Smooth individual level covariates adjustment in disease mapping.

Md Hamidul Huque[1, 2, 3], Craig Anderson[2,3], Richard Walton[4], Samuel Woolford[5], and Louise Ryan[2,3]

[1]*Murdoch Childrens Research Institute, 50 Flemington Road, parkville, VIC 3052, Australia.*
[2]*School of Mathematical and Physical Sciences, University of Technology Sydney. 15 Broadway, Ultimo, NSW, 2007, Australia.*
[3]*Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers.*
[4]*Cancer Institute NSW, 8 Central Avenue, Eveleigh, NSW, 2015, Australia.*
[5]*Department of Mathematical Sciences, Bentley University, Waltham, MA 02452 USA.*

January 19, 2018

Abstract

Spatial models for disease mapping should ideally account for covariates measured both at individual and area levels. The newly available "indiCAR" model fits the popular conditional autoregresssive (CAR) model by accommodating both individual and group level covariates while adjusting for spatial correlation in the disease rates. This algorithm has been shown to be effective but assumes log-linear associations between individual level covariates and outcome. In many studies, the relationship between individual level covariates and the outcome may be non-loglinear, and methods to track such non-linearity between individual level covariate and outcome in spatial regression modeling are not well developed. In this paper, we propose a new algorithm, smooth-indiCAR, to fit an extension to the popular conditional autoregresssive model which can accommodate both linear and non-linear individual level covariate effects while adjusting for group level covariates and spatial correlation in the disease rates. In this formulation the effect of a continuous individual level covariate is accommodated via penalized splines. We describe a two-step estimation procedure to obtain reliable estimates of individual and group level covariate effects where both individual and group level covariate effects are estimated separately. This distributed computing framework enhances its application in the Big Data domain with a large number of individual/group level covariates. We evaluate the performance of smooth-indiCAR through simulation. Our results indicate that the smooth-indiCAR method provides reliable estimates of all regression and random effect parameters. We illustrate our proposed methodology with an analysis of data on neutropenia admissions in New South Wales (NSW), Australia.

1 Introduction

The rapid growth of Geographic Information Systems (GIS) together with the advances in high performance computing environments, presents a unique opportunity to examine the relationship between risk factors and outcomes that vary across geographical locations. Careful analysis of spatial data can lead to useful explanation of the exposure and disease relationship through natural experimentation where individuals (or clusters of individuals) exposed to the experimental and control conditions are determined by nature or by other factors outside the control of the investigators, but the process governing the exposures arguably resembles random assignment (33; 31). It also helps in understanding the spatial variation of disease, disease clustering, distribution of socio-demographic characteristics, environmental exposure distribution and its impact on health outcomes (9).

This is the peer reviewed version of the following article: Huque MH, Anderson C, Walton R and Ryan L. Smooth individual level covariates adjustment in disease mapping, , which has been published in final form at Biometrical Journal,DOI:10.1002/bimj.201700143. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving."

Analysis of spatially indexed data is complicated by correlations among neighboring observations (6; 3; 7). Regression analysis which ignores this spatial correlation may lead to incorrect inference on the estimated regression coefficients due to the narrowing of associated confidence intervals (35). Mixed effects models provide a convenient way of adjusting for spatial correlation by incorporating spatially defined random effects in the model, with the most commonly used approach being the conditionally autoregressive (CAR) model (5; 21). Using such models allows us to map disease rates by borrowing information about each small area from its surrounding areas, leading to more stable estimation. The use of CAR-based random effects within a hierarchical generalized linear model offers a robust, flexible, and enormously popular class of models for the exploration and analysis incorporating small area rates for disease mapping.

In the case study that motivates this paper, researchers from the New South Wales (NSW) Cancer Institute were interested in exploring the geographical variation of neutropenia admissions across NSW, Australia. Neutropenia is a life threatening complication of cancer chemotherapy and a major cause of morbidity and associated health care resource costs. Geographical variations in neutropenia admissions are of particular interest because of non-uniform health care services across NSW resulting from uneven population concentration (2). Moreover, neutropenia experiences for each patient might also depend on their age and cancer type, as treatment modalities often vary across different types of cancer and age groups. Therefore, appropriate analysis of geographical variation of neutropenia admissions requires adjustment of both patient demographic characteristics and covariates reflecting patients geographic location of residence. Exploratory analysis reveals a non-linear association between the observed neutropenia rates and patient age. The non-linear age effect has also been noted in many previous studies (30; 24). While the theory has been developed for CAR models with smooth group level covariates (22), there is a lack of algorithms to fit smooth individual level covariates for CAR models. The *glmm* function in *mgcv* package in *R* offers functionality to fit an AR(1) model, but there no software to accommodate both nonlinear individual and group level covariate effects while adjusting for more general spatial structure.

Recently, Huque et al. proposed an individual level covariate adjusted CAR (indiCAR) model which can incorporate both individual and area level covariates while adjusting for spatial random effects (16). Although this approach is very useful in modeling a large number of individual and group level covariate effects, it relies on an assumption of log-linear dependence between the expected value of the outcome and covariates. In many epidemiological study settings, such linearity of the covariate effect may not be appropriate and instead other types of non-linearity may be in operation. For example, maternal age has a non-linear effect on gestational age (18). Although transformation of variable is a well-known approach for handling non-linearity in regression model, often transformation is not known or may not be adequate to induce linearity (32). In our case, commonly available square root or log transformation of age was inadequate to induce linearity suggesting a need to use spline based techniques for modeling the effect of covariates in a flexible non-linear fashion .

Therefore, in our current study we extend the indiCAR method (16) to incoprporate smooth non-linear covariate effect using penalized splines, termed as smooth-indiCAR. This additional sets of penalized splines in the indiCAR method requires solution of a Double Penalized Quasi-Likelihood (DPQL) equation involving both individual and group level data in order to obtain an estimate of the regression coefficients. However, no software/algorithms are currently available. Following indiCAR, we incorporate individual level smooth covariate information in a two-step iterative procedure following an initialization step. In this method, the individual level and the group level covariate effects are fitted in separate iterations by appropriate calculation of an offset at each step. We illustrate that the estimation and inference based on smooth-indiCAR can be carried out in a distributed computing framework, thus achieving a helpful reduction in computational cost and memory requirements.

We evaluate the performance of the smooth-indiCAR method through simulation studies. Our results show that the smooth-indiCAR is able to correctly estimate coefficients associated with both individual and group-level covariates. We further illustrate this method through the analysis of data on neutropenia admissions from the NSW Cancer Institute and conclude with some practical guidelines.

The structure of the paper is as follows: Section 2 describes the data set. Model formulation and estimation procedure are given in section 3. Section 4 describes the inference procedure. Data generation process for our simulations and results using simulated data are presented in section 5. An application of the proposed method to data on neutropenia is given in section 6. We conclude with general discussion in section 7.

2 Data

Our study population comprises all cancer patients that were diagnosed with cancer and were hospitalized in NSW, Australia during the period between 2001 and 2009. Data from the NSW Central Cancer Registry, linked to the NSW Admitted Patient Data Collection, were used to identify patients diagnosed with cancer and their associated treatment procedures and co-morbidities. Detailed descriptions of the data items can be obtained from the Centre for Health Record Linkage (CHeReL http://www.cherel.org.au/ master-linkage-key). Data were checked for consistency across data sources and linked by assigning a unique Project Person Number to each patient.

