_{2t}), \ldots, \min(Y^s_{+t} - k^s_t + Z^s_{1t}, Y_{1t} - Z^m_{1t})\big)$.

Also,

$$p(Y^s_{1t}, Y^s_{+t} | Y_{1t}, Y_{+t}, \boldsymbol{\theta}_t) = \binom{Y_{1t}}{Y^s_{1t}} (p^s_{1t})^{Y^s_{1t}} (1 - p^s_{1t})^{(Y_{1t} - Y^s_{1t})} \binom{Y_{+t} - Y_{1t}}{Y^s_{+t} - Y^s_{1t}} (p^s_{2t})^{Y^s_{+t} - Y^s_{1t}} (1 - p^s_{2t})^{(Y_{+t} - Y_{1t} - Y^s_{+t} + Y^s_{1t})},$$

where $Y_{1t} \in (Y^s_{1t} + Z^m_{1t}, Y_{+t} - Y^s_{+t} + Y^s_{1t} - k^m_t + Z^m_{1t})$.

D.2.2 Posterior

Let $\boldsymbol{Y}_1 = (Y_{11}, Y_{12}, \ldots, Y_{1t})$, $\boldsymbol{Z}^s_1 = (Z^s_{11}, Z^s_{12}, \ldots, Z^s_{1t})$, $\boldsymbol{Z}^m_1 = (Z^m_{11}, Z^m_{12}, \ldots, Z^m_{1t})$, $\boldsymbol{k}^s = (k^s_{11}, k^s_{12}, \ldots, k^s_{1t})$, $\boldsymbol{k}^m = (k^m_{11}, k^m_{12}, \ldots, k^m_{1t})$, $\boldsymbol{Y}^s_+ = (Y^s_{+1}, Y^s_{+2}, \ldots, Y^s_{+t})$, $\boldsymbol{Y}^m_+ =$

$(Y^m_{+1}, Y^m_{+2}, \ldots, Y^m_{+t})$ and $\boldsymbol{Y}_+ = (Y_{+1}, Y_{+2}, \ldots, Y_{+t})$. The parameters of interest are $\boldsymbol{q}_t = (q_1, q_2, \ldots, q_t)$, $\boldsymbol{p^s_1} = (p^s_{11}, p^s_{12}, \ldots, p^s_{1t})$ and $\boldsymbol{p^s_2} = (p^s_{21}, p^s_{22}, \ldots, p^s_{2t})$.

The joint posterior distribution is:

$$
\begin{aligned}
& p(\boldsymbol{Y}_1, \boldsymbol{q}, \boldsymbol{p^s_1}, \boldsymbol{p^s_2} | \boldsymbol{Z}^s_1, \boldsymbol{Z}^m_1, \boldsymbol{k}^s, \boldsymbol{k}^m, \boldsymbol{Y}^s_+, \boldsymbol{Y}^m_+, \boldsymbol{Y}_+) \\
= & \frac{p(\boldsymbol{Z}^s_1, \boldsymbol{Z}^m_1, \boldsymbol{k}^s, \boldsymbol{k}^m, \boldsymbol{Y}^s_+, \boldsymbol{Y}^m_+ | \boldsymbol{Y}_1, \boldsymbol{q}, \boldsymbol{p^s_1}, \boldsymbol{p^s_2}, \boldsymbol{Y}_+) \pi(\boldsymbol{Y}_1, \boldsymbol{q}, \boldsymbol{p^s_1}, \boldsymbol{p^s_2} | \boldsymbol{Y}_+)}{p(\boldsymbol{Z}^s_1, \boldsymbol{Z}^m_1, \boldsymbol{k}^s, \boldsymbol{k}^m, \boldsymbol{Y}^s_+, \boldsymbol{Y}^m_+ | \boldsymbol{Y}_+)} \\
\propto & p(\boldsymbol{Z}^s_1, \boldsymbol{Z}^m_1, \boldsymbol{k}^s, \boldsymbol{k}^m, \boldsymbol{Y}^s_+, \boldsymbol{Y}^m_+ | \boldsymbol{Y}_1, \boldsymbol{q}, \boldsymbol{p^s_1}, \boldsymbol{p^s_2}, \boldsymbol{Y}_+) \pi(\boldsymbol{Y}_1, \boldsymbol{q}, \boldsymbol{p^s_1}, \boldsymbol{p^s_2} | \boldsymbol{Y}_+) \\
\propto & p(\boldsymbol{Z}^s_1, \boldsymbol{Z}^m_1, \boldsymbol{k}^s, \boldsymbol{k}^m, \boldsymbol{Y}^s_+, \boldsymbol{Y}^m_+ | \boldsymbol{Y}_1, \boldsymbol{q}, \boldsymbol{p^s_1}, \boldsymbol{p^s_2}, \boldsymbol{Y}_+) \pi(\boldsymbol{Y}_1 | \boldsymbol{q}_t, \boldsymbol{Y}_+) \pi(\boldsymbol{q}, \boldsymbol{p^s_1}, \boldsymbol{p^s_2} | \boldsymbol{Y}_+) \\
& \text{assuming independence of } \boldsymbol{p^s_1} \text{ and } \boldsymbol{p^s_2} \\
\propto & \prod_{t=1}^{T} \left[p(Z^s_{1t}, Z^m_{1t}, \boldsymbol{k}_t, Y^s_{+t}, Y^m_{+t} | Y_{1t}, q_t, Y_{+t}) \pi(Y_{1t} | q_t, Y_{+t}) \right] \pi(\boldsymbol{p^s_1}) \pi(\boldsymbol{p^s_2}) \pi(\boldsymbol{q}) \\
& \text{introducing the auxiliary variable } Y^s_{1t} \\
\propto & \prod_{t=1}^{T} \left[\sum_{Y^s_1} \binom{Y^s_{1t}}{Z^s_{1t}} \binom{Y^s_{+t} - Y^s_{1t}}{k^s_t - Z^s_{1t}} \binom{Y_{+t} - Y_{1t} - Y^s_{+t} + Y^s_{1t}}{k^m_t - Z^m_{1t}} \binom{Y_{1t} - Y^s_{1t}}{Z^m_{1t}} \times \right. \\
& \binom{Y_{1t}}{Y^s_{1t}} (p^s_{1t})^{Y^s_{1t}} (1 - p^s_{1t})^{(Y_{1t} - Y^s_{1t})} \binom{Y_{+t} - Y_{1t}}{Y^s_{+t} - Y^s_{1t}} (p^s_{2t})^{Y^s_{+t} - Y^s_{1t}} (1 - p^s_{2t})^{(Y_{+t} - Y_{1t} - Y^s_{+t} + Y^s_{1t})} \times \\
& \left. \binom{Y_{+t}}{Y_{1t}} {q_t}^{Y_{1t}} (1 - q_t)^{Y_{+t} - Y_{1t}} \right] \pi(\boldsymbol{p^s_1}) \pi(\boldsymbol{p^s_2}) \pi(\boldsymbol{q})
\end{aligned}
$$

