Fall 2020

Incorporation and Measurement of Uncertainty in Clustered and Spatial Data

Yuan Hong

Incorporation and Measurement of Uncertainty in Clustered and Spatial Data

by

Yuan Hong

Master of Public Health
Drexel University, 2014

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Biostatistics

School of Public Health

University of South Carolina

2020

Accepted by:

Alexander C McLain, Major Professor

Bo Cai, Committee Member

Edward A Frongillo, Committee Member

Feifei Xiao, Committee Member

Jan M Eberth, Committee Member

Cheryl Addy, Dean of the Graduate School

Acknowledgments

I want to express my appreciation to my committee members, Drs. Alexander C McLain, Bo Cai, Feifei Xiao, Edward A Frongillo and Jan M. Eberth, for agreeing to serve on my committee and all the effort and time they put in my research. I wish to give a special thank my committee chair Dr. Alexander C McLain who has provided me his support, guidance and encouragement for the entire study.

I'm very grateful to Dr. Suzanne McDermott for giving me the opportunity to work for her and participating in many interesting projects where I was able to apply my statistical knowledge. I also want to thank Deborah Salzberg Clark for all her help during my stay.

I would like to mention my dear colleagues and classmates Yanan Zhang, Xinling (Claire) Xu, Xizhi (Adam) Luo and I feel very lucky to have them around to support me, listen to me.

Finally, I would like to thank my parents and family members back in China to support me for pursuing my PhD no matter what.

Abstract

Analyzing population representative datasets for local estimation and predictions over time is important for monitoring related public health issues, however, there are many statistical challenges associated with such analyses. Mixed effect models are one of the common options which can incorporate time and spatial effect in the model and related inference is well established.

In the first part of this dissertation, to estimate area-level prevalence using individual-level data, small area estimation (SAE) with post-stratified mixed effect models were used where sampling weights were also incorporated into it. However, if post-stratification which requires more computation effort can improve estimation accuracy is not clear given the complicated modelling framework. Therefore, comparing the mean squared prediction errors (MSPE) to evaluate the predictive ability of post-stratification is of interest. In this study, various bootstrap methods were also implemented to calculate confidence intervals for post-stratified estimates, and investigating and comparing the performances of different bootstrap methods is another aim of this study. Under different model complexity situations, we are able to identify the best-performed bootstrap methods in the simulation study.

The second part of the dissertation involves analyses and predictions of disease prevalence using a penalized B-spline model. A unique feature of the data is that the sampling standard errors (SSEs) coming with the prevalence estimates need to be incorporated into the model. In previous studies, the uncertainty of the SSE is ignored which could influence the reliability of the estimation. In this study, we incorporate the uncertainty of the SSE and proposed an approximated likelihood function for fast

computation. The performances of the proposed method were compared with some standard approaches in a simulation study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to longitudinal data and cluster data

Fitzmaurice et al. [15] defined a longitudinal study to be data where multiple measurements on the same subjects are taken repeatedly over time. Therefore, one of the distinct features of the longitudinal data is that they have a temporal order. Researchers can track the changes in responses over time and identify risk factors that have influences on the longitudinal trajectory of the outcome. Also, because the measurements were taken from the same subjects repeatedly, within-subject changes over time can be captured and also be a point of interest to researchers. However, the repeated measures indicate that the data points are clustered which brings in statistical challenges due to the violation of the standard independence assumption made in linear regression models. A huge amount of statistical research has focused on developing models that can accommodate clustered data including longitudinal data. In the scope of clustered data analyses, there are two important aspects including point estimation and corresponding inference.

The two-staged model proposed by Laird and Ware [24] is one of the models used for longitudinal analyses. This model assumes that all individuals follow the same distribution for multiple measurements, which is the first stage model. A random effect is a parameter that is allowed to vary over individuals. For each individual i in stage 1, $y_i = X_i\alpha + Z_i b_i + e_i$, where $e_i \sim N(0, R_i)$ and R_i is $n_i \times n_i$ positive-definite covariance matrix for individual i. At the first stage, α and b_i are considered fixed

for each individual i. In stage 2, $b_i \sim N(0, D)$ is a k-dimensional random effect and D is a $k \times k$ positive-definite matrix. Marginally $y_i \sim N(X_i\alpha, R_i + Z_i D Z_i^T)$, while conditionally $y_i|b_i \sim N(X\alpha + Z_i b_i, R_i)$.

Laird and Ware [24] proposed a unified approach to inference using two-staged models. For known variance matrices R_i and D, $Var(y_i) = V_i = R_i + Z_i D Z_i^T$, they proposed to estimate the parametesr via

$$\hat{\alpha} = \left(\sum_1^m X_i^T V_i^{-1} X_i \right)^{-1} \sum_1^m X_i^T V_i^{-1} y_i$$

and predict the random effects with

$$\hat{b}_i = D Z_i^T V_i^{-1} (y_i - X_i \hat{\alpha}).$$

For the unknown variance, α and b_i can be estimated similarly by replacing V_i^{-1} by the estimated version $\hat{V}_i^{-1}$, where $\hat{V}_i = \hat{R}_i + Z_i \hat{D} Z_i$. In this case, $\hat{\alpha}$ and $\hat{b}_i$ are known as the Empirical Best Linear Unbiased Predictors (EBLUP) of α and b_i.

Lindstrom and Bates [27] developed an efficient Newton-Raphson algorithm for estimating parameters in mixed-effect model. They provided derivatives used in the Newton-Raphson algorithm which improved the rates of convergence. Robinson et al. [36] summarized and discussed details about the best linear unbiased prediction (BLUP) estimate of the random effect in the mixed effect model.

Spatial statistics are in many situations a special case of clustered data analyses and have become a popular research area. Spatial statistical models shared some similarities with clustered data analysis as the samples from the same area can be dependent. However, the geographical structures determine the relationship among the areas which is one of the unique features of spatial data and needs to be taken into consideration. Therefore, appropriately incorporating spatial structure into spatial modeling is a current area of interest. One classic example can be the conditional autoregressive model [5] where the effect of one region depends on the data from neighboring regions and the distance between the neighbors.

1.2 Real-world examples

There are many real-world examples of cluster data including longitudinal data and spatial data among others. One of the studies discussed in this dissertation investigates the smoking related prevalence in each county of the South Carolina using the South Carolina Adult Tobacco Survey (SCATS). In this study, individual-level information was collected, and the geographic structure should be considered because neighboring areas tend to have similar smoking patterns. Also, some counties have a small number of observations where inference is more difficult given the small sample size. For these areas, it is important to obtain unbiased point estimates along with reasonable standard deviations and confidence intervals. Small area estimation is a technique that can be applied to use the patterns observed in areas with a large number of observations to overcome small sample sizes in other areas.

Another example in this dissertation is data used to track the prevalence of stunting disease in African countries and regions over time. In this example, spatial and time trends are taken into account by using a P-spline ANOVA-type interaction model which is further transformed to a mixed effect model.

Both studies used survey data where survey sampling designs have an impact on model estimation and inference. Therefore, in addition to clustered data, the model needs to account for various survey design issues discussed further below. In this dissertation, we focus on obtaining standard errors and confidence intervals given the complex data structures.

1.3 Aims

The dissertation has a wide spectrum of research interests including spatial small area estimation, missing heterogeneity information, and missing sampling weights with specific focus on mixed effect models. In Chapter 2, we focus on estimating

confidence intervals for post-stratified spatial small area estimates. This project was initiated by estimating the county-level prevalence of multiple smoking-related outcomes in South Carolina. We fit the multilevel logistic mixed effect models with ICAR random effects to estimate county-level smoking-related outcomes using the data with unequal sampling probabilities and applied post-stratification to obtain the aggregated prevalence. Standard errors and confidence intervals were estimated using bootstrap methods. We compared the performances of three different bootstrap methods: the classic bootstrap method, a Monte Carlo based bootstrap and a weighted bootstrap. In chapter 3, we focus on incorporating missing heterogeneity information in Spatio-temporal models. The motivation of this project was to track the trend of stunting disease prevalence over time in various countries/regions. The data were collected from different sources of surveys related to stunting prevalence and pooled into a joint dataset. A unique feature of the joint dataset is that prevalence point estimates come with the sampling standard errors (SSE). Both need to be incorporated into the prevalence analysis. However, about half of the SSEs might be missing which may also impact the final prevalence estimates. Additionally, spatial dependence should be taken into consideration as well as a non-linear trend over time. We propose a penalized mixed effect model with heterogeneous errors which can incorporate smoothing Spatio-temporal effects and the missing sampling standard errors.

Chapter 2

Estimating confidence intervals for spatial hierarchical mixed-effects models with post-stratification

2.1 Introduction

Epidemiologists and other public health practitioners are increasingly turning to large population-level representative datasets to measure or monitor area-level outcomes. However, there are many statistical challenges associated with analyses of this type of data including hierarchical covariate information, the limited sample size for some areas, complex survey designs and spatial correlations, among others. Small area estimation (SAE) techniques are often adopted to generate more reliable estimates for local areas [see 34]. Post-stratification is an SAE technique that combines area-level predictions that are conditional on sub-area-level (e.g., individual-level) data via auxiliary information on sub-area-level population counts [19, 28]. Post-stratification is a popular technique since it allows for models with, for example, individual-level covariate data (e.g., gender) that can have higher predictive power than models with only area-level covariates (e.g., the proportion of females). Valliant [38] studied the asymptotic and empirical properties of post-stratified direct (not model-based) estimators and found that resampling-based estimators can reasonably estimate the mean-squared error (MSE). However, it's not clear how these results would translate

to model-based estimators, which allow for more covariate categories [16] and are more common in recent applications of SAE.

A challenge for this statistical approach is that the data sources commonly used are surveys where informative sampling designs are based on gender, race, age, and other variables. Pfeffermann [31] argued that conditioning on all the survey design variables can be a plausible approach for controlling for unequal selection probability issues. However, it might not be realistic or possible to include all the relevant sampling variables in the model. If ignored, the distribution of the sample data may be very different from the distribution in the population, possibly biasing estimates. Incorporating sampling weights via the inverse probability of selection is one of the most common ways to correct this bias.

Hierarchical mixed-effects models with post-stratification have become a common SAE method in many areas of public health [40, 41, 18, 14, 10]. For example, Zhang et al. [40] expanded the hierarchical logistic regression model to a more flexible unit-level multilevel model and applied post-stratification using US census data to generate SAE estimates of chronic obstructive pulmonary disease. Zhang et al. [40] use a multilevel logistic regression model with variables at the individual, county, and state level and produced county-level estimates specific to each particular individual variable level, which we refer to as stratum-specific county-level estimates. Post-stratification is applied by leveraging population estimates from the US Census to aggregate stratum-specific estimates to the county level. Accurate estimation of standard errors and confidence intervals for such studies are crucial to understanding where predictions reflect real public health crises or rather, a lack of data.

Obtaining the mean squared prediction error (MSPE) for post-stratified estimates via asymptotic theory is difficult and may lead to estimators that are not feasible in practice. A Bayesian approach is another alternative option. However, it is very difficult to incorporate sampling weights and survey designs in a Bayesian approach

as Bayesian model sampling procedures condition on the samples. Therefore, it can result in large bias when the data arise from unequal probability sampling. Bootstrapping [12] has long been a common technique for estimating the mean squared prediction error of predictions in SAE [23, 30, 32]. A standard bootstrap method with random sampling with replacement can be the most straightforward form of bootstrapping. However, under a complex sampling design with unequal sampling weights, the classic bootstrapping method may result in bias due to the violation of equal sampling probabilities. Antal and Tillé [2] proposed a weighted bootstrapping method to correct for sampling bias. Another bootstrap method is the Monte Carlo-based bootstrapping approach [4], where model parameters are sampled from their estimated theoretical distributions.

There are two objectives to this study. First, post-stratification involves in bringing outside population information (i.e., US census) and more complicated computation, therefore, if post-stratification is necessary to improve predictive ability is of interest. In this study, we compare the predictive ability of estimators based on post-stratification versus those from non-post-stratified models in hierarchical mixed-effects models. Here, non-post-stratification models use the marginal proportion of each level of the variable. Non-post-stratified methods result in more straightforward forms for estimating confidence intervals, but average over individual level information which may be less accurate at the local level. The second objective is to evaluate the performances of different methods for estimating accurate confidence intervals under complex sampling designs. Various bootstrap methods are available, and some of them require intensive computational effort. It is of interest to determine which bootstrap methods that can provide accurate inference with less computational burden. In Section 2, we review a commonly used post-stratification model and detail the statistical challenges. In Section 3, we describe the various methods that can be used to estimate uncertainty under a complex sampling design. In Section 4, we

conduct simulation studies and present the simulation study results. In Section 5, we apply the methods to a recent study on smoking exposure and prevalence in South Carolina. Discussions and conclusions will be presented in Section 6.

2.2 Model specification

We are interested in estimating the prevalence of an outcome of interest in area j, denoted as p_j, for $j = 1, ..., J$. Here we consider a two-level logistic mixed effect model, which can also be generalized to other exponential family distributions. The two-level model is given by

$$logit(p_{ij}) = X_{ij}\beta + Z_j\alpha + b_{0j},$$

where $logit(p) = \log\{p/(1-p)\}$, p_{ij} is the probability of having the outcome for second-level i (i.e., individual-level) first-level j (i.e., area-level), X_{ij} is a second-level vector of covariates, Z_j is a first-level vector of covariates and b_j is a random intercept. Under a multilevel mixed effect model structure, spatial correlations among counties can be considered as part of the random effect b_{0j} to increase the efficiency of the model. For simplicity, we will use a spatial intrinsic conditional autoregressive (ICAR) model [35] where $b_j|b_l \in \delta_j \sim N(\bar{b}_j, \sigma_b^2/m_j)$, where δ_j is the set of indices of neighbors for area j and $\bar{b}_j = \sum_{l\in\delta_j} b_l/m_j$ with number of neighbor areas of area j denoted by m_j. Other spatial models are discussed in the Appendix 2.