Demographic variables including age at diagnosis, gender, residence at diagnosis, postal area of residence, and Accessibility/Remoteness Index of Australia (ARIA) based on patient residence were obtained from the Central Cancer Registry database. The ARIA variable was recorded at individual level rather than postal area level because the ARIA index varies within postal area. The Socio Economic Index For Areas (SEIFA; an index of social disadvantage) and the geo-coded shape files for mapping corresponding to 2006 census postal areas were obtained from the Australian Bureau of Statistics. Individual level clinical characteristics such as type of cancer were also obtained from Central Cancer Registry. The diagnosis of neutropenia admission and co-morbidity were obtained using data from the Admitted Patients Data Collection. The ICD-10-AM (International Statistical Classification of Disease and Related Health problems, 10th revision, Australian modification) code D70 (Agranulocytosis) was used to identify admissions with possible neutropenia.

3 Statistical model

Suppose the study area is divided into M contiguous regions and the number of neutropenia admissions for the i^{th} ($i = 1, 2, ..., n_j$) individual in the j^{th} ($j = 1, 2, ..., M$) area is denoted by $\{y_{ij}\}$. Let $\boldsymbol{Y} = (y_{11}, y_{21}, ..., y_{n_1 1}, ..., y_{1j}, y_{2j} ..., y_{n_j j}, ..., y_{1M}, y_{2M}, ..., y_{n_M M})^{\mathrm{T}}$ be a vector with elements $\{y_{ij}\}$ that represent the number of admissions for each individual in the study regions of interest. Similarly, let $\boldsymbol{X} = (X_1, X_2, ..., X_p)$ and $\boldsymbol{U} = (U_1, U_2, ..., U_q)$ represent individual and area level covariate matrices with dimensions $n \times p$ and $M \times q$, respectively, where n is the total sample size i.e., $n = \sum_{j=1}^{M} n_j$. Further suppose that in addition to the log-linear relationship of $\boldsymbol{X}$ and $\boldsymbol{U}$ with $\boldsymbol{Y}$, an additional individual level covariate, T exhibits a non-linear relationship with with the expectation of $\boldsymbol{Y}$. Under the above specifications, conditional on the area specific random effect vector, $\boldsymbol{b}$, the number of events for each cancer patient is assumed to be Poisson distributed with mean $\boldsymbol{\mu}$ where

$$\ln(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f}(T) + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}. \tag{1}$$

Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients associated with the individual level covariates, $\boldsymbol{f}(T)$ is an unknown smooth function, $\boldsymbol{\gamma}$ is a $q \times 1$ vector of area-specific regression coefficients and $\boldsymbol{Z} = \text{blockdiag}(Z_1, Z_2, ..., Z_M)$ is a replication matrix that replicates group level covariates and random effects to the individual level, where Z_j is a vector of length n_j with all elements equal to 1. We further assume that the unknown smooth function $\boldsymbol{f}(T)$ can be represented by a linear combination of spline basis functions, i.e., $\boldsymbol{f}(T) = \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu}$. Here $\boldsymbol{B}(T)$ is a vector of spline basis functions and $\boldsymbol{\nu}$ is a vector of corresponding basis coefficients. For simplicity in exposition we have only include a single non-linear covariate effect, however, the proposed method can be generalized to the case when more than one non-linear covariate effects is susceptible.

Note that the proposed model (1) represents various study designs, such as clustered, hierarchical and spatial designs depending on the specification of the random effect $\boldsymbol{b}$. For example, the random effect may represent specification for a) random slope and intercept for the multilevel (hierarchical) models (11), b) random intercept and stochastic process as of longitudinal studies (40) and c) modeling spatial correlation in disease mapping (21). Throughout this paper we will focus on modeling random effects so that they reflect spatial correlation. Our postulated model (1) is an extension of (22) that incorporates individual level predictors and area specific conditional auto-regressive random effects in the context of disease mapping literature.

To fit model (1), many different choices of random effects, $\boldsymbol{b}$ are available in the mapping literature (see (20) for a recent review). Among these, the method of Leroux el al. is appealing because it allows for a weighted combination of spatially structured and unstructured area-level variation (21). Within this

framework, the random effect vector, $\boldsymbol{b}$ has a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix, $\boldsymbol{D}$ with Moore-Penrose generalized inverse, $\boldsymbol{D}^{-} = \sigma^{-2}\{(1-\lambda)\boldsymbol{I} + \lambda\boldsymbol{R}\}$, where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{R}$ is the intrinsic auto-regression matrix reflecting neighborhood structure. Typically, neighbors are those areas which share a common boundary. The typical element of $\boldsymbol{R}$ is given by

$$\boldsymbol{R}_{jj'} = \begin{cases} m_j, & j = j' \\ -I\{j \sim j'\} & j \neq j', \end{cases}$$

where, m_j is the number of neighbors of region j, and $I\{j \sim j'\}$ is an indicator function that takes value 1 if regions j and j' are neighbors and 0 otherwise. Alternatively, a distance based neighborhood structure could be used (8). The parameters characterizing the random effect distribution, $\boldsymbol{\theta} = (\sigma^2 > 0, \lambda \in [0,1])$ quantify over-dispersion and spatial dependence, respectively. A larger value of $\lambda \in [0,1]$ indicates a higher degree of spatial dependence. This specification results in two extreme cases: i) completely independent random effects when $\lambda = 0$ and ii) the intrinsic auto-regressive model when $\lambda = 1$ (3). In general, a weighted combination of these two extreme (spatial independence and strong spatial dependence) is assumed (21).

Inference about $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ for model (1) can be made by integrating out or averaging over the distribution of the unobserved random effects, $\boldsymbol{b}$ when there are no non-linear predictors in model (1). The corresponding integrated quasi-likelihood function is equal to (see equation (2) of (5))

$$|\boldsymbol{D}|^{-\frac{1}{2}} \int \exp\left[-\frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{n_j} d_{ij}(Y_{ij}, \mu_{ij}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}^{-}\boldsymbol{b}\right] d\boldsymbol{b},$$

where $d(Y, \boldsymbol{\mu})$ is the deviance residual.

The maximum quasi-likelihood estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$ are those values of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ that maximize the above quasi-likelihood. However, no simple closed form solution exists. Instead, (5) proposed the penalized quasi-likelihood (PQL) approach for parameter estimation and inference. The PQL uses the Laplace method for integral approximation and jointly maximizes the above quasi-likelihood function to obtain estimates for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{b}(\boldsymbol{\theta})$.

In the presence of a non-linear predictor, however, statistical inference about $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ must account for the estimation of the basis coefficient, $\boldsymbol{\nu}$ and smoothing parameter δ (say). (22) showed that approximate estimates of the regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ can be obtained by maximizing the following Double Penalized Quasi-Likelihood equation with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{b}$ and $\boldsymbol{\nu}$:

$$-\frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{n_j} d_{ij}(Y_{ij}, \mu_{ij}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}}D^{-}\boldsymbol{b} - \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{\nu}, \tag{2}$$

where $\boldsymbol{S} = \delta\boldsymbol{K}$ with smoothing parameter δ and penalty matrix $\boldsymbol{K}$. Here, $\boldsymbol{K}$ is a $(q+N)\times(q+N)$ matrix where q is the number of knots and N is the dimension of the unpenalized function. Given the knot locations $\{x^*_{(k)} : k = 1, 2, ..., q\}$, the penalty matrix has zeros everywhere except in its lower right $q \times q$ block with $\boldsymbol{K}_{(ik)} = \|x^*_{j(i)} - x^*_{j(k)}\|^3$, for $k \leq q$. The penalty matrices map the spline basis functions to the data whereas the penalty parameters control the amount of smoothing (32; 37). For now, assume that the smoothing parameter δ is known.