Let $\pi(q_t) = \text{Beta}(1, 1)$ and

$$
\begin{aligned}
\text{logit}(p^s_{1t}) &= \alpha_1 + \gamma_{1t} \\
\text{logit}(p^s_{2t}) &= \alpha_2 + \gamma_{2t},
\end{aligned}
$$

where $\gamma_{1t} \sim \text{N}(0, \tau^{-1}_{\gamma_1})$ and $\gamma_{2t} \sim \text{N}(0, \tau^{-1}_{\gamma_2})$. We assign flat priors to the intercepts α_1 and α_2, and Gamma$(1, 0.026)$ priors to the precision parameters $\tau^{-1}_{\gamma_1}$ and $\tau^{-1}_{\gamma_2}$.

The conditional posterior distributions for α_1 is

$$
p(\alpha_1 | \gamma_{1t}, t, Y_{1t}, Y^s_{1t}) \propto \prod_{t=1}^{T} \left[\left(\frac{\exp(\alpha_1 + \gamma_{1t})}{1 + \exp(\alpha_1 + \gamma_{1t})} \right)^{Y^s_{1t}} \left(\frac{1}{1 + \exp(\alpha_1 + \gamma_{1t})} \right)^{(Y_{1t} - Y^s_{1t})} \right],
$$

and similarly for α_2.

The conditional posterior distributions for γ_{1t} is

$$
p(\gamma_{1t} | \alpha_1, Y_{1t}, Y^s_{1t}) \propto \left[\left(\frac{\exp(\alpha_1 + \gamma_{1t})}{1 + \exp(\alpha_1 + \gamma_{1t})} \right)^{Y^s_{1t}} \left(\frac{1}{1 + \exp(\alpha_1 + \gamma_{1t})} \right)^{(Y_{1t} - Y^s_{1t})} \right] \exp\left(-\frac{\tau_{\gamma_1}}{2} \gamma^2_{1t} \right),
$$

and similarly for γ_{2t}.

The conditional posterior distributions for τ_{γ_1} and τ_{γ_2} are Gamma$(1 + T/2, 0.026 + \sum_t \gamma_{1t}^s/2)$ and Gamma$(1 + T/2, 0.026 + \sum_t \gamma_{2t}^s/2)$, respectively.

Given the values of α_1, α_2, γ_{1t} and γ_{2t}, we can simultaneously obtain p_{1t}^s and p_{2t}^s. Together with q_t, the conditional distribution of Y_{1t} and Y_{1t}^s remain the same as in the model for a generic area and time period in Appendix B.

VITA

Cici Xi Chen Bauer was born Xi Chen in Hefei, China. In 2003, she earned her B.S. in Statistics from Anhui University. She then went to graduate school in Alaska and earned her M.S. in Statistics from the University of Alaska, Fairbanks in 2005. She worked as a Biometrician in Alaska Department of Fish and Game between 2005 and 2007. In 2012, she earned her Ph.D. in Statistics from the University of Washington, Seattle under the supervision of Prof. Jon Wakefield.