In this dissertation, we consider the most common case where second-level covariates are categorical (post-stratification for continuous second-level covariates is more challenging). For example, suppose there are three categorical second-level covariates with levels s ($s = 1, \ldots, S$), k ($k = 1, \ldots, K$) and l ($l = 1, \ldots, L$), respectively. The resulting model can be re-written as

$$logit(p_{sklj}) = \mu_s + \nu_k + \phi_l + Z_j\alpha + b_j, \tag{2.2.1}$$

where p_{sklj} is the stratum-specific prevalence for second-level covariate levels s, k and l in area j, with corresponding coefficients μ_s, ν_k and ϕ_l. Prediction of (2.2.1) can be made using the best linear unbiased prediction (BLUP). For generalized linear mixed effect models, Jiang and Lahiri [20] presented the best predictor (BP) to predict the outcome of as $logit(\hat{p}_{sklj}) = \hat{\mu}_s + \hat{\nu}_k + \hat{\phi}_l + Z_j\hat{\alpha} + \tilde{b}_j$ where $\tilde{b}_j = E(b_j)$. In this case, there is no explicit closed form for $\tilde{b}_j$ but the conditional expectation can be approximated as the ratio of two one-dimension integrals as mentioned previously.

To be able to aggregate the estimations for each stratum from the second level into the first level estimation, post-stratification can be used. With this approach, population level information is incorporated into stratum-specific estimates so that the final aggregated estimates are corrected by the population size of each stratum. That is, we can obtain the estimates for strata from the model, and post-stratification weights are calculated for each post-stratum based on population information (i.e., from census data). Then stratum-specific estimates are weighted by the post-stratification weights to obtain the final aggregated estimates. By adopting a similar idea from Gelman and Little [16], we can aggregate each stratum in each county to obtain county-level probability

$$\hat{p}_j = E(\hat{p}_{sklj}|\boldsymbol{Pop}_j) = \sum_s \sum_k \sum_l \hat{p}_{sklj} \frac{Pop_{sklj}}{Pop_j}, \tag{2.2.2}$$

where $\boldsymbol{Pop}_j = \{Pop_{sklj}; s = 1, \ldots, S, k = 1, \ldots, K, l = 1, \ldots, L\}$, Pop_{sklj} is the population size in area j for second-level covariates s, k and l, and $\sum_s \sum_k \sum_l Pop_{sklj} = Pop_j$ is the total population for area j.

As discussed in the introduction, an important aspect of the study is to estimate the MSPE. Here, MSPE is defined as the mean squared error of $\hat{p}_{sklj}$ as $\text{MSPE}(\hat{p}_{sklj}) = E(\hat{p}_{sklj} - p_{sklj})^2$. Jiang et al. [22] and Jiang and Lahiri [21] proposed jackknife and Taylor series expansion methods, respectively, to estimate the MSPE of $\hat{p}_{sklj}$ that can correct bias up to the second order. After the MSPE of $\hat{p}_{sklj}$ is obtained, the MSPE of $\hat{p}_j$ can be calculated using delta-method. However, this involves

calculating the covariances between each stratum-specific $\hat{p}_{sklj}$, which are not available in most software. Further, it's not clear how accurate the delta-method would be since the normality of (2.2.1) is questionable. As a result, we seek alternatives. There are several ways of bootstrapping including the standard, weighted and Monte Carlo based bootstrap methods. Details on how to estimate the MSPE using various bootstrap methods will be presented in the next section.

2.3 Estimating the MSPE

2.3.1 Standard bootstrap and weighted bootstrap method

From the finite population with size N sample data $\mathcal{D}$ is selected with size n. The standard bootstrap method used the simple random sampling with replacement (SRSWR) algorithm. However, under a complex sampling design with unequal sampling weights, the classic bootstrapping methods may result in bias due to the violation of the independence assumption. Antal and Tillé [2] proposed a weighted bootstrapping method to correct for the sampling bias. Conceptually, their weighted bootstrap method is an attempt to select bootstrap samples from the original sample so that scaling, weighting and using artificial population are not needed. Here we will adopt Algorithm 4 of the [2] to resample with unequal probability sampling without replacement. The key point of this algorithm is to yield unbiased or an approximation of unbiased variance estimator of the outcome of interest using resampling methods. The technical details of this algorithm are beyond the scope of this discussion and we refer the readers to Algorithm 4 page 539 of [2]. The weighted bootstrap sampling algorithm for the case $n = \sum_{s\in S} \phi_k \geq 2$ are summarized below:

(1) Select a sample S^*_{kA} without replacement with unequal inclusion probabilities ϕ_k with fixed sample size $n^* = \sum_{k\in S} \phi_k$ from the original dataset $\mathcal{D}$. We will discuss the choice of ϕ_k later in this section. If n^* is not an integer then we have

$$m = \begin{cases} m_1 = \lfloor n^* \rfloor, & \text{with prob. } q \\ m_2 = \lfloor n^* + 1 \rfloor, & \text{with prob } 1-q, \end{cases}$$

where $q = \lfloor n^* \rfloor + 1 - n^*$. The value of ϕ_k can be chosen by taking $1 - \phi_k = \tilde{D}_{kk}$, where $\tilde{D}_{kk}$ can be viewed as an approximated variance estimator that $\tilde{D}_{kl} = \begin{cases} c_k - \frac{c_k^2}{\sum_{j \in U} S_j c_j}, & \text{if } k = l \\ -\frac{c_k c_l}{\sum_{j \in U} S_j c_j}, & \text{if } k \neq l, \end{cases}$ where c_k are weights. There are several options that have been proposed for computing the values of c_k and the simplest one is $c_k = \frac{n}{n-1}(1 - \pi_k)$, where $\pi_k = E(S_k)$.

(2) From the units that $S_{kA}^* = 0$, a sample of S_{kB}^* can be selected according to a one-one design. The one-one re-sampling design is to randomly select n_B units from a sample size of n_B so that $E(S_{kB}^*) = Var(S_{kB}^*) = 1$, for all $k \in S_{kB}$.

The re-sampling algorithm can be shown below for the case that total sample size $n_B \geq 3$.

First, compute: $p = \lfloor \frac{1}{2} \left(1 + \sqrt{\frac{4n_B^2 + 5n_B - 1}{n_B - 1}} \right) \rfloor$ and

$\alpha_B = \frac{p(n_B - 1)(p+1) - n_B(n_B+1)}{2p(n_B - 1)}$

$$\tilde{n}_B = \begin{cases} p, & \text{with a prob. } \alpha_B \\ p+1, & \text{with a prob. } 1 - \alpha_B \end{cases}$$

Then, select a simple random sample with over-replacement with sample size $\tilde{n}_B$, denoted by S_{B1} from S_{kB}^*. The over-replacement sampling is designed as $Pr(S_1 = x_1, ..., S_N = x_N) = \binom{N+n-1}{n}^{-1}$. The marginal distribution of S_k is inverse hypergeometric distribution with $E(S_k) = n/N$,

and covariance matrix $\Delta_{kl} = \frac{(N-1)(N+n)n}{N^2(N+1)} x \begin{cases} 1, & \text{if } k = l \\ -\frac{1}{N-1}, & \text{if } k \neq l \end{cases}$.

The final step is to select a simple random sample with replacement with sample size $n_B - \tilde{n}_B$ from S_{kB}^*, denoted by S_{B2}.

In this case, the final sample is $S^*_{kB} = S_{B1} + S_{B2}$.

(3) The complete resampling dataset for the weighted bootstrap method is $\mathcal{D}^b = S^*_{kA} + S^*_{kB}$.

(4) Use $\mathcal{D}^b$ from the weighted bootstrap sampling to fit the logistic mixed effect model and get the stratum-specific prevalence estimates $(\hat{p}^w_{sklj})^b$. The post-stratification approach can be applied to aggregate the stratum-specific prevalence to the county-level prevalence estimates $\hat{p}^b_j$.

(5) Repeat steps (1)-(5) for B times to obtain a collection of B county level prevalence estimates $\hat{\boldsymbol{p}}_j = (\hat{p}^1_j, ..., \hat{p}^B_j)$. Then $\hat{\boldsymbol{p}}_j$ can be used to calculate the standard deviation and the percentile-based empirical 95%CI of $\hat{p}_j$.

To calculate the standard deviation and confidence intervals of the area-level prevalence p_j, the standard and weighted bootstrap methods have similar algorithms:

(1) The original dataset is sampled with same sample size using either SRSWR or weighted bootstrap method, yielding a bootstrap dataset $\mathcal{D}^b$.

(2) Use $\mathcal{D}^b$ to fit the logistic mixed effect model with ICAR random effect $logit(p^w_{sklj}) = \mu_s + \nu_k + \phi_l + Z_j\alpha + b_j$ and obtain the stratum-specific prevalence estimates $(\hat{p}^w_{sklj})^b$.

(3) Apply the post-stratification to 2.2.2 to get the aggregated prevalence $\hat{p}^b_j$.

(4) Repeat the steps (1) - (3) for B times so that we can have a collection of $\hat{\boldsymbol{p}}_j = \{\hat{p}^1_j, ..., \hat{p}^B_j\}$.

After obtaining $\hat{\boldsymbol{p}}_j$, we can calculate the confidence intervals. There are several bootstrap CI calculation methods including the percentile-based empirical 95% confidence interval or $\hat{\boldsymbol{p}}_j$ is assumed to follow some known distribution, e.g., Normal distribution or t-distribution. For instance, t-CI has the form $\hat{p}_j \pm t_{\alpha/2}\hat{\sigma}_B$, where $t_{\alpha/2}$ is the t

distribution critical value and $\hat{\sigma}_B$ is standard error of $\hat{p}_j$. These methods require the estimates from the bootstrap samples to be symmetrically distributed. There are methods that can correct for the bias and skewness if the bootstrap estimates are asymmetric. The bias-corrected and accelerated (BCa) bootstrap confidence interval proposed by Efron [11] includes a bias correction constant z_0 and some acceleration constant a. Briefly, let $\hat{\phi} = g(\hat{\theta})$ be the transformation of bootstrap estimate θ and $\phi = g(\theta)$ be the transformation for the true value. By using the transformation, we can have $(\hat{\phi} - \phi)/\tau \sim N(-z_0\sigma_\phi, \sigma_\phi^2)$, where $\sigma_\phi = 1 + a\phi$ and τ is a constant standard error of $\hat{\phi}$. Therefore, the confidence interval for the transformed function ϕ can be given as $(\hat{\phi} + \tau z_0) \pm \tau z_\alpha$ and the confidence interval for the parameter θ can be calculated by taking the inverse transformation $\theta = g^{-1}(\phi)$. [9] proposed an approximate bootstrap confidence intervals (ABC) which are an analytic approximation to BCa intervals. Here, we found the BCa and ABC methods to perform similarly to the standard bootstrap approach (results not shown).

2.3.2 Monte Carlo based parametric bootstrap method

The Monte Carlo based bootstrap method [33] samples the model parameters via parametric distributional assumptions. Let $\hat{\boldsymbol{\beta}}$ be a vector including all the estimated parameters from the model. In our situation, $\hat{\boldsymbol{\beta}} = (\hat{\mu}_s, \hat{\nu}_k, \hat{\phi}_l, \hat{\alpha})$ with the corresponding covariance matrix $\hat{\Sigma}_\beta$ and let $\hat{\boldsymbol{b}}$ be the EBLUP random effect with corresponding covariance matrix $\hat{\Sigma}_b$. Since covariances for $\hat{\boldsymbol{b}}$ are difficult to obtain, we consider diagonal $\hat{\Sigma}_b$. We assume that $\boldsymbol{\beta} \sim MVN(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_\beta)$ and $b_j \sim N(\hat{b}_j, \hat{\sigma}_{bj}^2)$ where $\hat{\sigma}_{bj}^2$ is the (j, j) element of $\hat{\boldsymbol{\Sigma}}_b$.

The Monte Carlo based bootstrap method can be described as

(1) Fit the original dataset $\mathcal{D}$ to model (2.2.1) and obtain the parameter estimations and the BLUP of the random effects.

(2) Generate $\boldsymbol{\beta}^b$ and $\boldsymbol{b}^b$ from their corresponding estimated distributions given above.

(3) The stratum-specific prevalence are calculated as

$$(\hat{p}^w_{jskl})^b = \frac{\exp(\hat{\mu}^b_s + \hat{\nu}^b_k + \hat{\phi}^b_l + Z_j\hat{\alpha}^b + \hat{b}^b_{0j})}{1 + \exp(\hat{\mu}^b_s + \hat{\nu}^b_k + \hat{\phi}^b_l + Z_j\hat{\alpha}^b + \hat{b}^b_{0j})}, \tag{2.3.1}$$

for each sampled $\boldsymbol{\beta}^b$ and $\boldsymbol{b}^b_0$.

(4) Apply the post-stratification 2.2.2 to get $\hat{p}^b_j$ estimates.

(5) Repeat steps (2) - (4) for $b = 1, 2, ..., B$ to obtain a collection of $\hat{\boldsymbol{p}}_j = (\hat{p}^1_j, ..., \hat{p}^B_j)$ for $j = 1, \ldots, J$.

The methods discussed in the previous section are applied to the bootstrap sample $\hat{\boldsymbol{p}}_j$ to construct prediction intervals for $\hat{p}_j$ for $j = 1, \ldots, J$.

2.4 Simulation study

To compare the effect of post-stratification under the logistic mixed effect model, we conducted numerous simulation studies to evaluate their MSPEs for models with and without post-stratification. Meanwhile, it is also of interest to compare the performances of bootstrap methods via coverage probabilities. We considered various scenarios of true prevalences, the number of covariates in the model and sample sizes.

To investigate the performances of the model for common or rare outcomes, we set the true prevalence to be $p = 0.5$ or 0.1, with $J = 20$ or 40 areas and sample size $n = 500$ or 1000. It is hypothesized that the complexity of the model may play a role in which method performs the best. Therefore, two separate covariate scenarios were considered with two or three second-level covariates. The covariates were generated as $X_{1ij} \sim Bern(1/2)$, $X_{2ij} \sim Multi(1/3, 1/3, 1/3)$ for the scenarios with 2 covariates. The model with 3 covariates is motivated by the real data where $X_{1ij} \sim Bern(0.5)$, $X_{2ij} \sim Multi(0.63, 0.12, 0.15, 0.1)$ and $X_{3ij} \sim Multi(0.1, 0.4, 0.3, 0.2)$ are designed to

mirror gender, race, and age categories, respectively, which are popular individual level covariates to use in post-stratification since population level data are commonly available.