Under the above specification the approximate log likelihood can be expressed as

$$\begin{aligned} const + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}) - \\ \mathbf{1}^{\mathrm{T}}\exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}^{-}\boldsymbol{b} - \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{\nu}, \end{aligned} \tag{3}$$

where $\mathbf{1}^{\mathrm{T}} = (1, 1, ..., 1)$, is a vector of 1's. Differentiating (3) with respect to $\boldsymbol{\beta}$, $\boldsymbol{\nu}$, $\boldsymbol{\gamma}$ and $\boldsymbol{b}$ using vector matrix calculus (36), we obtain the following score equations

$$\left\{\boldsymbol{Y} - \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\right\}^{\mathrm{T}}\boldsymbol{X} = 0, \tag{4}$$

$$\left\{\boldsymbol{Y} - \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\right\}^{\mathrm{T}}\boldsymbol{B}(T) = \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{S}, \tag{5}$$

$$\left\{\boldsymbol{Y} - \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\right\}^{\mathrm{T}} \boldsymbol{Z}\boldsymbol{U} = 0, \tag{6}$$

and

$$\left\{\boldsymbol{Y} - \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\right\}^{\mathrm{T}} \boldsymbol{Z} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}^{-}. \tag{7}$$

In principle, Penalized Iteratively Re-weighted Least Squares (P-IRLS) can be applied to solve the above equations for $\boldsymbol{\beta}$, $\boldsymbol{\nu}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ (37). However, lack of available software along with high computational costs and memory space constraints make it difficult to apply these iterative procedures to data sets with a large number of group level covariates and a large sample size. An alternative computational strategy is the use of the Gauss-Seidel algorithm to obtain the same estimates of the associated parameters as can be obtained by P-IRLS (13). In this approach, at each iteration one of the parameters is estimated while keeping others fixed at current values. Within this framework, we first initialize $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ and then obtain updated estimates for $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ in the following two step procedure:

Step 0: To initialize, set the coefficients of area level covariates, $\boldsymbol{\gamma}$ and random effects, $\boldsymbol{b}$ to zero in equation (4) and (5). Then we have

$$\left\{\boldsymbol{Y} - \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu})\right\}^{\mathrm{T}} \boldsymbol{X} = 0,$$

and,

$$\left\{\boldsymbol{Y} - \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu})\right\}^{\mathrm{T}} \boldsymbol{B}(T) = \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{S}.$$

If the value of penalty parameter δ is known, the solution of the above equations can be computed by a penalized version of the iterative re-weighted least square method used for GLM estimation (37; 12). The smoothing parameter can be estimated using the Generalized Cross Validation score (GCV) or the generalized Akaike's Information Criterion (37). Computationally, the estimation of the regression parameters $\boldsymbol{\nu}$ associated with non-linear function and smoothing parameter δ can be obtain using a penalized splines approach with the existing *gam* function in the *mgcv* package (37) in $\boldsymbol{R}$ (28). Thus we can obtain an estimate of the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ associated with individual level covariates and the penalty parameter δ. Although these parameters can also be estimated using penalized splines (P-splines/ B-splines) techniques, such an approach require careful selection of the number of knots and knot location. Instead *gam* function in the *mgcv* package uses thin plate regression splines as the default choice which doesn't require the estimation of knot locations and are computationally efficient. The choice of the number of knots (basis dimensions for the thin plate regression splines) is not generally critical for thin-plate regression splines as long as this is large enough to represents the degrees of freedom of the underlying true model.

This step provides initial estimates of the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$.

Step 1. Now substitute the current estimated individual level coefficients, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\nu}}$ into equations (6) and (7). With some simple algebra, we have

$$\{\boldsymbol{Y}_c - \exp(\mathrm{O}_1 + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{b})\}^{\mathrm{T}} \boldsymbol{U} = 0$$

and,

$$\{\boldsymbol{Y}_c - \exp(\mathrm{O}_1 + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{b})\}^{\mathrm{T}} = \boldsymbol{b}^{\mathrm{T}}D^{-},$$

where, $\boldsymbol{Y}_c^{\mathrm{T}} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Z}$, is a vector of aggregated outcome counts of length M at the group level and $\mathrm{O}_1 = \log\{\boldsymbol{Z}^{\mathrm{T}} \exp(\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{B}^{\mathrm{T}}(T)\widehat{\boldsymbol{\nu}})\}$ is a vector of offsets.

The above two equations are well known PQL estimating equations associated with the Poisson mixed model (5). Since, the outcome $\boldsymbol{Y}_c$, offset O_1, covariate $\boldsymbol{U}$ and random effects $\boldsymbol{b}$ are all available at the group level, estimates of parameters for the group level coefficient $\widehat{\boldsymbol{\gamma}}$ and random effects $\boldsymbol{\theta}$ can be estimated using the existing PQL method (5; 21) with only group level data. This results in substantial computational savings.

Step 2. Substitute the estimated area-specific regression coefficient, $\widehat{\boldsymbol{\gamma}}$ and random effect parameter, $\widehat{\boldsymbol{\theta}}$ estimated at Step 1 into (4) & (5). With some simple algebra, we have

$$\left\{\boldsymbol{Y} - \exp(\mathrm{O}_2 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu})\right\}^{\mathrm{T}} \boldsymbol{X} = 0,$$

and

$$\left\{\boldsymbol{Y} - \exp(\mathrm{O}_2 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{\nu})\right\}^{\mathrm{T}} \boldsymbol{B}(T) = \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{S},$$

where $\mathrm{O}_2 = \boldsymbol{Z}(\boldsymbol{U}\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{b}})$ is a offset vector of dimension $n \times 1$. Under the above specification, the individual level coefficient estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\nu}}$ and smoothing parameter δ can then be updated using the *gam* function with individual level data.

Step 1 and 2 are then repeated until the algorithm converges. The estimated coefficients and random effect parameters obtained by this iterative procedure will be similar (aside from rounding error) to the estimates of the true regression coefficients and random effect parameters which would be obtained by solving equations (3) - (7) directly.

4 Inference

Model fitting at Step 1 and Step 2 assumes fixed $(\boldsymbol{\beta}, \boldsymbol{\nu})$ and fixed $\boldsymbol{\gamma}$, respectively. Therefore the corresponding standard errors of $(\boldsymbol{\beta}, \boldsymbol{\nu})$ obtained from *gam* and $(\boldsymbol{\gamma}, \boldsymbol{\theta})$ obtained from the PQL method based on Step 2 and Step 1 will not be exactly correct. We re-calculate the standard error of these regression coefficients by adjusting the estimated standard errors of $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\nu}}$. This can be done via the IRLS estimation based on score equations (4, 5, 6 and 7) for a known smoothing parameter, δ (22). The IRLS estimation requires us to define a working dependent variable and a weight matrix that are updated at each iteration and solved via Fisher scoring (5; 37).

Now assume that an unpenalized linear combination of basis functions is adequate to represent the nonlinear function $\boldsymbol{f}(T)$. In this case the linear combination of the basis functions contributes to the log-likelihood equation (3) via the fixed effect components only. Hence the model (1) can be represented by a more general model, in which the conditional mean vector $\boldsymbol{\mu}$ is modeled via a log-link function as $\eta = \ln(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}$ and can be estimated using the PQL approach. The PQL fitting approach requires calculation of the GLM adjusted dependent variable $\boldsymbol{Y}_{pseudo} = \widehat{\eta} + (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})\frac{\delta\widehat{\eta}}{\delta\widehat{\boldsymbol{\mu}}}$ at each step of the iteration (26). The GLM adjusted dependent variable , $\boldsymbol{Y}_{pseudo}$ corresponding to (1) is given by

$$\boldsymbol{Y}_{pseudo} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \tag{8}$$

where $\boldsymbol{W}$ is a $n \times n$ diagonal matrix with diagonal term $\boldsymbol{\mu}$. Following (14) and (29), it can be shown that the Fisher scoring corresponding to the score equations (4, 5, 6 and 7) and GLM dependent variable as in (8), is identical to the normal equation of the best linear unbiased predictors (BLUPs) of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ corresponding to the following linear mixed model

$$\boldsymbol{Y}_{pseudo} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}(T)\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}_{pseudo},$$

where the pseudo-error, $\boldsymbol{\epsilon}_{pseudo}$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{W}^{-1}$. The estimated regression coefficients for the fixed effect, $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu})$ and BLUP estimate for the random effect $\boldsymbol{b}$ can be obtained as (29)

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\nu}}) = (\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{C})^{-1}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}_{pseudo})$$

and

$$\widehat{\boldsymbol{b}} = \boldsymbol{D}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{U}\widehat{\boldsymbol{\gamma}} - \boldsymbol{B}(T)\widehat{\boldsymbol{\nu}}\}, \tag{9}$$

where $\boldsymbol{C} = [\boldsymbol{X}|\boldsymbol{Z}\boldsymbol{U}|\boldsymbol{B}(T)]$ is a design matrix consisting of the individual level covariate matrix, group level covariates and basis functions, and $\boldsymbol{V} = \boldsymbol{Z}D\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{W}^{-1}$ is the variance of pseudo response $\boldsymbol{Y}_{pseudo}$.

Now consider the fact that the nonlinear function $\boldsymbol{f}(T)$ is represented using spline regression bases, with associated roughness penalties in the log-likelihood equation (3). Following (37) and (25), it can be shown that the maximum penalized likelihood estimate, $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\nu}})$ can be obtained as

$$\begin{aligned}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\nu}}) &= (\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{C} + \boldsymbol{S}_1)^{-1}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}_{pseudo}) \\ \widehat{\boldsymbol{b}} &= \boldsymbol{D}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{U}\widehat{\boldsymbol{\gamma}} - \boldsymbol{B}(T)\widehat{\boldsymbol{\nu}}\},\end{aligned} \tag{10}$$

where $\boldsymbol{S}_1$ is the smooth matrix consisting of 0s except in the block corresponding to the basis coefficients $\boldsymbol{\nu}$, where it is replaced by smoothing matrix $\boldsymbol{S}$. Thus, the frequentist variance-covariance matrix for the fixed effect ($\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\nu}}$) can be estimated by

$$\boldsymbol{Q}_{Freq} = (\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{C} + \boldsymbol{S}_1)^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{C}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{C} + \boldsymbol{S}_1)^{-1}. \tag{11}$$

However, (37) noted that the above estimated standard error for non-parametric functions is only useful when testing basis coefficients which are equal to zero. He also suggests the use of an alternative Bayesian approach to calculate uncertainty, which results in a Bayesian posterior covariance matrix for the parameters as

$$\boldsymbol{Q}_{Bayes} = (\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{C} + \boldsymbol{S}_1)^{-1}. \tag{12}$$

Note that the above frequentist and the Bayesian estimates of standard error differ in the inference about basis coefficients, but are virtually identical for linear individual and group level covariate effects. Further note that reliable estimates of the regression coefficients and variance components can also be obtained using the model of (22) with appropriate specification of the design matrix ($\boldsymbol{Z}$) associated with the spatial random effect model (1). However, their formulation requires the representation of smooth terms as a linear combination of fixed and random effect covariates. Back-fitting approaches such as the smooth-indiCAR method calculate the tuning parameters at Step 1 and will be effective in situations where memory constraints prohibit the fitting of a single model consisting of a large numbers of individual and group level covariates. Smooth-indiCAR not only provides a convenient way of fitting large numbers of individual and group level covariates in a distributed computing framework, it also allows us to calculate the standard error in a distributed computing framework. This is because $\boldsymbol{V}^{-1}$ can be expressed as $\boldsymbol{W} - \boldsymbol{W}\boldsymbol{Z}\boldsymbol{D}(I + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}$ (15). Therefore, the above Bayesian variance-covariance matrix can be written as

$$\boldsymbol{Q} = \left(\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \boldsymbol{S}_1 \right)^{-1},$$

where,

$$\begin{aligned}
a_{11} &= \bar{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\bar{\boldsymbol{X}} - \bar{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D}(\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D})^{-1} \times \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\bar{\boldsymbol{X}} \\
a_{12} &= \bar{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{U} - \bar{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D}(\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D})^{-1} \times \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{U} \\
a_{21} &= a_{12}^{\mathrm{T}} \\
a_{22} &= \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{U} - \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D} \times (\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{U}.
\end{aligned}$$

Here $\bar{\boldsymbol{X}} = [\boldsymbol{X}|\boldsymbol{B}(T)]$, is the design matrix combining individual level covariates and basis functions. Thus, among the various components of the above variance-covariance matrix, $\bar{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}$ are the only terms involving individual level data, and the rest of the terms involve a lower dimension corresponding to the group level data. Hence, upon convergence, calculation of the variance covariance matrix can also be carried out in a distributed computing framework for individual and group-level data separately.

However, the above standard errors for the non-linear function, $\boldsymbol{f}(T)$ rely heavily on the large sample assumption and treating the smoothing parameter as a known quantity (37). In reality, the smoothing parameter is estimated from the data, hence the confidence intervals for the non-linear function based on the standard errors as calculated above may not be appropriate. (25) proposed and developed an alternative confidence interval based on the above frequentist and Bayesian variance-covariance matrices. In this formulation the Bayesian and frequentist confidence intervals for the non-linear function, $\boldsymbol{f}(T)$ can be obtained as $\widehat{f}(T) \pm z_{\alpha/2}\sqrt{[\boldsymbol{V}_f]_{ii}}$, where $\widehat{f}(T) = \boldsymbol{B}^{\mathrm{T}}(T)\widehat{\boldsymbol{\nu}}$, $\boldsymbol{V}_f = \boldsymbol{B}(T)\boldsymbol{V}_{\boldsymbol{\nu}}\boldsymbol{B}^{\mathrm{T}}(T)$, and $z_{\alpha/2}$ is the critical value of the standard normal distribution with level of significance, α. Here, $\boldsymbol{V}_{\boldsymbol{\nu}}$ is the variance covariance matrix of $\widehat{\boldsymbol{\nu}}$ that can be obtained from the corresponding blocks of $\boldsymbol{Q}_{Freq}$ in (11) and of $\boldsymbol{Q}_{Bayes}$ in (12) in order to obtain frequentist and Bayesian confidence intervals, respectively. These confidence intervals are known to exhibit good coverage probabilities (39). In this paper, we have presented the smooth-indiCAR approach as a method of adjusting for a non-linear covariate term within our model. Here, the non-linear term is considered as a confounding variable, and the main interest lies in inference on other model parameters. More detailed analysis of this non-linear term, such as formally testing for equality to zero, is beyond the scope of this paper.