The individual level data are generated for $N = \sum_j N_j$ subjects where the sample size in each area was generated as $N_j \sim Uniform(1000, 1500)$ and rounded up and this is the population data. The individual-specific probability of having the outcome p_{ij} for first-level i in second-level j is given by

$$logit(p_{ij}) = \boldsymbol{\beta}^T \boldsymbol{X}_{ij} + \alpha W_{ij} + b_j \quad i = 1, \ldots, N_j, \ j = 1, \ldots, J \tag{2.4.1}$$

where $W_{ij} \sim N(0, 1)$ is a covariate that is related to the sampling weights, b_j is the spatial random intercept which follows the ICAR model, and $\boldsymbol{X}_{ij}$ are the fixed effects. For all models $\beta_1 = 0.5$ and $\beta_2 = -0.5$, the 3 covariate model adds $\beta_3 = -0.5$ and β_0 was set so that $E(p_{ij}) = p$ where $p = 0.1$ or 0.5. The individual level outcome values can be generated using $Y_{ij} \sim Bern(p_{ij})$. To be clear, the population prevalence p and p_j are obtained from the stratum-specific prevalence values. THey are calculated as the observed prevalence across the entire population and areas. In each iteraction of the simulation study, the sample is selected using probabilities q_{ij} with $logit(q_{ij}) = \eta + W_{ij}$ where η was set so that $n = 500$ or 1000 samples are selected. Note that W is related to both the probability of subjects being selected from the population and the outcome. Thus, when W is unobserved sampling weights need to be used. In all simulations, we do not adjust for W in our regression model and use inverse probability weighting based on q_{ij}. A total of 500 iterations of each simulation were implemented.

All approaches were estimated using adaptive Gaussian quadrature approximation with PROC GLMMIX in SAS v9.4 (SAS Institute, Cary NC). The post-stratification weights were calculated using the population frequencies for each covariate. Post-stratification was applied to the stratum-specified prevalence estimates to get $\hat{p}_j^{post}$. For the case without post-stratification, the probability of each level of the covariate

used to generate the covariates was used to calculate the stratum-specific prevalence. That is, the prevalence is calculated by inputting the proportion of each covariate category for $\boldsymbol{X}$ yielding the value $\hat{p}_j^{std}$.

Let $\hat{p}_j^t$, $t = 1, ..., T$ be the estimated prevalence for simulation iteration t in area j. The averaged MSPE is calculated has $MSPE = \frac{1}{J}\sum_{j=1}^{J} MSE_j$, where $MSE_j = \frac{1}{T}\sum_{t=1}^{T}(\hat{p}_j^t - p_j)^2$. To evaluate the performances of MSPE estimation, the methods discussed in Section 3 were applied to obtain the 95% confidence intervals of $\hat{p}_j^t$. By checking the empirical distributions of $\hat{p}_j^t$ using each bootstrap method, there were some cases where normality was not satisfied. Therefore, the empirical percentile-based 95% prediction intervals were used. Coverage probability (CP) was calculated as the percentage of iterations where the true prevalence p_j was covered by prediction intervals.

2.4.1 Averaged MSPE Results

In Table 2.1, we present the averaged MSPE for different covariate settings, sample sizes, and prevalence. We also compared the averaged MSEs for post-stratified prevalence estimates and the naive estimates (without post-stratification). The values in Table 2.1 were obtained from a single GLMM and post-stratification was applied or the mean values of the covariates were used for prediction.

For all scenarios, the averaged MSPEs became smaller when the sample size was larger. For two covariates, the averaged MSPE was smaller in the post-stratified case than the one without post-stratification for all settings. In addition, the differences were larger when the prevalence was low. For instance, the ratio of averaged MSE for the non-post-stratified case was 31.7% larger than that for the post-stratified case given that the sample size was 1000 in 20 counties when $p = 0.1$. The ratio was 8% larger when the prevalence $p = 0.5$. For the three covariate settings, the averaged MSPE does not depend on the average sample size per area. For example,

Table 2.1: Averaged MSPE $\times 1000$ for a GLMM with and without post-stratification by the number of covariates in the model, the total sample size (n) and the number of counties (J).

		Two Covariates		Three Covariates	
n	J	$p = 0.1$	$p = 0.5$	$p = 0.1$	$p = 0.5$
			Post-stratified		
500	20	0.324	1.26	3.03	7.99
500	40	0.378	1.40	1.69	6.15
1000	20	0.209	0.860	1.43	5.23
1000	40	0.233	0.949	1.04	4.81
			Not Post-stratified		
500	20	0.442	1.36	2.94	9.59
500	40	0.472	1.51	2.07	6.91
1000	20	0.276	0.929	1.26	5.93
1000	40	0.294	0.984	0.77	5.21

the averaged MSPE is smaller for the case with $n = 1000$ in 40 areas as compared to that with $n = 1000$ in 20 areas, despite the former having fewer people per area. This is possibly due to the estimates of the random effect covariance parameters being more accurate with more areas. It appears that applying post-stratification was beneficial for $p = 0.5$, however, when $p = 0.1$ post-stratification was only beneficial when $J = 40$. From the simulations (results not shown), the difference of MSPE between post-stratification and non-post-stratification was primarily driven by the standard deviations and the biases did not make an obvious difference across all settings.

2.4.2 Coverage probability results

Table 2.2 displays the CPs for all models and settings tested. For the two covariate setting with $p = 0.5$, both standard bootstrap and weighted bootstrap methods performed equally well and had coverage probabilities close to the nominal 0.95 level. However, for $p = 0.1$ the CPs for the standard and weighted bootstrap were well below the nominal level dropping as low as 0.713 and 0.706, respectively, for the $n = 500$ and $J = 40$ case with no post-stratification. The Monte Carlo based bootstrap method had relatively conservative coverage probabilities ranging from 0.977 to 0.999. The

results for $p = 0.1$ were markedly more conservative than those with $p = 0.5$. In general, all three bootstrap methods performed better when sample sizes were larger compared to the smaller sample sizes.

For the two covariate setting, post-stratification generally performed better than the naive method for both the standard bootstrap and weighted bootstrap methods, while it had little impact on the Monte Carlo approach. In addition, the CPs were much lower for the cases without post-stratification when the sample sizes were small. For the case where$n = 500$ in 40 counties, CP for post-stratification was 0.885 compared to 0.713 for the case without post-stratification when using standard bootstrap.

Overall, the CPs for three covariates were lower than those for the two covariate setting. When $p = 0.5$, Monte Carlo based bootstrap methods performed the best among the three bootstrap methods with coverage probabilities ranging from 0.886 to 0.969. Both the standard and weighted bootstrap methods had CPs well below the nominal level. When $p = 0.1$, the Monte Carlo based bootstrap method had CPs close to the nominal level and performed the best among the three bootstrap methods with or without post-stratification. In this low prevalence setting, post-stratification resulted in CPs closer to the nominal level for both standard and weighted bootstrap.

2.4.3 Bayesian Approach

Bayesian approach via MCMC is an alternative to the frequentist's bootstrap methods to calculate confidence intervals, where estimating the MSPE of $\hat{p}_j$ is straightforward. In this approach, we can specify flat priors to the parameter vector $\boldsymbol{\beta}$ and random effect $\boldsymbol{b}$. The original dataset is fit to the multilevel logistic regression model and samples of the parameters $\hat{\boldsymbol{\beta}}$ and random effect $\hat{\boldsymbol{b}}$ can be drawn from the posterior distribution via MCMC. In each iteration, stratum-specific prevalence can be calculated as in 2.2.1 and then apply the post-stratification to get the aggregated estimates

Table 2.2: Empirical coverage probabilities for the standard bootstrap (Standard), the Monte Carlo based bootstrap (MC), the weighted bootstrap (Weighted) by the total sample size (n) and the number of counties (J). Displayed is the proportion of times the estimated 95% confidence or credible interval contained the true value.

		Two Covariates			Three Covariates		
n	J	Standard	MC	Weighted	Standard	MC	Weighted
$p = 0.1$							
				Post-stratified			
500	20	0.909	0.993	0.902	0.898	0.944	0.893
1000	20	0.926	0.993	0.924	0.929	0.942	0.915
500	40	0.885	0.999	0.739	0.811	0.969	0.811
1000	40	0.883	0.992	0.873	0.906	0.959	0.891
				Not Post-stratified			
500	20	0.834	0.991	0.824	0.636	0.962	0.613
1000	20	0.875	0.991	0.836	0.747	0.947	0.711
500	40	0.713	0.992	0.706	0.636	0.995	0.602
1000	40	0.834	0.986	0.820	0.783	0.989	0.770
$p = 0.5$							
				Post-stratified			
500	20	0.950	0.989	0.942	0.867	0.925	0.854
1000	20	0.948	0.977	0.933	0.884	0.886	0.877
500	40	0.938	0.990	0.933	0.859	0.956	0.847
1000	40	0.941	0.982	0.925	0.855	0.923	0.818
				Not Post-stratified			
500	20	0.949	0.989	0.943	0.892	0.940	0.879
1000	20	0.943	0.978	0.934	0.903	0.909	0.895
500	40	0.937	0.991	0.933	0.879	0.969	0.870
1000	40	0.938	0.981	0.931	0.877	0.948	0.848

of $\hat{p}_j^b$. Table 2.3 presents the coverage probabilities for Bayesian models with 2 covariates and 3 covariates where the simulation settings are analogous to previous simulations.

2.5 Application to Adult Tobacco Survey data

We used the 2014-2015 South Carolina (SC) Adult Tobacco Survey (ATS), which is a large state-based telephone survey collecting tobacco related information from noninstitutionalized SC adults. Specifically, the dataset consists of multiple tobacco-related

Table 2.3: Coverage probabilities for Bayesian models with 2 or 3 covariates

		Two Covariates		Three Covariates	
n	J	$p = 0.1$	$p = 0.5$	$p = 0.1$	$p = 0.5$
Post-stratified					
500	20	0.876	0.925	0.872	0.878
500	40	0.991	0.895	0.776	0.851
1000	20	0.915	0.932	0.897	0.850
1000	40	0.862	0.900	0.849	0.834
Not Post-stratified					
500	20	0.920	0.994	0.946	0.999
500	40	0.776	0.973	0.833	0.993
1000	20	0.973	0.999	0.990	1.000
1000	40	0.908	0.991	0.948	0.999

variables including a binary outcome of smoking status assessed among 7503 survey participants. We included individual-level sex (2 categories), race (5 categories), and age (4 categories) as fixed effect variables. We also included two-way interactions between the variables as they also showed a significant impact on the smoking status. One of the primary goals was to examine the smoking prevalence at the county-level, and we used the spatial logistic mixed effects model with post-stratification. SAE techniques were necessary given that there were some counties with limited sample sizes which can lead to unstable estimates. Specifically, we used 2-level spatially intrinsic conditional autoregressive random intercept by assuming that neighboring counties might share some similarities. Stratum-specific estimates for each county were estimated and averaged based on population size using information from the US Census. The standard errors and confidence intervals for county-level smoking prevalence were calculated using three different bootstrap methods. In order to compare the performances of the 4 methods, we calculated the widths of CIs of the counties and then averaged them to obtain the averaged CIs.

The averaged estimated smoking prevalence is approximately 18% (results not shown) for all 46 counties. Table 2.4 presents the averaged widths of CIs from the 3 aforementioned methods. The widths of CIs for the standard and weighted boot-

Table 2.4: Average width of post-stratified confidence intervals and correlation of estimated values between methods: standard bootstrap (STANDARD), the Monte Carlo bootstrap (MC), the weighted bootstrap (WEIGHTED).

	Bootstrap methods		
	STANDARD	**MC**	**WEIGHTED**
Averaged CI width	0.144	0.247	0.145
	Pearson Correlation coefficients		
MC	0.893		
WEIGHTED	0.981	0.909	

strap methods are similar (0.144 and 0.145, respectively), which is consistent with the simulation results. The Monte Carlo-based bootstrap methods had wider CIs, compared to the standard and weighted bootstrap methods. This is also expected from the simulation study where the coverage probability for both Monte Carlo-based bootstrap was close to 1.

2.6 Discussion

2.6.1 Conclusions

Bootstrap methods provide a general and straightforward way of calculating SE and CIs without involving further theoretical derivations. The multilevel logistic mixed effects model with a complex sampling design has broad application in many areas including prevalence estimation. We have found that incorporating sampling weights via IPW leads to estimates with low bias and good precision. Besides obtaining unbiased point estimates, calculating MSPE and CIs are vital to accurate quantification of error. Bootstrap methods are preferable in many cases, however, how they should be used with complex sampling designs is uncertain. In this study, we evaluated the performances of three bootstrap methods for calculating CIs of a logistic mixed effects model with a spatial random effect and we also compared the effect the post-stratification under such complicated models.

Our simulation results suggest that when two covariates are used, post-stratification had beneficial effects as the averaged MSPE of prevalence estimates were smaller as compared to the cases without post-stratification. Post-stratification might not be necessary in some cases with large sample sizes or when many covariates are used. For the three-covariate setting, we found that the parametric Monte Carlo bootstrap approach was the closest to the nominal 0.95 value in all settings with or without post-stratification. This is interesting considering the Monte Carlo based bootstrap is not as computationally intensive as the other bootstrap methods as it only requires implementing the analysis once. The Monte Carlo based bootstrap was conservative for two covariates but always held its coverage probability. Overall, we did not see any benefit to using the weighted bootstrap method over the standard bootstrap.

The weighted hierarchical logistic model with spatial random effects and post-stratification is a powerful approach to implementing SAE on large survey's with complex sampling. More work is needed on graphical approaches to jointly display point estimates and measures of uncertainty, as uncertainty is commonly ignored in such analysis. This paper has demonstrated that the Monte Carlo based bootstrap, an approach that is less computationally intensive than most others, can provide confidence intervals that hold their coverage probabilities in a wide variety of settings.

Chapter 3

Incorporating missing heterogeneity information into spatial-temporal model

3.1 Introduction

Stunting is the most prevalent form of child malnutrition and one of the best overall indicators of a child's well-being. The United Nations have proposed Sustainable Development Goals (SDGs), one of which calls for a reduction of stunting prevalence in children under 5, with a target of a 40% reduction in children under 5 by 2025. Therefore, monitoring and assessing the progress towards these targets is important in terms of informing public health practitioners and policymakers. However, tracking the trend of prevalence over time can be challenging. One of the challenges is the non-linear time trend. This trend may be very flexible, and it is not appropriate to assume linearity. Consequently, a flexible statistical model is needed to smooth the non-linear trend over time. Similarly, spatial dependency among the neighboring areas/regions also needs to be considered. The dependence of one area to its neighboring areas may depend on the distances between the areas. Alkema and New [1] proposed a Bayesian B-spline bias-reduction model that was able to flexibly smooth the trend of under 5 mortality rate (U5MR). In their approach, the outcome log-transformed U5MR was modeled by a penalized B-spline model and the model was fitted in a Bayesian framework.