The covariance matrix for $\widehat{\boldsymbol{\theta}}$ was obtained from the Fisher information matrix from Step 1 in the usual way, assuming that the parameters for the individual and area specific covariates are fixed. Of course there is additional variability due to the fact that the individual and area specific covariate parameters are estimated. However, following (5) we ignore the additional variability due to estimation of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$ for inference about estimated random effect parameter, $\widehat{\boldsymbol{\theta}}$. Further details are given in Appendix A.

In the next section we describe a simulation study to evaluate the performance of the smooth-indiCAR method.

5 Simulation study

To evaluate our proposed smooth-indiCAR approach, we design a simulation study involving 400 regions in a 20×20 square lattice grid with varying sample sizes. To evaluate the smooth-indiCAR in both a large and a smaller sample setting, we divided a total of 20000 individuals (scenario i) and 5000 individuals (scenario ii) randomly among 400 areas. In this allocation, we ensured that each region contained at least one individual. We define two regions as neighbors if they share a common border. The random effects are then generated following a multivariate normal distribution with mean 0 and covariance matrix $D = [\sigma^{-2}\{(1-\lambda)\boldsymbol{I} + \lambda R\}]^{-1}$. The value of σ is set to 0.4 and five different values of spatial dependence parameters, $\lambda = \{0, 0.25, 0.50, 0.75, 0.99\}$ are considered in order to represent different strengths of spatial correlation ranging from the independent case (no spatial correlation, $\lambda = 0$, to a very high spatial correlation, $\lambda = 0.99$). We then generate three individual level covariates (one binary, one categorical and one continuous) and one group level covariate. The binary covariate represents the distribution of sex in the area and is generated following a Bernoulli random variable with probability ranging from 0.45 to 0.55 across groups. The categorical variable with five categories is generated with pre-specified probabilities. The continuous individual level covariate, T is generated using a univariate bump function as $f(t) = \frac{1}{1+t} - 2e^{-20(t-1)^2}$, to represent an age effect. The group level covariate is generated as a standard normal random variable. The outcome variable is then generated using model (1). The overall intercept of the model is set to zero in this simulation. This is because the overall intercept parameters of the model is not identifiable in the presence of non-linear function using *mgcv* package in R, as the current implementation assumes non-linear function decomposed into the intercept and linear combination of basis function (25). The full list of the parameters used to generate the simulated data is given in the header row of Table 1.

5.1 Simulation Results

Table 1 displays the average of the estimated regression coefficients for the linear individual level covariates (βs), group level covariate (γ) and parameters in the spatial random effects (σ and λ) corresponding to model (1) along with their estimated standard errors based on 500 simulation runs of scenario (i) with various specification of the spatial random effect. We calculate two different standard errors for the estimated regression coefficients: namely, (a) empirical standard errors i.e., taking the standard deviation of the 500 simulated regression coefficient estimates, (b) average of model based standard errors. The first column of Table 1 specifies the spatial dependence parameter (λ) used in that particular simulation. The second column represents the estimated coefficients corresponding to the binary variable, the next four columns list the estimated regression coefficients for the categorical individual level covariates. The last three columns list the estimated regression coefficients for the group specific covariate, estimated over-dispersion parameter and spatial dependence parameter. The true and estimated non-linear curves associated with scenario (i) were presented in Figure 1. The solid line represents the true non-linear curve and the dotted lines represent the estimated non-linear functions from the first 50 simulations. Only the first 50 simulations were plotted to enhance the visibility of each estimated graph around the true curve.

As expected, the smooth-indiCAR method provides reliable estimates of the individual level and region specific regression coefficients and the spatial random effect parameters in all of the spatial dependence settings considered. In general, the average of the estimated regression and random effect coefficients matches well with the true value of the coefficients used in the simulation study. Moreover, the estimated standard error matches well with the empirical standard error for the individual level and region specific regression coefficients. However, there is a slight underestimation of the standard errors for the spatial random effect parameters. The estimated non-linear functions also approximate the true non-linear function

well (Figure 1) while the variability of the fitted curve increases with the degree of the spatial dependence parameter.

To evaluate the performance of the proposed method in a small sample setting, we also conducted another simulation study with 5000 subjects distributed randomly in 400 regions (scenario (ii)). The estimates of the regression coefficients are given in Table 2. The true and estimated non-linear curves associated with scenario (ii) are given in Figure 2. As indicated in the Table 2 and Figure 2, the proposed method also performs well in the case when the number of individuals in a group is low.

We have also compared the computational time between the smooth-indiCAR and a generalized additive mixed model with cluster specific random intercept using *mgcv* package in R (37). Data were generated using two different spatial correlation parameters, $\lambda = 0$, which means that a random intercept only model is appropriate and $\lambda = 0.75$, which means that a CAR component is necessary for an accurate model fit. Varying sample sizes are considered both in 100 and 400 regions. One hundred simulation datasets were generated and fitted using smooth-indiCAR and GAMM with random intercept. The median computational time between these two approaches are presented in Table 4. It is clear for the table that in the case of both independent and spatially correlated data the GAMM with random intercept model is faster when the number of subjects within each group is low. However, with the increase of number of subjects within each group the smooth-indiCAR outperform the GAMM with random intercept model in terms of computational time.

6 Application to the neutropenia data

One of the key objectives of this analysis is to assess the geographical variation of neutropenia admission rates and its association with area level measures of socio-economic status. Data also includes patient age, gender, year of diagnosis, ARIA, cancer types at diagnosis, number of major comorbidities other than cancer and geographic location reported via postcode of residence. The descriptive data on the patients characteristics are given in the web Appendix Table 5. Briefly, data were collected from 279,623 cancer patients who received chemotherapy in NSW between 2001 and 2009. Among them only 4.5% were diagnosed with neutropenia. Exploratory analysis of our data using categorized values of age (not shown) suggested a non-linear association between age and the risk of chemotherapy-induced neutropenia. Indeed a plot of the predicted smooth curve and associated confidence bounds obtained by our method confirm this (see Figure 3), showing a sharp decline in neutropenia rates beyond age 65. We explored the use of transformation based approaches such as taking the log or square root of age instead (details obtainable from authors on request). However, none of these were able to adequately capture the pattern present in the data.

Table 3 reports the multivariable analysis of neutropenia admissions using the smooth-indiCAR and indiCAR method (16). The key difference between these two approaches is that the former includes a smooth term to capture non-linear age effect and the latter includes the age effect as linear (misspecified age effect). In general, the results are quite similar although the magnitude of the regression coefficients differ. The stronger effect for male, hematological malignant, and patients with higher co-morbidities as observed in smooth-indiCAR over the indiCAR might reflect the differences in age profile of patients across various categories of these covariates. Further analysis of neutropenia data using age as a categorical variable also exhibit very similar results as of smooth-indiCAR (result not shown). Thus smooth-indiCAR adequately captures the non-linear age effect as difference in estimates are captured more strongly in this approach compared with the indiCAR approach. Therefore, we believe the addition of a smooth non-linear term is a worthwhile and valuable extension of the indiCAR as modeling age as a categorical/linear variable might not represent the relationship as accurately. Moreover, categorization of age is often challenging as this may introduce residual confounding effect and result in spurious relationships between age and the outcome variable (31).

We also calculate standardized incidence ratios (SIR) corresponding to each postcode by dividing the observed number of neutropenia admissions to the model based expected number of neutropenia admissions (4). Figure 4 shows the distribution of estimated SIR following smooth-indiCAR and indiCAR approaches. We obtained similar distributions of estimated SIR of neutropenia admissions in NSW for the smooth-indiCAR method and indiCAR method. The strong spatial correlation after adjusting for individual and group specific covariates indicates that geographical variation of neutropenia might be due to differences in health care practices or access to care across NSW.

7 Discussion

In this paper, we have developed a framework to semi-parametrically adjust for a non-linear individual level covariate effect in spatial disease mapping. Our results suggest that smooth-indiCAR provides reliable estimates of the true regression parameters. One of the key advantage of the smooth-indiCAR approach is that this method can easily be adopted to situation where more than one non-linear covariate effect might be susceptible. This is due to the fact that the implementation of non-linear function in the smooth-indiCAR relies on the the existing *gam* function in the *mgcv* package which can readily accommodate more than one smooth terms. Moreover, this method has potential for Big Data implementations due to the natural applicability of the smooth-indiCAR method in a distributed computing framework . One of the key problems in Big Data analysis is to divide the data so that this division retains the inherent correlation structure of the data. Our proposed methodology provides a convenient way for such division by separating data according to the natural characteristics of the data, based on individual and group level covariates. The individual level covariate data can then be analyzed with the recent development of generalized additive models for large data sets (38). Thus our proposed smooth-indiCAR provides a convenient way to extend recent developments in Big Data for independent responses to the spatially correlated responses. This could also speed up the process and reduce computational costs.

Although the software implementation is quite similar between indiCAR and smooth-indiCAR except for the fact that the generalized linear model and the generalized additive model are required to fit the individual level data for indiCAR and smooth-indiCAR, respectively, the two approaches differ by the fact that smooth-indiCAR can incorporate the smooth function in the model to characterize the non-linear covariate effects. The inclusion of the smooth function in the model requires a solution of a double penalized quasi-likelihood equation. As a result the estimation of the regression parameters using these two approaches are quite different and current indiCAR algorithm is unable to handle this additional complexity. To obtain a solution of the double penalized quasi-likelihood equation, (22) suggested a mixed model based approach, however no software is currently available to adopt their approach. Therefore, without the theoretical development presented in this paper it was not straightforward to incorporate smooth-individual level covariates in the existing indiCAR model.

Health registries routinely collect geo-coded information for patient's residence at diagnosis and their individual level socio-demographic and clinical characteristics and thus could benefit by using our proposed method to incorporate individual level information in the analysis and mapping of disease rates. We illustrated the proposed approaches using the analysis of neutropenia data. By accounting for both individual-level and area-level effects, our model would represent an improvement on any analysis which focused on just one of these. The ability to incorporate individual level covariates in the disease mapping models provides an additional opportunity to investigate causal relationships without getting caught in the web of ecological fallacy (34). In our case we however do not have the individual level measurement of socio-economic status to estimate the true causal effect of individual socio-economic status on the neutropenia rates. Additional study that collect both individual and area level measures of exposure might provide better insights in regards to the ecological fallacy benefit of the smooth-indiCAR.

Likewise, accounting for age in a non-linear manner is likely to provide a better insight into age effects that appear to contradict conventional wisdom to some extent (e.g. increasing age appears to have a declining risk for neutropenia). Since we have the descriptive data indicating this relationship and our model is clearly able to capture the effect, this would clearly be a potential improvement over modeling age as a categorical variable which might not represent the relationship as accurately. Additional simulation suggests non-convergence of group level and random effect parameters when true non-linear covariates is misspecified as a categorical variable (results not shown in a table). Further confirmation of these results would need future research. There are also a number of areas where future study would be useful.

In our simulation study we note that we have slightly underestimated the spatial random effect parameters within our model. This is not a direct limitation of our smooth-indiCAR method, but instead a symptom of the limitations of the PQL approach when fitting spatial models with small expected counts (21). In most applications, including our neutropenia study, this bias is small when compared to the empirical standard deviation. However, we note that in cases where very small counts are present, such as the study of rare diseases, it may be more appropriate to adopt alternative inferential techniques such as a fully Bayesian approach (20).

In our application to the neutropenia admission data, we observed a lower risk of neutropenia admissions associated with increasing age, although advanced age has been identified as a significant predictor

for neutropenia admissions in previous studies (19). This might be due to the fact that the current risk based prophylactic administration of Colony Stimulating Factor guidelines account for patients advanced age (1). The lower than expected counts near the borders of NSW are almost certainly a result of some patients being admitted to hospitals in neighboring states. Therefore, the presence of edge effect is inevitable regardless of the choice of model. As from the map it is clear that the edge effect is mostly presented near the major metropolitan cities where better access to care is available. Without access to neutropenia data from these state, the edge effect cannot be fully ascertained. Although adjacency based weight matrix might influence edge effect but this is still the most commonly used weight matrix for spatial modelling. Various methods of calculation of weight matrix has been proposed in the literature (8), thus can be employed to explore in detailed. However, exploration of edge effect based on different weight matrices is beyond the scope of the present paper. We also identified that both the main regions with higher than expected counts are located in areas of higher population, and this pattern invites further investigation.

The dependence between area-specific neutropenia rates on ARIA and SEIFA are in the opposite direction. This is counter intuitive as remote areas in NSW are mostly associated with disadvantaged SEIFA categories. However, the observed contrast in estimated regression coefficients might be due to differences in the health care practices. Patients in the remote areas where the patients are geographically distant to the treating medical oncologist are best managed by their primary care physicians, and therefore may be treated with lower doses of chemotherapy (10). On the contrary, patients in the major cities might get intensive chemotherapy to treat them early, and are better managed due to availability of resources. Previous studies also indicate that remoteness has the greatest effect in quality of cancer treatment (17) and it affects treatment choices made by both patients and clinicians (27).

In our application, we observed a very high spatial dependence parameter estimate, $\hat{\lambda}$, despite that fact that we adjust for both individual and group level covariates in our analysis. In order to help explain the estimation of the high spatial autocorrelation parameter, following the reviewer suggestion, we considered an additional simulation study using the NSW as the geography consisting of 600 postcodes. A sample of 20000 subjects were considered for this simulation study. The results are very similar to that presented in Section 5 and are provided in Appendix Table 6. Although there are some variability in the estimated spatial regression parameters, the result suggests that the smooth-indiCAR method can adequately estimate the spatial dependence parameters in accordance with the strength of spatial correlation in the simulated data. The strong spatial correlation after adjusting for individual and group specific covariates as obtained from our neutropenia data analysis might indicates that geographical variation of neutropenia might be due other factors such as differences in health care practices, patients preferences or access to care across NSW. However, we do not have any data that measure these differences to examine this. Of note, (16) also reported similar estimates of the spatial dependence parameters following (21) approach that does not include individual level covariates.

Variation with respect to clinical practices in the treatment of neutropenia has been identified in Australia in a previous survey (23). This survey showed that the treatment approach for management of neutropenia varies across oncologists, hematologists and clinicians as well as different sectors of cancer care. Therefore, it might be interesting to explore whether the observed variation is due to variation across different hospitals (eg., metropolitan hospital vs. Non-metropolitan hospitals) in NSW or across various health care providers. However, data items for such analysis are not collected in the registry and are beyond the scope of our present paper.

Smooth-indiCAR is a useful addition to the existing methodology to explore clinical variation across geographical locations where covariates might have non-linear effects. One of the major advantages of our proposed method is the ability to obtain both individual and group level covariate effects when employing spatial regression models for disease mapping.

7.1 acknowledgement

MHH, CA and LR were supported by the University of Technology Sydney and by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). The authors thank the Cancer Institute NSW and the Ministry of Health for making the data available.