Another challenge is the sparse amount of health information in consecutive years. For example, our motivating data contains a total of 260 survey observations for 54

countries over 10 years of time. Sample data can yield large standard errors and wide confidence intervals for small sample sizes. Statistical models that are capable of borrowing strength from the available information to make predictions are needed. De Onis et al. [8] combined multiple national and sub-national surveys from various countries and used a linear mixed-effect model to estimate the trend of child growth and malnutrition for over 100 countries and sub-regions for 15 years.

A joint data set has been constructed by the World Health Organization, United Nations Children's Fund, and the World Bank that used the information from more than 700 surveys from 150 countries in Africa. By using the different sources of surveys and datasets, we expect to produce more reliable predictions of prevalence. In this joint dataset, malnutrition estimates were collected from different sources of surveys, such as demographic and health surveys (DHS), multiple indicator cluster surveys (MICS), among other types of surveys (Others) over time. Although great effort has been put to collect and standardize the data, the difference in survey quality is non-ignorable. Therefore, sampling standard errors (SSE) of the prevalence estimates from each survey have been incorporated in the joint dataset which needs to be considered in the data analyses. The magnitudes of the SSEs will vary due to the data having different sampling designs and sample sizes, which have a large influence on the uncertainty in the of prevalence estimates. Preferably, the studies with smaller SSEs (better data quality) have larger inflence on the prevalence estimates than the ones with poorer data quality. Alkema and New [1] proposed to use a survey type-specific bias parameters to control for variation of the data quality. The idea is based on the assumption that the data quality from the same survey type are comparable. In this study, a further difficulty with incorporating the SSEs is that approximately half of the SSEs are missing. The estimates could be biased if the missing heterogeneity information was estimated and simply imputed, especially when the SSEs are potentially large. Some studies may incorporate observed and missing

SSEs in the analyses by simply imputing the missing SSEs using observed SSEs and other information. This method ignores the uncertainty in the missing SSEs.

This study aimed to develop a statistical model that can fit flexible spatio-temporal trends and incorporate heterogeneity information with potential missingness. McLain et al. [29] used a penalized longitudinal model that used penalized B-splines (P-spline) to model the flexible time effect, where SSE information was incorporated in the residual variance. The missing SSEs were imputed using single imputation depending on the sample size (when known), estimated prevalence and survey types observed. This method does not account for the uncertainty in the SSE values which could also influence the reliability of the estimation. There are some similarities between their methods and our proposed method in terms of using penalized B-splines to model the flexible spatio-temporal effects and the idea of incorporating SSEs into the residual variance. However, our model also considers spatial information and we are able to use an approximated likelihood function for prevalence estimation that incorporates SSE uncertainty. Our procedure has a closed form optimization function and requires less computational effort. In addition, in our method, we take the uncertainty of the SSEs into consideration by assuming a gamma distribution which is later incorporated into the mixed effect model.

The outline for this paper is as follows. In Section 2, we introduce the P-spline model we use to smooth spatio-temporal effect and give the mixed effect model representation of the P-spline model. In Section 3, we introduce our method to incorporate the heterogeneity information with potential missingness into a mixed effect model. In Section 4, we present the results of a simulation study to validate our proposed model and compare it to a standard approach. In Section 5, we apply our proposed method for real data analysis and validate our method compared with classical methods. We finalize the article with concluding marks in Section 6.

3.2 P-spline model specification and its mixed effect model representation

3.2.1 P-spline model

As discussed above, non-linear time effects and spatial dependency can be challenging and we proposed to use a spline-type model to smooth spatio-temporal effects. The spline regression model has the general form

$$f(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{B\theta}, \tag{3.2.1}$$

where $\boldsymbol{B}$ is the regression basis of the smoothing variables, $\boldsymbol{\theta}$ is a vector of parameters and $\boldsymbol{\mu}$ is a function of spatial and time information. Here, $\boldsymbol{\theta}$ can be estimated by minimizing the sum of squares: $S = (\boldsymbol{y}-\boldsymbol{B\theta})'(\boldsymbol{y}-\boldsymbol{B\theta})$. However, the choices of knots greatly inflence the B-spline model fitting: the model could overfit the data if too many knots are selected leading to high variance and underfit the data; and if too few knotsare selected which could result in high bias [13]. Eilers and Marx [13] proposed to include many knots with a penalty term to control for over-fitting via penalized sum of squares $S_p(\boldsymbol{\theta}) = (\boldsymbol{y}-\boldsymbol{B\theta})'(\boldsymbol{y}-\boldsymbol{B\theta})+\boldsymbol{\theta'P\theta}$. The penalty $\boldsymbol{P}$ proposed by Eilers and Marx [13] used a difference of the adjacent coefficients, while Currie and Durban [6] used a second order difference penalty which will be adopted here. In this approach, the penalty term is $\boldsymbol{P} = \lambda\boldsymbol{D'D}$ with $\boldsymbol{D}$ being the second order difference matrix of the regression coefficients $\boldsymbol{\theta}$ and λ is the smoothing parameter. In the case of the second order difference, $D_{jj} = 1$, $D_{jj+1} = -2$, $D_{jj+2} = 1$ and all other $D_{jj'} = 0$. The penalized sum of squares can be re-written as: $S_p(\boldsymbol{\theta}) = (\boldsymbol{y}-\boldsymbol{B\theta})'(\boldsymbol{y}-\boldsymbol{B\theta})+\lambda\boldsymbol{\theta'D'D\theta}$.

Currie and Durban [6] used a penalized spline (P-spline) smoothing method to flexibly model the effect of time. P-spline smoothing methods can also be expanded to multi-dimensional smoothing. Lee and Durbán [25] proposed to use P-spline smooth-

ing for spatial data, where longitude and latitude information were used for smoothing

$$\boldsymbol{y} = \mu_s(\boldsymbol{V}_1, \boldsymbol{V}_2) + \boldsymbol{\epsilon} = \boldsymbol{B}(\boldsymbol{V}_1, \boldsymbol{V}_2)\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{3.2.2}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{I})$ and $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are vectors of longitude and latitude information, respectively. Suppose that $\boldsymbol{B}(\boldsymbol{V}_1)$ and $\boldsymbol{B}(\boldsymbol{V}_2)$ have dimensions of $n_s \times c_1$ and $n_s \times c_2$, respectively, where n_s is the number of unique spatial longitude and latitude points and c_1 and c_2 are the number of knots for the B-spline basis along latitude and longitude, respectively. Here, $\boldsymbol{B}(\boldsymbol{V}_1, \boldsymbol{V}_2) = \boldsymbol{B}(\boldsymbol{V}_2)\square\boldsymbol{B}(\boldsymbol{V}_1)$, where $\boldsymbol{B}(\boldsymbol{V}_1)$ and $\boldsymbol{B}(\boldsymbol{V}_2)$ are the B-spline basis for the $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$, respectively. The tensor product $\square$ operation is defined as $\boldsymbol{B}_2\square\boldsymbol{B}_1 = (\boldsymbol{B}_2 \otimes \boldsymbol{1}'_{c_1}) \odot (\boldsymbol{1}'_{c_2} \otimes \boldsymbol{B}_1)$, where $\otimes$ is the kronecker product and $\odot$ is the element-wise product. If $\boldsymbol{A}$ is an $m \times n$ matrix and $\boldsymbol{C}$ is a $p \times q$ matrix, then the Kronecker product $\boldsymbol{A} \otimes \boldsymbol{C}$ is a $mp \times nq$ block matrix that $\boldsymbol{A} \otimes \boldsymbol{C} = \begin{bmatrix} a_{11}\boldsymbol{C} & \dots & a_{1n}\boldsymbol{C} \\ \dots & & \\ a_{m1}\boldsymbol{C} & \dots & a_{mn}\boldsymbol{C} \end{bmatrix}$, and $\boldsymbol{1}_{c_1}$, and $\boldsymbol{1}_{c_2}$ are vectors of ones with length of c_1 and c_2, which match the column numbers of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$. In the end, $\boldsymbol{B}(\boldsymbol{V}_1, \boldsymbol{V}_2)$ has the dimension of $n_s \times c_s$, where $c_s = c_1c_2$.

Lee and Durbán [26] expanded it to P-spline ANOVA-type interaction models that smoothed the effect of time, spatial effects and the interactions between the time and spatial effects. That is, their model allows for an additive relationship of time, space and time-space interaction:

$$\boldsymbol{y} = \gamma + \boldsymbol{\mu}_s(\boldsymbol{V}_1, \boldsymbol{V}_2) + \boldsymbol{\mu}_t(\boldsymbol{t}) + \boldsymbol{\mu}_{st}(\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{t}) + \boldsymbol{\epsilon}, \tag{3.2.3}$$

where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{I}\sigma^2)$. Here, $\mu_s(\boldsymbol{V}_1, \boldsymbol{V}_2) = \boldsymbol{B}_s(\boldsymbol{V}_1, \boldsymbol{V}_2)\boldsymbol{\theta}_s$ is the spatial smoothing function, which is analogous to the P-spline model in (3.2.2) with dimension of $n_tn_s \times c_s$. Note that n_t is the number of unique time points, which is allowed to be different from n_s. Further, $\mu_t(\boldsymbol{t}) = \boldsymbol{B}_t(\boldsymbol{t})\boldsymbol{\theta}_t$ is the smoothing function for the time effect, where $\boldsymbol{B}_t(\boldsymbol{t})$ is the B-spline basis for time $\boldsymbol{t}$, which has dimension of $n_tn_s \times c_t$. Lastly,

$\mu_{st}(\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{t}) = \boldsymbol{B}_{st}(\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{t})\boldsymbol{\theta}_{st}$ is the interaction between the spatial and time effects and $\boldsymbol{B}_{st} = \boldsymbol{B}_s \otimes \boldsymbol{B}_t$. The general form for the B-spline basis matrix $\boldsymbol{B}$ in the spline model is defined as $\boldsymbol{B} = [\mathbf{1}_{n_s n_t} : \boldsymbol{B}_s \otimes \mathbf{1}_{n_t} : \mathbf{1}_{n_s} \otimes \boldsymbol{B}_t : \boldsymbol{B}_s \otimes \boldsymbol{B}_t]$, where $\mathbf{1}_{n_s n_t}$, $\mathbf{1}_{n_t}$ and $\mathbf{1}_{n_s}$ are vectors of ones with length of $n_s n_t$, n_t and n_s, respectively and the $\boldsymbol{B}$ matrix has dimension of $n_s n_t \times (1 + c_t + c_s + c_t c_s)$.

The smoothing model is penalized on $\boldsymbol{B}_s$, $\boldsymbol{B}_t$ and $\boldsymbol{B}_{st}$, separately using $\boldsymbol{P}^{(s)}$, $\boldsymbol{P}^{(t)}$ and $\boldsymbol{P}^{(st)}$. The spatial penalty term $\boldsymbol{P}^{(s)} = \lambda_1 \boldsymbol{I}_{c_2} \otimes \boldsymbol{D}_1' \boldsymbol{D}_1 + \lambda_2 \boldsymbol{D}_2' \boldsymbol{D}_2 \otimes \boldsymbol{I}_{c_2}$, where $\boldsymbol{D}_1$ is the second order difference matrix corresponding to the latitude information $\boldsymbol{V}_1$ and $\boldsymbol{D}_2$ is the difference matrix corresponding to the longitude $\boldsymbol{V}_2$. The time penalty term $\boldsymbol{P}^{(t)} = \lambda_t \boldsymbol{D}_t' \boldsymbol{D}_t$, where $\boldsymbol{D}_t$ is the difference matrix for time $\boldsymbol{t}$. The spatial and time interaction penalty term is given by $\boldsymbol{P}^{(st)} = \tau_2 \boldsymbol{D}_2' \boldsymbol{D}_2 \otimes \boldsymbol{I}_{c_1} \otimes \boldsymbol{I}_{c_t} + \tau_1 \boldsymbol{I}_{c_2} \otimes \boldsymbol{D}_1' \boldsymbol{D}_1 \otimes \boldsymbol{I}_{c_t} + \tau_3 \boldsymbol{I}_{c_2} \otimes \boldsymbol{I}_{c_1} \otimes \boldsymbol{D}_t' \boldsymbol{D}_t$. The penalty term $\boldsymbol{P}$ is a block diaganol martrix with $\boldsymbol{P}^{(s)}$, $\boldsymbol{P}^{(t)}$ and $\boldsymbol{P}^{(st)}$ on the diaganol such that $\boldsymbol{P} = \text{blockdiag}(0, \boldsymbol{P}^{(s)}, \boldsymbol{P}^{(t)}, \boldsymbol{P}^{(st)})$.

3.2.2 Mixed effect model representation of the P-spline model

Connections between spline models and linear mixed effect models have drawn attentions from various aspects. Brumback et al. [3] used truncated polynomials to represent P-spline models and made a connection between P-spline models, mixed effect models and the BLUP estimator. Verbyla et al. [39] derived the mixed effect model representation of the cubic P-spline model. Currie and Durban [6] developed a flexible P-spline model that adopted the connection between mixed effect models and cubic spline models. Lee and Durbán [26] proposed a method of ANOVA-type interaction models for transforming spatio-temporal effects to a mixed effect model representation.

The standard form of the linear mixed effect model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{3.2.4}$$

with $\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{G})$ being the random effect coefficients and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Lambda})$ being the error term where $\boldsymbol{\Lambda}$ is a positive-definite matrix that can be used to incorporate hetereogeneity information.