Conflict of Interest *The authors have declared no conflict of interest.*

References

[1] MS Aapro, J Bohlius, DA Cameron, Lissandra Dal Lago, J Peter Donnelly, Nora Kearney, GH Lyman, R Pettengell, VC Tjan-Heijnen, J Walewski, et al. 2010 update of EORTC guidelines for the use of granulocyte-colony stimulating factor to reduce the incidence of chemotherapy-induced febrile neutropenia in adult patients with lymphoproliferative disorders and solid tumours. *European Journal of Cancer*, 47(1):8–32, 2011.

[2] Australian Bureau of Statistics. Regional Population Growth, Australia, 2013-14 (cat. no. 3218.0), March 2015.

[3] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

[4] NE Breslow and NE Day. *Statistical Methods in Cancer Research. Volume II–The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Oxford University Press, New York, U.S.A., 1987.

[5] Norman Breslow and David Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

[6] David Clayton and John Kaldor. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681, 1987.

[7] Noel Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. John Wiley & Sons, New York, U.S.A., 1993.

[8] Arul Earnest, Geoff Morgan, Kerrie Mengersen, Louise Ryan, Richard Summerhayes, and John Beard. Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International Journal of Health Geographics*, 6(1):54–65, 2007.

[9] Paul Elliott and Daniel Wartenberg. Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives*, 112(9):998–1006, 2004.

[10] Peter Fox and Adam Boyce. Cancer health inequality persists in regional and remote Australia. *Medical Journal Australia*, 201(8):445–446, 2014.

[11] Andrew Gelman. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

[12] Peter J Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 55(3):245–259, 1987.

[13] Subharup Guha, Louise Ryan, and Michele Morara. Gauss–Seidel Estimation of Generalized Linear Mixed Models With Application to Poisson Modeling of Spatially Varying Disease Rates. *Journal of Computational and Graphical Statistics*, 18(4):818–837, 2009.

[14] David A Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[15] Harold V Henderson and Shayle R Searle. On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60, 1981.

[16] Md Hamidul Huque, Craig Anderson, Richard Walton, and Louise Ryan. Individual level covariate adjusted Conditional Autoregressive (indiCAR) model. *International Journal of Health Geographics*, 15(25):1–13, 2016.

[17] Katharine E Jong, David P Smith, Q Yu Xue, Dianne L O'Connell, David Goldstein, and Bruce K Armstrong. Remoteness of residence and survival from cancer in New South Wales. *Medical Journal of Australia*, 180(12):618–622, 2004.

[18] EE Kammann and Matthew P Wand. Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):1–18, 2003.

[19] Jean Klastersky, Marianne Paesmans, Edward B Rubenstein, Michael Boyer, Linda Elting, Ronald Feld, James Gallagher, Jorn Herrstedt, Bernardo Rapoport, Kenneth Rolston, et al. The multinational association for supportive care in cancer risk index: a multinational scoring system for identifying low-risk febrile neutropenic cancer patients. *Journal of Clinical Oncology*, 18(16):3038–3051, 2000.

[20] Duncan Lee. A comparison of conditional autoregressive models used in Bayesian disease mapping.

Spatial and Spatio-temporal Epidemiology, 2(2):79–89, 2011.

[21] Brian G Leroux, Xingye Lei, and Norman Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. Elizabeth Halloran and Donald Berry, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191. Springer, New York, 1999.

[22] Xihong Lin and Daowen Zhang. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):381–400, 1999.

[23] S Lingaratnam, MA Slavin, L Mileshkin, Benjamin Solomon, K Burbury, JF Seymour, R Sharma, B Koczwara, SW Kirsa, ID Davis, et al. An Australian survey of clinical practices in management of neutropenic fever in adult cancer patients 2009. *Internal Medicine Journal*, 41(1b):110–120, 2011.

[24] Ying C MacNab. Bayesian spatial and ecological models for small-area accident and injury analysis. *Accident Analysis & Prevention*, 36(6):1019–1028, 2004.

[25] Giampiero Marra and Simon N Wood. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53–74, 2012.

[26] Peter McCullagh and John A Nelder. *Generalized linear models*. Chapman and Hall., London, England., 1989.

[27] Ann Butler Nattinger, Ronald T Kneusel, Raymond G Hoffmann, and Mary Ann Gilligan. Relationship of distance from a radiotherapy facility and initial breast cancer treatment. *Journal of the National Cancer Institute*, 93(17):1344–1346, 2001.

[28] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[29] George K Robinson. That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

[30] Paul R Rosenbaum and Donald B Rubin. Difficulties with regression analyses of age-adjusted rates. *Biometrics*, 40(2):437–443, 1984.

[31] K.J. Rothman, S. Greenland, and T.L. Lash. *Modern Epidemiology, 3rd Edition*. Philadelphia, PA: Lippincott, Williams & Wilkins., U.S.A, 2008.

[32] D. Ruppert, P. Wand, and R.J. Carroll. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press., New York, U.S.A., 2003.

[33] John Snow. *On the mode of communication of cholera*. John Churchill, London, England., 1855.

[34] Jon Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183, 2007.

[35] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*, volume 368. John Wiley & Sons, New Jersey, U.S.A., 2004.

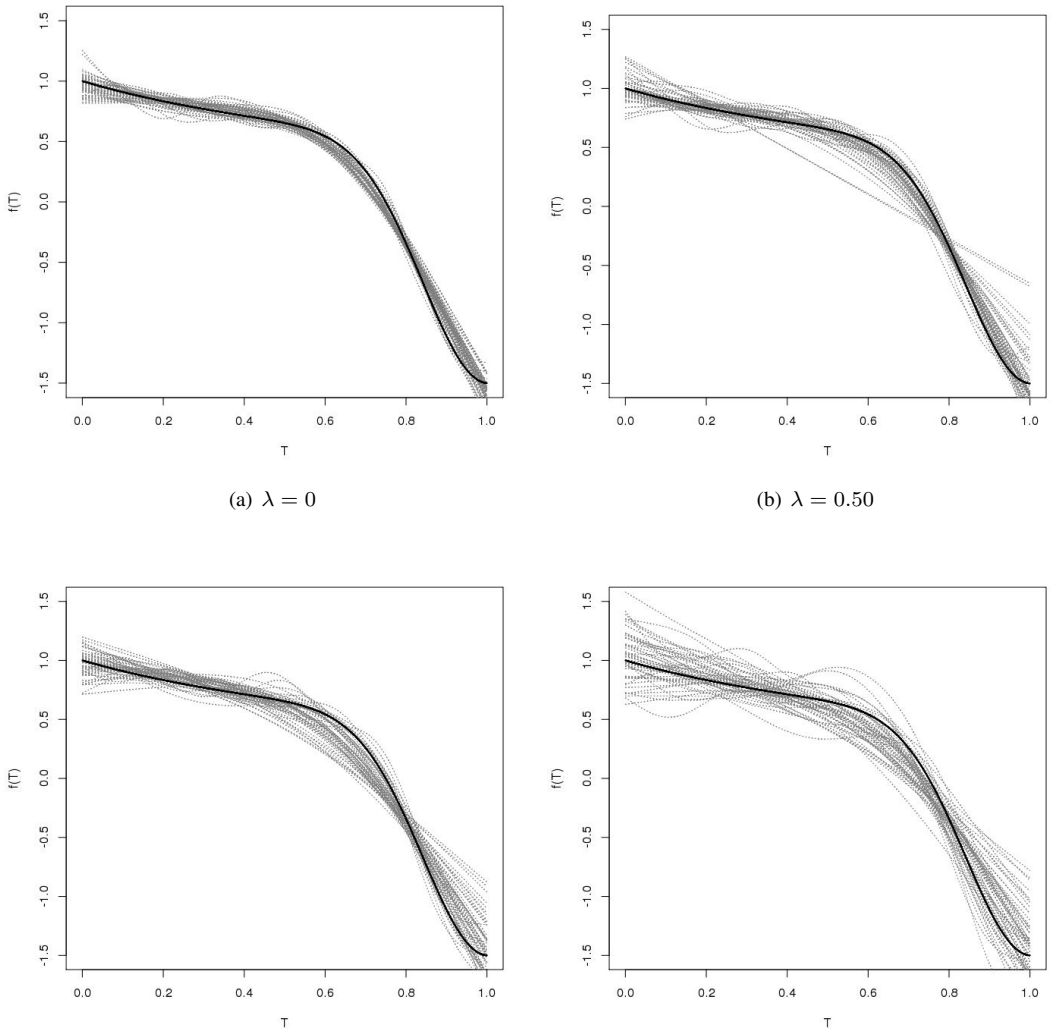
[36] MP Wand. Vector differential calculus in statistics. *The American Statistician*, 56(1):55–62, 2002.