The main idea of the representation is to rewrite the non-parametric spline model as the sum of fixed and random effects. That is, we would like to transform B-spline basis $\boldsymbol{B}$ into $[\boldsymbol{X} : \boldsymbol{Z}]$ matrices. The transformation from B-spline to mixed effect model is not the focus of this paper, and we recommand readers to [6] for more details. In summary, a one-to-one transformation matrix $\boldsymbol{T}$ is constructed by reformating the penalty $\boldsymbol{P}$ such that $\boldsymbol{BT} = [\boldsymbol{X} : \boldsymbol{Z}]$. The penalty $\boldsymbol{P}$ (or the matrix $\boldsymbol{D}'\boldsymbol{D}$) can be decomposed as $SVD(\boldsymbol{D}'\boldsymbol{D}) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}'$ using singular value decomposition (SVD), where $\boldsymbol{\Sigma}$ is a diagnal matrix with the eigenvalues of $\boldsymbol{D}'\boldsymbol{D}$ on the diagonal and $\boldsymbol{U}$ is the matrix of eigenvectors. The matrix $\boldsymbol{U}$ can be partitioned into $\boldsymbol{U} = [\boldsymbol{U}_n : \boldsymbol{U}_s]$ where $\boldsymbol{U}_n$ corresponds to the zero eigenvectors and $\boldsymbol{U}_s$ corresponds to the non-zero eigenvector. The transformation matrix $\boldsymbol{T} = [\boldsymbol{T}_n : \boldsymbol{T}_s]$ can be built as functions of $\boldsymbol{U}_n$ and $\boldsymbol{U}_s$, respectively. In the end, we can have $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{T}_n$ and $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{T}_s$.

For estimation purposes, by using this transformation minimizing the penalized sum of squares S_p is equivalent to minimizing the residual maximum log-likelihood (REML):

$$L_R = -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}'(\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1})\boldsymbol{y}, \tag{3.2.5}$$

where $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Lambda} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$ is the marginal variance of $\boldsymbol{y}$.

3.2.3 Estimation and asymptotic inference

Under mixed effect model framework, the empirical best linear unbiased predictors (EBLUP) are $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{\Psi}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Psi}}^{-1}\boldsymbol{y}$ and the predicted random effect $\hat{\boldsymbol{b}} = \hat{\boldsymbol{G}}\boldsymbol{Z}'\hat{\boldsymbol{\Psi}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{G}} = \hat{\sigma}^2\boldsymbol{F}^{-1}$ is the estimated $\boldsymbol{G}$ and $\hat{\boldsymbol{\Psi}} = \hat{\sigma}^2\hat{\boldsymbol{\Lambda}} + \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}'$ is the estimated $\boldsymbol{\Psi}$. For a new subject with covariate matrix $\boldsymbol{C}_i = [\boldsymbol{X}_i, \boldsymbol{Z}_i]$, the

expected value of the new observation $\boldsymbol{\phi}_i = \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i$. The variance of the $\boldsymbol{\phi}_i$ is

$$\hat{\Sigma}_{\boldsymbol{\phi}_i} = \boldsymbol{C}_i(\boldsymbol{C}'\hat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{C} + \hat{\boldsymbol{K}})^{-1}\boldsymbol{C}_i', \tag{3.2.6}$$

[37] where $\boldsymbol{K}$ is a block diagonal matrix of a $(p+q)\times(p+q)$ zeros and $\hat{\boldsymbol{G}}^{-1}$. Here, p is the number of columns in $\boldsymbol{X}$ and q is the number of columns in $\boldsymbol{Z}$. In this case, the variance of the new observation y_i can be calculated as

$$\hat{\Sigma}_{y_i} = \hat{\sigma}^2\hat{\boldsymbol{\Lambda}}_i + \hat{\Sigma}_{\phi_i}, \tag{3.2.7}$$

Assuming a normal distribution, the predicted interval of the new observation are given by $(\hat{\phi} - Z_{1-\alpha/2}\hat{\sigma}_{y_i}, \hat{\phi} + Z_{1-\alpha/2}\hat{\sigma}_{y_i})$, where $\hat{\sigma}_{y_i}$ is the standard error of y_i which is the square root of $\hat{\Sigma}_{y_i}$.

3.3 Incorporating heterogeneity information into mixed effect model

We assume the sampling precision S_{ij} (which is the reciprocal of the variance) for each area i and data point j follows a gamma distribution $S_{ij} \sim \Gamma\left(\nu_0, \frac{\nu_{ij}}{\nu_0}\right)$ with mean $E(S_{ij}) = \nu_{ij}$ and variance $Var(S_{ij}) = \nu_{ij}^2/\nu_0$. We allow the expected precision ν_{ij} to vary across the areas and time points. We also assume that the expected sampling precision ν_{ij} is related to a known covariate matrix $\boldsymbol{W}_{ij}$, such that $\nu_{ij} = \exp(\eta_0 + \boldsymbol{\eta_1}\boldsymbol{W}_{ij})$ where η_0 and $\boldsymbol{\eta_1}$ are corresponding coefficients. The sampling precision can be incorporated into the mixed effect model in (3.2.4) by letting the error term $\epsilon_{ij} \sim N(0, \sigma^2/S_{ij})$.

In terms of missing heterogeneity information, we will assume some S_{ij} are missing at random (MAR). The whole dataset can be divided into two parts: data with observed S_{ij} denoted as $\mathcal{D}^o$ and data with missing S_{ij} denoted as $\mathcal{D}^m$. Let the observed sampling precision be $\boldsymbol{S}^o$ and the missing sampling precision be $\boldsymbol{S}^m$. Now, the S_{ij} becomes the combination of observed and predicted sampling precision $S_{ij} = (S_{ij}^o, \hat{S}_{ij}^m)$.

We can also define analogously $\boldsymbol{y}^o$ and $\boldsymbol{y}^m$ to be the outcome with and without heterogeneity information, that is $\boldsymbol{y} = (\boldsymbol{y}^o, \boldsymbol{y}^m)^T$. Similarly, let $\boldsymbol{C}^o = [\boldsymbol{X}^o, \boldsymbol{Z}^o]$ and $\boldsymbol{C}^m = [\boldsymbol{X}^o, \boldsymbol{Z}^m]$ be the covariate matrices for fixed and random effects. Under the MAR assumption, the observed data could be used to fit a gamma regression model, yielding consistent paramter estimates $\hat{\eta}_0$, $\hat{\eta}_1$ and $\hat{\nu}_0$, and $\hat{\nu}_{ij}$ can be calcuated. Therefore the missing sampling variance can be predicted $\hat{S}_{ij}^m$ using estimated parameters.

Under the general mixed effect model framework, we assume that the conditional distribution, $f_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{b})$ is Gaussian with mean $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$ and variance $\boldsymbol{\Sigma} = \boldsymbol{R}$. Here $\boldsymbol{b} \sim MVN(0, \boldsymbol{G})$ are the random effects and $\boldsymbol{R} = \sigma^2 \boldsymbol{S}^{-1}$, where $\boldsymbol{S}^{-1}$ is a diagonal matrix with $1/S_{ij}$ on the diagonal. From there, the likelihood function for subject i can be written as

$$\begin{aligned} L_i &= \int\int f_{\boldsymbol{y}|\boldsymbol{S},\boldsymbol{b}}(\boldsymbol{y}_i|\boldsymbol{b}_i) dF_{\boldsymbol{b}}(\boldsymbol{b}_i) dF_{\boldsymbol{S}}(\boldsymbol{S}_i^m) \\ &= \int\int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f_{\boldsymbol{y}^m|\boldsymbol{y}^o,\boldsymbol{S}^m,\boldsymbol{b}}(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{b}_i) dF_{\boldsymbol{S}}(\boldsymbol{S}_i^m) dF(\boldsymbol{b}_i), \quad (3.3.1) \end{aligned}$$

with full likelihood $L = \prod_i L_i$. Since the first part of the integration is not related to the unknown sampling precision $\boldsymbol{S}^m$ and $\boldsymbol{y}^m|\boldsymbol{b} \perp\!\!\!\perp \boldsymbol{y}^o|\boldsymbol{b}$, the likelihood can be written as

$$L = \int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i) \int f_{\boldsymbol{y}^m|\boldsymbol{S}^m,\boldsymbol{b}}(\boldsymbol{y}_i^m|\boldsymbol{S}_i^m, \boldsymbol{b}_i) dF(\boldsymbol{S}_i^m) dF(\boldsymbol{b}_i)$$

Following standard results that $\int f_{\boldsymbol{y}^m|\boldsymbol{S}^m,\boldsymbol{b}}(\boldsymbol{y}^m|\boldsymbol{S}_i^m, \boldsymbol{b}_i) dF(\boldsymbol{S}_i^m) = f^t_{\boldsymbol{y}^m|\boldsymbol{b}}(\boldsymbol{y}_i^m|\boldsymbol{b}; 2\hat{\nu}_0, \boldsymbol{\mu}_t^m, \boldsymbol{\Sigma}_t^m)$, where $f^t_{\boldsymbol{y}^m|\boldsymbol{b}}(\boldsymbol{y}_i^m|\boldsymbol{b}_i, 2\hat{\nu}_0, \boldsymbol{\mu}_t^m, \boldsymbol{\Sigma}_t^m)$ is a generalized multivariate t-distribution with degrees of freedom $df = 2\hat{\nu}_0$, mean $\boldsymbol{\mu}_t^m = \boldsymbol{X}^m\boldsymbol{\beta} + \boldsymbol{Z}^m\boldsymbol{b}$ and variance $\boldsymbol{\Sigma}_t^m = \sigma^2(\hat{\boldsymbol{S}}^m)^{-1}$ where $(\hat{\boldsymbol{S}}^m)^{-1}$ contains the predicted heterogeneity information, that is, a diagonal matrix with $\hat{\nu}_0/(\hat{\nu}_{ij}(\hat{\nu}_0 - 1))$ on the diagonal. The likelihood function can be written as

$$L_i = \int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f^t_{\boldsymbol{y}^m|\boldsymbol{b}}(\boldsymbol{y}_i^m|2\hat{\nu}, \boldsymbol{\mu}_t^m, \boldsymbol{\Sigma}_t^m) dF_{\boldsymbol{b}}(b_i). \quad (3.3.2)$$

Here, $f_{\boldsymbol{y}^o,\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i)$ has a normal distribution with mean $\boldsymbol{\mu}^o = \boldsymbol{X}^o\boldsymbol{\beta} + \boldsymbol{Z}^o\boldsymbol{b}$ and variance $\boldsymbol{\Sigma}^o = \sigma^2(\boldsymbol{S}^o)^{-1}$, where $(\boldsymbol{S}^o)^{-1}$ is a diagonal matrix with S_{ij} on the diagonal. In this

way, the likelihood function has a multivariate normal-t mixture distribution inside of the integral and it does not have a closed form when integrating with respect to the random effect. One way to solve it is to compute numerically, which may be time intensive and computationally demanding.

Here we seek an approximation that will result in a closed form for the integral. To this end, we temporarily approximate the t-distribution using a normal distribution with the same mean and variance. This results in an integral of a multivariate normal distributions which can be easily integrated. To be specific, $f^t_{\boldsymbol{y}^m,\boldsymbol{b}}(\boldsymbol{y}^m|\boldsymbol{b}; 2\hat{\nu}, \boldsymbol{\mu}^m_t, \boldsymbol{\Sigma}^m_t) \approx f^N_{\boldsymbol{y}^m,\boldsymbol{b}}(\boldsymbol{y}^m|\boldsymbol{b}; \boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)$. Then we have the approximated likelihood function

$$\tilde{L} = \int f^N_{\boldsymbol{y}^o,\boldsymbol{b}}(\boldsymbol{y}^o|\boldsymbol{\mu}^o, \boldsymbol{\Sigma}^o) f^N_{\boldsymbol{y}^m,\boldsymbol{b}}(\boldsymbol{y}^m|\boldsymbol{\mu}^m_t, \boldsymbol{\Sigma}^m_t) dF(\boldsymbol{b}). \tag{3.3.3}$$

In this case, it is easy to integrate out the random effects and we can calculate the likelihood as

$$\tilde{L} = f^N_{\boldsymbol{y}}(\boldsymbol{y}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \tag{3.3.4}$$

a multivariate normal distribution, where $\tilde{\boldsymbol{\mu}} = \boldsymbol{X}\boldsymbol{\beta}$, $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \sigma^2\boldsymbol{S}^{-1}$, and $\boldsymbol{S} = (\boldsymbol{S}^o, \hat{\boldsymbol{S}}^m)$. Let the covariance matrix of the multivariate normal distribution be $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \sigma^2\boldsymbol{S}^{-1} = \begin{pmatrix} \Sigma^o & \Sigma^{om} \\ \Sigma^{om} & \Sigma^m \end{pmatrix}$, where Σ^o and Σ^m are the variances for the $\boldsymbol{y}^o$ and $\boldsymbol{y}^m$ and Σ^{om} is the covariance matrix. Note that the covariance Σ^{om} is non-zero since the clusters can have observations in $\boldsymbol{y}^o$ and $\boldsymbol{y}^m$.

The original likelihood function has normal and t distribution mixture which corresponds to the observed and missing sampling precisions parts, respectively and we would like to convert the approximated likelihood function back to a normal-t mixture to improve the accuracy of estimation. The approximated normal likelihood function (3.3.4) can work as a standard comparison to compare the performance of our proposed method in the later simulation study. The first step is to separate the

data with observed and missing sampling precision, and the approximated likelihood function $\tilde{L}$ can then be given by

$$\tilde{L} = f^N_{\boldsymbol{y}^o}(\boldsymbol{y}^o) f^N_{\boldsymbol{y}^m|\boldsymbol{y}^o}(\boldsymbol{y}^m|\boldsymbol{y}^o), \tag{3.3.5}$$

where $f^N_{\boldsymbol{y}^m|\boldsymbol{y}^o}(\boldsymbol{y}^m|\boldsymbol{y}^o)$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}^{m|o} = \boldsymbol{X}^m\boldsymbol{\beta} + \Sigma^{om}(\Sigma^o)^{-1}(\boldsymbol{y}^o - \boldsymbol{X}^o\boldsymbol{\beta})$ and variance $\boldsymbol{\Sigma}^{m|o} = \Sigma^m - \Sigma^{om}(\Sigma^o)^{-1}\Sigma^{om}$, such that $f^N_{\boldsymbol{y}^m|\boldsymbol{y}^o}(\boldsymbol{y}^m|\boldsymbol{y}^o) = f^N_{\boldsymbol{y}^m|\boldsymbol{y}^o}(\boldsymbol{y}^m|\boldsymbol{y}^o, \boldsymbol{\mu}^{m|o}, \boldsymbol{\Sigma}^{m|o})$. Then we can switch the second part $f^N_{\boldsymbol{y}^m|\boldsymbol{y}^o}(\boldsymbol{y}^m|\boldsymbol{y}^o, \boldsymbol{C}^m)$ back to the t distribution that has degrees of freedom $2\hat{\nu}$, mean = $\boldsymbol{\mu}^{m|o}$, and variance $\boldsymbol{\Sigma}^{m|o}$ as defined above. As a result, we have the likelihood

$$\hat{\tilde{L}} = f^N_{\boldsymbol{y}^o}(\boldsymbol{y}^o|\boldsymbol{\mu}^o, \boldsymbol{\Sigma}^o) f^t_{\boldsymbol{y}^m}(\boldsymbol{y}^m|2\hat{\nu}, \boldsymbol{\mu}^{m|o}, \boldsymbol{\Sigma}^{m|o}). \tag{3.3.6}$$

We assume that after the approximations, the likelihood function L is approximately equal to the new normal-t mixture $\hat{\tilde{L}}$. More detailed derivation of the likelihood function can be found in the Appendix C.

3.4 Simulation study

To test the performance of our proposed method, numerous simulation studies have been performed. For each simulation, we generated a training set $\mathcal{D}$ which is used for estimation, and a test set $\mathcal{D}^{pred}$ used to evaluate the properties of the predictions. The data was generated for $N = 50$ countries from the years 2010 to 2021. The first 6 years were used as the training dataset $\mathcal{D}$ and last 6 years were used as the test dataset $\mathcal{D}^{pred}$. The knots for time and spatial splines for both datasets were chosen to can be the same. The time information (years) was smoothed using penalized-B spline model and the location of knots were chosen to be 3 years apart. The spatial longitude and latitude coordinates information is assumed to spread evenly over $[0, 50]$ range for the 50 countries and is also smoothed by penalized-B spline in which location of knots were chosen to be 10 points apart. In this simulation study, we will mainly focus on

the additive time and spatial interaction model

$$\boldsymbol{y} = \gamma + \boldsymbol{\mu}_s(\boldsymbol{V}_1, \boldsymbol{V}_2) + \boldsymbol{\mu}_t(\boldsymbol{t}) + \boldsymbol{b}_0 + \boldsymbol{\epsilon},$$

where $\boldsymbol{\mu}_t(\boldsymbol{t})$ is the p-spline function for time, $\boldsymbol{\mu}_s(\boldsymbol{V}_1, \boldsymbol{V}_2)$ is the interactive effect of longitude and latitude, $\boldsymbol{b}_0 \sim N(0, \sigma^2_{b0})$ is a country-specific random intercept, $\epsilon_{ij} \sim N(0, \sigma^2/S_{ij})$ is the error term where sampling precision $S_{ij} \sim \Gamma(\nu_0, \nu_{ij}/\nu_0)$ is incorporated. Here, ν_{ij} is generated from $\nu_{ij} = \exp(\eta_0 + \eta_1 * W_{ij})$ with $\eta_0 = -1$, $\eta_1 = 0.5$ and W_{ij} with $W_{ij} \sim N(0, 0.5^2)$. Different proportions of missing p^{miss} were used for the sampling precision S_{ij} that are dependent on W_{ij}. The small proportion of missing had $p^{miss} = \exp(-1-W_{ij})/(1+\exp(-1-W_{ij}))$ which corresponds to about 20% missing, while the large proportion of missing had $p^{miss} = \exp(1 - W_{ij})/(1 + \exp(1 - W_{ij}))$ which gives about 60% missing. The rationale behind this setting is that W_{ij} is negatively associated with the missing probability p^{miss} and positively related to ν_{ij}, so when W_{ij} increases the p^{miss} decreases and ν_{ij} increases which results in the increases in the expectation and variance of S_{ij}. In this setting, the proportion of missing p^{miss} is negatively associated with the covariate matrix $\boldsymbol{W}_{ij}$, which is positively related to ν_{ij}. Given the fixed value of ν_0, the variance of S_{ij} is also positively related to ν_{ij}, which results in the possitive association between the variance of S_{ij} and the proportion of missingness. The gamma parameter ν_0 is set to different values to compare the performances for large ($\nu_0 = 2$ or $\nu_0 = 5$) and small ($\nu_0 = 10$) variance of S_{ij} which is negatively related to the variance of error term ϵ_{ij}.

The model was estimated using mixed effect representation of the P-spline model and we can have $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{b}_0 + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the fixed and random effect design matrices that transformed the mixed-effect model to the P-spline model. The dimensions of $\boldsymbol{X}$ and $\boldsymbol{Z}$ matrices depend on the knots of P-spline model, $\boldsymbol{\beta}$ is the corresponding coefficients and $\boldsymbol{\alpha} \sim N(0, \sigma^2_{b1})$. Here, σ^2_{b0}, σ^2_{b1} and σ^2 were set to $0.005, 0.005$ and 2×10^{-7}, respectively which tried to mirror the parameters in the joint dataset. By minimizing the normal-t approximation likelihood function (3.3.6), we

are able to estimate the parameters along with making inference and predictions. A classical normal likelihood function (referring as a normal-normal model) was used as comparison where the missing sampling precision $\boldsymbol{S}^m$ was replaced by its expectation ν_{ij}. The degrees of freedom of the t part of the normal-t model is a function of the gamma parameter ν_0, therefore, we would expect more similar results from normal-normal model and normal-t model when ν_0 is large.

Let $\boldsymbol{Y} = (y_1, ..., y_M)$ be the true outcome with corresponding expectation $\boldsymbol{\mu} = (\mu_1, ..., \mu_M)$ in dataset $\mathcal{D}^{pred}$ and $\hat{\boldsymbol{\mu}}^{pred} = (\hat{\mu}_1, ..., \hat{\mu}_M)$ be the predicted outcome using either normal-t or normal-normal likelihood function. Root mean squared prediction error (RMSPE) and prediction coverage probability (CP) can be used to validate the prediction. RMSPE is defined as $\text{RMSPE}_Y = \sqrt{\sum_{i=1}^{M}(\hat{\mu}_i - y_i)^2/M}$ and $\text{RMSPE}_\mu = \sqrt{\sum_{i=1}^{M}(\hat{\mu}_i - \mu_i)^2/M}$.

Table 3.1 shows the width of prediction intervals (PIs) were also narrower for normal-t model than those in normal-normal model across all simulation settings. The differences in prediction interval widths between the normal-t and normal-normal models are more obvious when the missing proportion is low. When the missing proportion is about 20%, the prediction interval widths for the normal-t model is 0.136 as compared to 0.364 for the normal-normal model given $\nu_0 = 2$. However, when the missing proportion is high, the PI width for normal-t model is 0.347 as compared to 0.390 for normal-normal model. The significant difference is primarily driven by the difference between estimated $\hat{\sigma}^2$ from the error term. The ratio of $\hat{\sigma}^2$ from the normal-normal and normal-t models ($\hat{\sigma}^2_N/\hat{\sigma}^2_t$) is about 6 when the missing probability is low and about 1.2 when the missing probability is high (results not shown). The coverage probabilities for the normal-t approximation are close to 0.95 when the missing proportion for SSEs is low as compared to the normal-normal model which has higher coverage probabilities. The normal-t model tended to have lower coverage probability as compared to the normal-normal model when the missing proportion is

Table 3.1: Summarized results of simulation study. The results contain the width of prediction confidence intervals, prediction interval coverage probability, root mean squared prediction error of estimates Y (RMSPE_Y), and μ (RMSPE_μ)

	Low missing probability			High missing probability		
	$\nu_0 = 2$	$\nu_0 = 5$	$\nu_0 = 10$	$\nu_0 = 2$	$\nu_0 = 5$	$\nu_0 = 10$
Width of PI						
Normal	0.364	0.318	0.304	0.390	0.337	0.326
Normal-t	0.136	0.123	0.123	0.347	0.319	0.315
Prediction CP						
Normal	0.974	0.972	0.979	0.958	0.961	0.947
Normal-t	0.950	0.979	0.967	0.905	0.941	0.934
RMSPE_y						
Normal	0.060	0.055	0.053	0.072	0.066	0.071
Normal-t	0.039	0.028	0.029	0.086	0.073	0.074
RMSPE_μ						
Normal	0.047	0.047	0.046	0.068	0.060	0.066
Normal-t	0.017	0.013	0.015	0.075	0.066	0.068

high. Both RMSPE_Y and RMSPE_μ were smaller in the normal-t model versus those in the normal-normal when the missing proportion was low. These results suggest that normal-t model is preferred when missing proportion is low.

3.5 Real data analysis

In this real data analysis, we are interested in investigating the prevalence trend of stunting over time in African countries using the joint data set. There were 260 stunting survey estimates available from 54 countries from years 1993 to 2015 and we would like to predict the prevalence for years with no data in these countries over a 23-year period, along with prediction confidence intervals. Among the 260 survey points, 92 surveys were from DHS, 36 were from MICS and 132 were from other survey sources. To predict the stunt prevalence, we adjusted for gross domestic product (GDP), fertility rate and life expectancy. Overall, 12 survey points were excluded from the analysis due to missing covariate information. To better fit the

model, teh GDP variable was scaled to make the mean of GDP as 0 and standard deviation as 1. In addition, there were 119 SSEs missing.

A penalized B-spline model was used to smooth over longitude and latitude information as well as time (in years). The location of the spatial knots was chosen to be 10 points apart from the minimum value to maximum value for both longitude and latitude, and the knot location for time was chosen to be 10 years apart. The sampling precision is assumed to follow a gamma distribution, such that Precision $\sim \Gamma(\nu_0, \nu_{ij}/\nu_0)$, and survey categories were considered to well explain the precision, $\exp(E(\text{precision})) = \eta_0 + \eta_1 *$ (Survey categories). By using observed precisions, the parameters $\hat{\eta}_0, \hat{\eta}_1, \hat{\nu}_0$, and $\hat{\nu}_{ij}$ can be estimated and the missing precisions can be predicted using the estimated parameters. Data was fit to our proposed normal-t approximation likelihood and the standard approach to obtain the regression parameters and predict the stunting disease prevalence over year with prediction intervals.

In general, we found that stunting prevalence was decreasing, with some of countries having increasing trend until around years 2000 to 2005 and decreasing trend afterward. Some of the countries had overall stunting prevalence as low as 0.2, as compared to other countries that had stunting prevalence consistently higher than 0.4. In addition, GDP and life expectancy were negatively associated with stunting prevalence and fertility rate was positively associated with stunting prevalence. With a 1 unit increase in GDP there would be about a 2.4% decrease in prevalence, and with a 1 year increase in life expectancy there would be about a 0.37% decrease in stunting. On the other hand, a 1% increase in fertility rate, results in the stunting prevalence increasing by 4.7% on average. In addition, survey category is used to model the sampling precision and the precision for MICS surveys has 1.05 times the precision of DHS and the precision for other types of surveys is 0.56 times of that for DHS on average. The estimated $\hat{\nu}_0$ is 4.

To compare the performance of our proposed method, Figure 3.1 presents the predicted stunting prevalence for Algeria, Benin, Botswana, and Burkina Faso over time with predicted confidence intervals as well as the observed points. The left panel presented the results using the standard normal approach and the right panel are the ones using our proposed normal-t approximation method. The prediction intervals for all 4 countries are narrower for our method as compared to the standard normal approach which is consistent with the findings from the simulation study.

3.5.1 Validation

A 10-fold cross-validation (CV) analysis was conducted to validate the difference between the normal-normal model and the normal-t model as did in the simulation study. In the CV analysis, the joint dataset was used and the countries with only one observed stunting estimate were excluded from the analysis. We did not adjust any covariates in the real data cross-validation to avoid losing more survey estimates. The included survey estimates were divided into 10 folds. The point prediction with prediction intervals for out-of-sample data was computed using observed data as well as SSE values. Both normal-normal and normal-t models were fit and the coverage probability and RMSPE were calculated to compare the performances of the two models.

Table 3.2 presents the results of the 10-fold CV analyses. From the results, both the normal-normal and the normal-t models had coverage probability close to 0.95 and the normal-t model had smaller bias. The normal-t model has smaller RMSPE and narrower prediction interval width compared to the normal-normal model which, is consistent with our simulation results where missing proportion is low. Therefore, it appears that the normal-t model out-performed the normal-normal model.

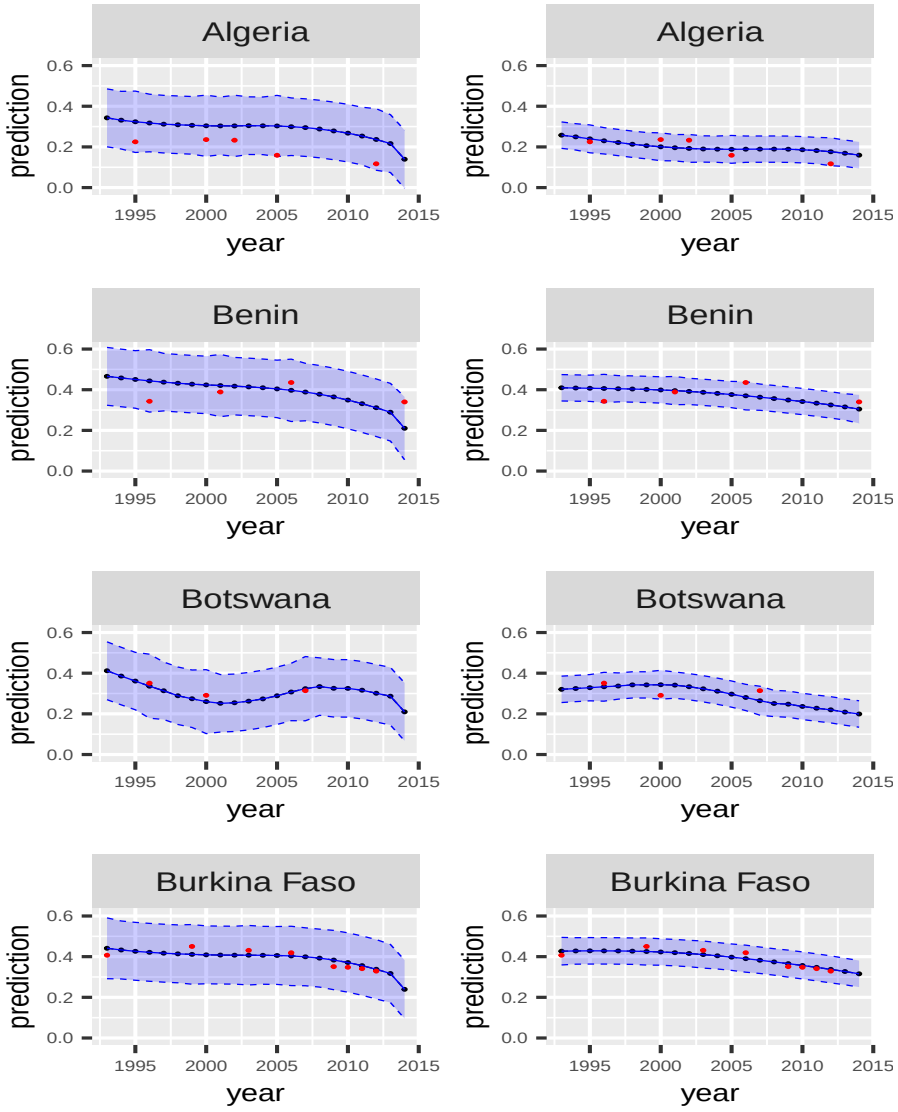

Figure 3.1: Predicted stunting prevalence from 1993-2015 for Algeria, Benin, Botswana, and Burkina Faso. The figure contains the predicted stunting prevalence with prediction confidence intervals and observed stunting prevalence. Left panel is the prevalence prediction using the standard normal approach and right panel is using normal-t approximation

3.6 Conclusion

In this paper, we developed a normal-t mixture model to fit flexible spatio-temporal trend with partial missing heterogeneity information. In this model, longitude and latitude information as well as repeated time can be modelled flexibly using penalized-B spline model which can be converted into a mixed effec model. In addition, by

Table 3.2: Summarized results of 10-fold cross-validation. The results contain the prediction confidence interval coverage probability, Bias, prediction confidence interval width, and root mean squared prediction error of stunting prevalence estimates (RMSPE_Y)

	Normal-t	Normal-normal
Mean CP	0.946	0.961
Bias	-0.001	0.002
CI width	0.185	0.442
RMSPE_Y	0.054	0.123

assuming to follow gamma distribution, the missing heterogeneity information (i.e., SSEs) are incorporated into the mixed effect model framework. For computational simplicity, we proposed a normal-t approximation of the likelihood function to estimate the parameters. We compared our method to a standard normal likelihood function which assumed the missing precision $\boldsymbol{S}_{ij}^m$ to be its predicted expectation $\hat{\nu}_{ij}$. The predicted $\boldsymbol{S}_{ij}$ for our normal-t model is $\frac{\nu_{ij}(\nu_0-1)}{\nu_0}$ (derivation details can be found in Appendix C). In the cases where ν_0 is large, we would expect simular variance estimates for both methods, and for the cases with small ν_0, the corresponding variance estimates for the normal-t model are smaller than the ones from the normal-normal method. From the simulation study, we found that our proposed method has narrower prediction intervals as well as smaller RMSPE across all settings with small missing proportions. The width of PI was slightly narrower for the normal-t model but the differences were not obvious. RMSPEs were larger for the normal-t model as compared to the normal-normal model when the missing proportion is slightly large across all $\hat{\nu}_0$ values. When the missing proportion is high $\hat{\sigma}^2$ from the normal-normal model tended to be over-estimated to have larger values, resulting in larger RMSPE estimates.

In both the simulation study and real data analysis, η_0 and η_1 can be estimated using known covariates and observed sampling precision, and are treated as fixed values without incorporating the uncertainty of their values. The uncertainty can

be incorporated via bootstrapping methods. One way is to assume distributions for both of the parameters and sample them using a parametric bootstrap. Alternatively, by resampling the covariates data, we could also measure the uncertainty of the parameters.

For the real data analysis, the results were consistent with the simulation study when $\nu = 5$ with low missing SSE proportion and our proposed method performed better than the standard normal model.

Bibliography

[1] Alkema, L. and J. R. New (2014). Global estimation of child mortality using a bayesian b-spline bias-reduction model. *The Annals of Applied Statistics 8*(4), 2122–2149.

[2] Antal, E. and Y. Tillé (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association 106*(494), 534–543.

[3] Brumback, B. A., D. Ruppert, and M. P. Wand (1999). Comment. *Journal of the American Statistical Association 94*(447), 794–797.

[4] Buckland, S. T. (1984). Monte Carlo confidence intervals. *Biometrics 40*(3), 811–817.

[5] Clayton, D. and J. Kaldor (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 671–681.

[6] Currie, I. and M. Durban (2002). Flexible smoothing with p-splines: a unified approach. *Statistical Modelling 2*(4), 333–349.

[7] Datta, G. S. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 613–627.

[8] De Onis, M., M. Blössner, E. Borghi, R. Morris, and E. A. Frongillo (2004). Methodology for estimating regional and global trends of child malnutrition. *International journal of epidemiology 33*(6), 1260–1270.

[9] Diciccio, T. and B. Efron (1992). More accurate confidence intervals in exponential families. *Biometrika 79*(2), 231–245.

[10] Dwyer-Lindgren, L., A. Bertozzi-Villa, R. W. Stubbs, C. Morozoff, S. Shirude, M. Naghavi, A. H. Mokdad, and C. J. Murray (2017). Trends and patterns of differences in chronic respiratory disease mortality among us counties, 1980-2014. *Jama 318*(12), 1136–1149.

[11] Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association 82*(397), 171–185.

[12] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 569–593. Springer.

[13] Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 89–102.

[14] Eke, P., X. Zhang, H. Lu, L. Wei, G. Thornton-Evans, K. Greenlund, J. Holt, and J. Croft (2016). Predicting periodontitis at state and local levels in the united states. *Journal of dental research 95*(5), 515–522.

[15] Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2012). *Applied longitudinal analysis*, Volume 998. John Wiley & Sons.

[16] Gelman, A. and T. C. Little (1997). Poststratification into many categories using hierarchical logistic regression. *Survey methodology 23*(2), 127–135.

[17] Hall, P. and T. Maiti (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(2), 221–238.

[18] Hao, Y., L. Balluz, H. Strosnider, X. J. Wen, C. Li, and J. R. Qualters (2015). Ozone, fine particulate matter, and chronic lower respiratory disease

mortality in the united states. *American journal of respiratory and critical care medicine 192*(3), 337–341.

[19] Holt, D. and T. F. Smith (1979). Post stratification. *Journal of the Royal Statistical Society: Series A (General) 142*(1), 33–46.

[20] Jiang, J. and P. Lahiri (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics 53*(2), 217–243.

[21] Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation. *Test 15*(1), 1.

[22] Jiang, J., P. Lahiri, S.-M. Wan, et al. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics 30*(6), 1782–1810.

[23] Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, 199–210.

[24] Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.

[25] Lee, D. J. and M. Durbán (2009). Smooth-car mixed models for spatial count data. *Computational Statistics & Data Analysis 53*(8), 2968–2979.

[26] Lee, D. J. and M. Durbán (2011). P-spline anova-type interaction models for spatio-temporal smoothing. *Statistical modelling 11*(1), 49–69.

[27] Lindstrom, M. J. and D. M. Bates (1988). Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association 83*(404), 1014–1022.

[28] Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association 88*(423), 1001–1012.

[29] McLain, A. C., E. A. Frongillo, J. Feng, and E. Borghi (2019). Prediction intervals for penalized longitudinal models with multisource summary measures: An application to childhood malnutrition. *Statistics in medicine 38*(6), 1002–1012.

[30] Molina, I., N. Salvati, and M. Pratesi (2009). Bootstrap for estimating the MSE of the spatial EBLUP. *Computational Statistics 24*(3), 441–458.

[31] Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical methods in medical research 5*(3), 239–261.

[32] Pfeffermann, D. and S. Correa (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika 99*(2), 457–472.

[33] Preacher, K. J. and J. P. Selig (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures 6*(2), 77–98.

[34] Rao, J. N. (2014). Small-area estimation. *Wiley StatsRef: Statistics Reference Online 1*(1), 1–8.

[35] Rasmussen, S. (2004). Modelling of discrete spatial variation in epidemiology with SAS using GLIMMIX. *Computer Methods and Programs in Biomedicine 76*(1), 83–89.

[36] Robinson, G. K. et al. (1991). That blup is a good thing: the estimation of random effects. *Statistical science 6*(1), 15–32.

[37] Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric regression*. Number 12. Cambridge university press.

[38] Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association 88*(421), 89–96.

[39] Verbyla, A. P., B. R. Cullis, M. G. Kenward, and S. J. Welham (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 48*(3), 269–311.

[40] Zhang, X., J. B. Holt, H. Lu, A. G. Wheaton, E. S. Ford, K. J. Greenlund, and J. B. Croft (2014). Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American journal of epidemiology 179*(8), 1025–1033.

[41] Zhang, X., J. B. Holt, S. Yun, H. Lu, K. J. Greenlund, and J. B. Croft (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American journal of epidemiology 182*(2), 127–137.

Appendix A

More discussions about SAE methods

SAE methods can be divided into “design-based" methods and “model-based" methods. Here, we will focus on model-based methods. For model-based methods, commonly used models include two main types: area-level models and unit-level models. For the unit-level models, the nested error unit level model for a continuous outcome is one of the most popular models. It takes the form $y_{ij} = x'_{ij}\beta + b_i + \epsilon_{ij}$, where the random effect $b_i \sim N(0, \sigma_b^2)$ and the error $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ are independent. The true area mean of the outcome is $E(y_i|b_i) = \theta_i = X_i'\beta + b_i$. The model requires that the area level mean of the covariates $\bar{X}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$ are known. The best linear unbiased predictor (BLUP) for θ_i can be given as $\hat{\theta} = \lambda_i \left[\bar{y}_i + (\bar{X}_i - \bar{x}_i)'\hat{\beta}_{GLS}\right] + (1-\lambda_i)\bar{X}_i\hat{\beta}_{GLS}$, where $\hat{\beta}_{GLS}$ is the estimated coefficients from the Generalized Least Square (GLS) estimation using all observations and $\lambda_i = \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2/n_i)$. For non-linear outcomes, specifically for binary outcomes, the logistic mixed effect model can be used, where $logit(p_{ij}) = x_{ij}\beta + u_i$, where x_{ij} and y_{ij} are covariates and the outcome for individual j in area i, respectively, and u_i is the random effect with $u_i \sim N(0, \sigma_u^2)$. [20] presented the best predictor (BP) for the mixed logistic model as $logit(\hat{p}_{ij}) = x_{ij}\beta + E(u_i|\sum_{j=1}^{n_i} y_{ij})$. There is no explicit closed form for $\hat{p}_{ij}$, but the conditional expectation can be approximated as the ratio of two one-dimension integrals. Therefore, statistical methods are needed to solve SAE problems which actually have two aspects: producing reliable estimates based on small sample sizes and assessing the estimation errors.

Another aspect of the SAE problem is to assess the error in the predictions using prediction mean squared errors (PMSE). Let the variance components be $\psi_i = (\sigma_\epsilon^2, \sigma_b^2)$ and the EBLUP of the outcome be $\hat{\theta}(\hat{\psi})$ obtained from $\hat{\theta}(\psi)$ with ψ replaced by the estimated $\hat{\psi}$, the PMSE can be defined as the MSE of $\hat{\theta}(\hat{\psi})$ that is $\text{MSE}(\hat{\theta}(\hat{\psi})) = E[\hat{\theta}(\hat{\psi}) - \theta]^2$. For the general mixed effect model, [7] proposed a unified PMSE estimator for the EBLUP when the unknown parameters were estimated using MLE or REML. Re-sampling procedures including the jackknife method and bootstrap methods are often used in the estimation of PMSE in GLMM models versus LMM due to the complex form of the estimator. [17] proposed a double-bootstrap method to estimate PMSE.

Appendix B

Spatial correlations

Under a multilevel mixed effect model structure, spatial correlations among counties can be considered as part of the random effect b_{0j} to increase the efficiency of the model. One option is to use the methods that incorporate the spatial dependency into the variance matrix $Var(\boldsymbol{Y})$, such as simultaneous autoregressive (SAR) model. In SAR model, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where the error term $\boldsymbol{\epsilon} = \rho \boldsymbol{W\epsilon} + \boldsymbol{\xi}$ incorporates the spatial weight matrix $\boldsymbol{W}$ that contains the weights of the linked neighbors. A generalization of SAR is the weighted SAR that includes the weights inversely proportional to the population sizes of the neighbors. Another option is to model the spatial dependency conditional on the observations of the related neighbors, such as conditional autoregressive (CAR) model. The joint distribution of the random effects $\boldsymbol{b}_0$ for the CAR model is assumed that $\boldsymbol{b}_0 \sim N(0, \sigma_b^2(\boldsymbol{I} - \phi \boldsymbol{W})^{-1})$. $\boldsymbol{W}$ is the adjacency matrix with dimension $J \times J$, where entries $w_{r,j}$ and $w_{j,r}$ are positive when region r and j are neighbors and zero otherwise. The parameter ϕ controls the spatial correlations between the neighbors. To simplify the model, we will use a spatial intrinsic conditional autoregressive (ICAR) model, which is a generalization of the CAR model with the parameter $\phi = 1$. In this case, the spatial dependency is only depends on the neighborhood structure of the regions. The ICAR model assumes that $b_j | b_l \in \delta_j \sim N(\bar{b}_j, \sigma_b^2/m_j)$, where the random effect for county j given its neighbors l follows normal distribution with mean $\bar{b}_j = \frac{\sum_{l \in \delta_j} b_l}{m_j}$ where m_j is the number of neighbor counties of county j, and δ_j is the set of indices of neighbors for county j.

Appendix C

Derivations of normal-t likelihood approximation

The whole dataset $\mathcal{D}$ contains two parts including the one $\mathcal{D}^o$ with observed sampling precision $\boldsymbol{S}^o$ and the one $\mathcal{D}^m$ with missing sampling precision $\boldsymbol{S}^m$. The outcome $\boldsymbol{y}$, covariate matrix $\boldsymbol{X}$ have analogous definitions. Let $\boldsymbol{C}^o = [\boldsymbol{X}^o, \boldsymbol{Z}^o, \boldsymbol{W}^o]$ and $\boldsymbol{C}^m = [\boldsymbol{X}^m, \boldsymbol{Z}^m, \boldsymbol{W}^m]$ be the covariate matrix for mixed and random effects and known information for sampling precision.

Under general mixed effect model framework, we can have following distribution defined

- $f_{\boldsymbol{y}|\boldsymbol{b},\boldsymbol{S}}(\boldsymbol{y}|\boldsymbol{b}) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$ and $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{S}^{-1}$ with $\boldsymbol{S} = (\boldsymbol{S}^o, \boldsymbol{S}^m)$, .
- $S_{ij}^m \sim \Gamma(\nu_0, \frac{\nu_{ij}}{\nu_0})$ is a diagonal elements of the matrix $\boldsymbol{S}^m$ and the pdf takes the form $f(S_{ij}^m) = \frac{1}{\Gamma(\nu_0)(\nu_{ij}/\nu_0)^{\nu_0}}(S_{ij}^m)^{\nu_0}\exp(-\frac{S_{ij}^m}{\nu_{ij}})$. We allow that the parameter ν_{ij} varies with S_{ij}.
- The random effect $\boldsymbol{b} \sim MVN(0, G)$.

The full likelihood function is

$$\begin{aligned} L &= \int\int f_{\boldsymbol{y}|\boldsymbol{S},\boldsymbol{b}}(\boldsymbol{y}_i|\boldsymbol{S},\boldsymbol{b})dF_{\boldsymbol{b}}(\boldsymbol{b})dF_{\boldsymbol{S}}(\boldsymbol{S}^m) \\ &= \int\int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}^o|\boldsymbol{b})f_{\boldsymbol{y}^m|\boldsymbol{y}^o,\boldsymbol{S},\boldsymbol{b}}(\boldsymbol{y}^m|\boldsymbol{y}^o,\boldsymbol{S}^m,\boldsymbol{b})dF_{\boldsymbol{S}}(\boldsymbol{S}^m)dF_{\boldsymbol{b}}(\boldsymbol{b}) \end{aligned} \tag{C.0.1}$$

For each subject, the likelihood function has the independent and identical distribution.

$$
\begin{aligned}
L_i &= \int\int f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f_{\boldsymbol{y}^m|\boldsymbol{y}^o,\boldsymbol{S}^m,\boldsymbol{b}}(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{S}_{ij}^m,\boldsymbol{b}_i) f_{\boldsymbol{S}^m}(\boldsymbol{S}_{ij}^m) d(S_i^m) dF_{\boldsymbol{b}}(\boldsymbol{b}_i) \\
&= \int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i) \int f_{\boldsymbol{y}^m|\boldsymbol{y}^o,\boldsymbol{S}^m,\boldsymbol{b}}(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{S}_{ij}^m,\boldsymbol{b}_i) f_{\boldsymbol{S}^m}(S_i^m) f_{\boldsymbol{b}}(\boldsymbol{b}_i) d(S_{ij}^m) d(\boldsymbol{b}_i) \\
&= \int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}) \int \frac{\sqrt{S_{ij}^m}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_{ij}^m-\mu_{ij})^2}{2\sigma^2} S_i^m\right) \frac{\nu_0^{\nu_0}}{\Gamma(\nu_0)\nu_{ij}^{\nu_0}} (S_{ij}^m)^{\nu_0-1} \exp(-\frac{S_{ij}^m \nu_0}{\nu_{ij}}) d(S_{ij}^m) dF_{\boldsymbol{b}_i}(\boldsymbol{b}) \\
&= \underbrace{\int f_{\boldsymbol{y}^o|\boldsymbol{b}}(y_{ij}^o|\boldsymbol{b}_i)}_{\text{L1}} \\
&\underbrace{\left[\int \frac{\sqrt{S_{ij}^m}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_{ij}^m-\mu_{ij})^2}{2\sigma^2} S_i^m\right) \frac{\nu_0^{\nu_0}}{\Gamma(\nu_0)\nu_{ij}^{\nu_0}} (S_{ij}^m)^{\nu_0-1} \exp\left(-\nu_0 \frac{S_{ij}^m}{\nu_{ij}}\right) d(S_{ij}^m)\right]}_{\text{L2}} dF_{\boldsymbol{b}}(\boldsymbol{b})
\end{aligned}
\tag{C.0.2}
$$

$$
\begin{aligned}
L2 &= \int \frac{\nu_0^{\nu_0}}{\sqrt{2\pi\sigma^2}\Gamma(\nu_0)\nu_{ij}^{\nu_0}} (S_{ij}^m)^{(\nu_0+\frac{1}{2})-1} \exp\left[-\frac{(y_{ij}^m-\mu_{ij})^2 S_{ij}^m}{2\sigma^2} - \frac{\nu_0 S_{ij}^m}{\nu_{ij}}\right] d(S_{ij}^m) \\
&= \int \frac{\nu_0^{\nu_0}}{\sqrt{2\pi\sigma^2}\Gamma(\nu_0)\nu_{ij}^{\nu_0}} (S_{ij}^m)^{(\nu_0+\frac{1}{2})-1} \exp\left[-\frac{(y_{ij}^m-\mu_{ij})^2 \nu_{ij} + 2\sigma^2\nu_0}{2\sigma^2\nu_{ij}} S_{ij}^m\right] d(S_{ij}^m)
\end{aligned}
\tag{C.0.3}
$$

Let $\Theta_{ij} = \frac{2\sigma^2\nu_{ij}}{(y_{ij}^m-\nu_{ij})^2\nu_{ij}+2\sigma^2\nu_0}$

$$
\begin{aligned}
L2 &= \int \frac{\nu_0^{\nu_0}}{\sqrt{2\pi\sigma^2}\Gamma(\nu_0)\nu_{ij}^{\nu_0}} (S_{ij}^m)^{(\nu_0+\frac{1}{2})-1} \exp\left(-\frac{S_{ij}^m}{\Theta_{ij}}\right) d(S_{ij}^m) \\
&= \frac{\nu_0^{\nu_0}}{\sqrt{2\pi\sigma^2}\Gamma(\nu_0)\nu_{ij}^{\nu_0}} \int \frac{\Gamma(\nu_0+\frac{1}{2})\Theta_{ij}^{(\nu_0+\frac{1}{2})}}{\Gamma(\nu_0+\frac{1}{2})\Theta_{ij}^{(\nu_0+\frac{1}{2})}} (S_{ij}^m)^{\nu_0-\frac{1}{2}} \exp\left(-\frac{S_{ij}^m}{\Theta_{ij}}\right) d(S_{ij}^m) \\
&= \frac{\nu_0^{\nu_0}}{\sqrt{2\pi\sigma^2}\Gamma(\nu_0)\nu_{ij}^{\nu_0}} \Gamma(\nu_0+\frac{1}{2})\Theta_{ij}^{(\nu_0+\frac{1}{2})}
\end{aligned}
\tag{C.0.4}
$$

Replace Θ_{ij} back into the formula

$$
\begin{aligned}
L2 &= \frac{\nu_0^{\nu_0}\Gamma(\nu_0+\frac{1}{2})}{\sqrt{2\pi\sigma^2}\Gamma(\nu_0)\nu_{ij}^{\nu_0}} \Gamma(\nu_0+\frac{1}{2}) \left[\frac{(y_{ij}^m-\mu_{ij})^2\nu_{ij}+2\sigma^2\nu_0}{2\sigma^2\nu_{ij}}\right]^{-(\nu_0+\frac{1}{2})} \\
&= \frac{\nu_0^{\nu_0}\Gamma(\nu_0+\frac{1}{2})}{\sqrt{2\pi\sigma^2}\Gamma(\nu_0)\nu_{ij}^{\nu_0}} \Gamma(\nu_0+\frac{1}{2}) (\frac{\nu_0}{\nu_{ij}})^{-(\nu_0+\frac{1}{2})} \left[\frac{(y_{ij}^m-\mu_{ij})^2\frac{\nu_{ij}}{\nu_0}+2\sigma^2}{2\sigma^2}\right]^{-(\nu_0+\frac{1}{2})} \\
&= \frac{\Gamma(\nu_0+\frac{1}{2})}{\Gamma(\nu_0)} \left(\frac{\nu_{ij}}{2\pi\sigma^2\nu_0}\right)^{\frac{1}{2}} \left[1+\frac{(y_{ij}-\mu_{ij})^2\nu_{ij}}{2\sigma^2\nu_0}\right]^{-(\frac{2\nu_0+1}{2})}
\end{aligned}
\tag{C.0.5}
$$

L2 has the generalized t-distribution kernel.

Replace L2 back to the full likelihood

$$L_i = \int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i) \underbrace{\frac{\Gamma(\nu_0+\frac{1}{2})}{\Gamma(\nu_0)}\left(\frac{\nu_{ij}}{2\pi\sigma^2\nu_0}\right)^{\frac{1}{2}}\left[1+\frac{(y_{ij}-\mu_{ij})^2\nu_{ij}}{2\sigma^2\nu_0}\right]^{-(\frac{2\nu_0+1}{2})}}_{\text{P2: Generalized t-distritbution}} dF_{\boldsymbol{b}}(\boldsymbol{b}_i) \tag{C.0.6}$$

$P2$ is a generalized t distribution with degrees of freedom $df = 2\nu_0$, location parameter $\boldsymbol{\mu}_t = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$, and variance $\boldsymbol{\Sigma}_t$ that is a diagonal matrix with the diagonal elements $\boldsymbol{\Sigma}_{ii} = \nu_0\sigma^2/\nu_{ij}(\nu_0 - 1)$.

The full likelihood L is a normal-t mixture. Let $f^t(\cdot|df, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ stand for the probability density function (pdf) of a multivariate generalized t-distribution function with degrees of freedom df, mean $\boldsymbol{\mu}_t$ and variance $\boldsymbol{\Sigma}_t$. Let $f^N(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the pdf of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The likelihood function can be written as

$$\begin{aligned} L &= \int f^N_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f^t_{\boldsymbol{y}^m|\boldsymbol{b}}(\boldsymbol{y}_i^m|\boldsymbol{b}_i) dF_{\boldsymbol{b}}(\boldsymbol{b}) \\ &= \int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f^t_{\boldsymbol{y}^m|\boldsymbol{b}}(\boldsymbol{y}^m|2\hat{\nu}, \boldsymbol{\mu}_t^m, \boldsymbol{\Sigma}_t^m) dF_{\boldsymbol{b}}(\boldsymbol{b}), \end{aligned} \tag{C.0.7}$$

Here, $f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}_i^o|\boldsymbol{b}_i)$ has a normal distribution with mean $\boldsymbol{\mu}^o = \boldsymbol{X}^o\boldsymbol{\beta} + \boldsymbol{Z}^o\boldsymbol{b}$, and the covariance matrix for the multivariate normal distribution is $\boldsymbol{\Sigma}^o = \sigma^2(\boldsymbol{S}^o)^{-1}$. For the multivariate t-distribution, we have all the parameters $\boldsymbol{\mu}_t^m$ and variance matrix $\boldsymbol{\Sigma}_t^m$ as in $P2$.

We want to approximate the multivariate t-distribution with a multivariate normal distribution with same mean and variances, that is $f^t_{\boldsymbol{y}^m|\boldsymbol{b}}(\boldsymbol{y}^m|\boldsymbol{b}; 2\hat{\nu}_0, \hat{\boldsymbol{\mu}}_t^m, \hat{\boldsymbol{\Sigma}}_t^m) \approx f^N_{\boldsymbol{y}^m|\boldsymbol{b}}(\boldsymbol{y}^m|\boldsymbol{b}_i; \hat{\boldsymbol{\mu}}^m, \hat{\boldsymbol{\Sigma}}^m)$. The likelihood function can be approximate as

$$\tilde{L} = \int f_{\boldsymbol{y}^o|\boldsymbol{b}}(\boldsymbol{y}^o|\boldsymbol{b}_i) f^N_{\boldsymbol{y}^m|\boldsymbol{b}_i}(\boldsymbol{y}^m|\boldsymbol{b}_i; \hat{\boldsymbol{\mu}}^m, \hat{\boldsymbol{\Sigma}}^m) dF(\boldsymbol{b}_i) \tag{C.0.8}$$

In this case, it is easy to integrate out the random effect and we can have the likelihood function

$$\tilde{L} = f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y}|\boldsymbol{b}; \tilde{\boldsymbol{\mu}} = \boldsymbol{X}\boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}} = (\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \sigma^2\boldsymbol{S}^{-1})), \tag{C.0.9}$$

where $\boldsymbol{S} = (\boldsymbol{S}^o, \boldsymbol{S}^m)$. The likelihood follows a multivariate normal distribution with mean $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ and variance $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \sigma^2\boldsymbol{S}^{-1}$.

Let the covariance matrix of the likelihood function $\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma^o & \Sigma^{om} \\ \Sigma^{om} & \Sigma^m \end{pmatrix}$, where Σ^o and Σ^m are the variance matrices for $\boldsymbol{y}^o$ and $\boldsymbol{y}^m$, and Σ^{om} is the covariance matrix for the two. Note that the covariance Σ^{om} is non-zero since the outcome can have observations in both $\boldsymbol{y}^o$ and $\boldsymbol{y}^m$.

The likelihood function $\tilde{L}$ can be re-written as

$$\tilde{L} = f(\boldsymbol{y}^o) f(\boldsymbol{y}^m | \boldsymbol{y}^o), \tag{C.0.10}$$

where $f(\boldsymbol{y}^m|\boldsymbol{y}^o)$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}^{m|o} = \boldsymbol{X}^m\boldsymbol{\beta} + \Sigma^{om}\Sigma^{o^{-1}}(\boldsymbol{y}^o - \boldsymbol{X}^o\boldsymbol{\beta})$ and variance $\boldsymbol{\Sigma}^{m|o} = \Sigma^m - \Sigma^{om}\Sigma^{o^{-1}}\Sigma^{om}$. Therefore, we can write the likelihood function $\tilde{L}$ as

$$\tilde{L} = f_{\boldsymbol{y}^o}(\boldsymbol{y}^o) \underbrace{f_{\boldsymbol{y}^m|\boldsymbol{y}^m}(\boldsymbol{y}^m|\boldsymbol{y}^o; \hat{\boldsymbol{\mu}}^{m|o}, \hat{\boldsymbol{\Sigma}}^{m|o})}_{\text{conditional distribution}} \tag{C.0.11}$$

We will switch the conditional distribution back to the multivariate generalized t distribution with the same mean and variance and the likelihood function can be written as

$$\hat{\tilde{L}} = f_{\boldsymbol{y}^o}(\boldsymbol{y}^o) f^t_{\boldsymbol{y}^m|\boldsymbol{y}^o}(\boldsymbol{y}^m|\boldsymbol{y}^o; 2\hat{\nu}, \boldsymbol{\mu}^{m|o}, \boldsymbol{\Sigma}^{m|o}) \tag{C.0.12}$$

We assume that after the approximation, the likelihood function L is approximately equal to the new normal-t mixture $\hat{\tilde{L}}$