[37] Simon Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC., Florida, U.S.A., 2006.

[38] Simon Wood, Yannig Goude, and Simon Shaw. Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155, 2015.

[39] Simon N Wood. On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228, 2013.

[40] Daowen Zhang, Xihong Lin, Jonathan Raz, and MaryFran Sowers. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93(442):710–719, 1998.



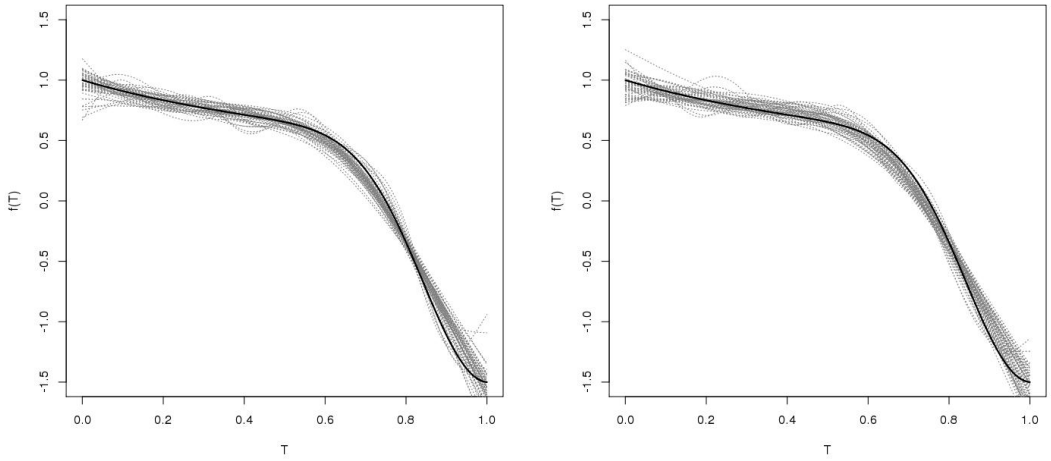
(a) $\lambda = 0$

(b) $\lambda = 0.50$

(c) $\lambda = 0.75$

(d) $\lambda = 0.99$

Figure 1: Fitted nonlinear curves based on first 50 simulations under scenario (i) with different values of spatial dependence parameter. The solid line indicates true curve.



(a) $\lambda = 0$ (b) $\lambda = 0.50$

(c) $\lambda = 0.75$ (d) $\lambda = 0.99$

Figure 2: Fitted nonlinear curves based on first 50 simulations under scenario (ii). The solid line indicates true curve.

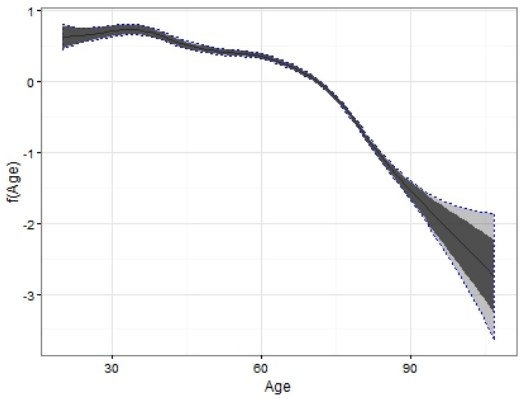
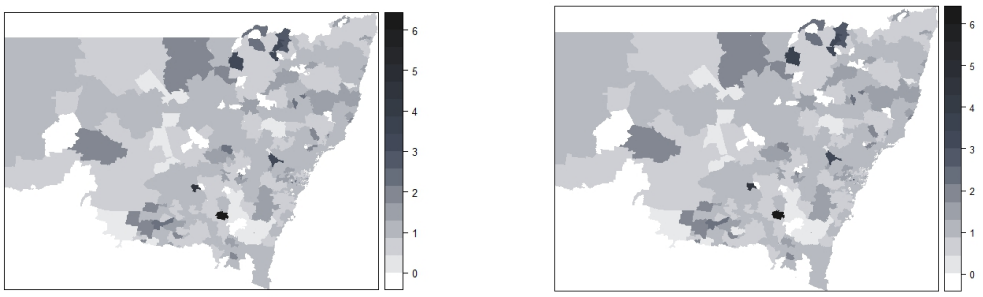


Figure 3: The estimated effect of age on neutropenia admission rates with associated 95 % Bayesian (light gray region) and frequentist (dark region) confidence intervals.



(a) Smooth-indiCAR. (b) indiCAR.

Figure 4: Distributions of estimated Standardized Incidence Ratios of neutropenia admissions in NSW, Australia following (a) smooth-indiCAR (b) indiCAR.

Table 1: Estimated regression coefficients and variance parameters for the proposed smooth-indiCAR method using simulation scenario (i).

True value	β_2 -2.00	β_{32} -1.50	β_{33} 0.15	β_{34} 0.50	β_{35} 0.20	γ 0.20	σ 0.40	λ
λ	Estimated coefficient							
0.00	-2.000	-1.499	0.150	0.501	0.201	0.198	0.391	0.021
0.25	-2.001	-1.499	0.150	0.500	0.200	0.197	0.392	0.249
0.50	-1.999	-1.500	0.149	0.499	0.199	0.198	0.392	0.496
0.75	-2.000	-1.499	0.150	0.500	0.202	0.200	0.391	0.723
0.99	-2.000	-1.502	0.147	0.497	0.197	0.198	0.392	0.932
	Empirical standard error							
0.00	0.020	0.053	0.030	0.028	0.029	0.023	0.036	0.035
0.25	0.020	0.053	0.030	0.029	0.030	0.018	0.041	0.133
0.50	0.020	0.055	0.031	0.028	0.029	0.016	0.042	0.189
0.75	0.020	0.052	0.032	0.029	0.030	0.015	0.034	0.164
0.99	0.020	0.053	0.032	0.031	0.030	0.014	0.024	0.066
	Average of the simulated standard error							
0.00	0.020	0.052	0.030	0.028	0.029	0.021	0.030	0.049
0.25	0.020	0.053	0.030	0.029	0.029	0.018	0.031	0.095
0.50	0.020	0.053	0.030	0.029	0.029	0.016	0.027	0.113
0.75	0.020	0.053	0.030	0.029	0.029	0.015	0.023	0.099
0.99	0.020	0.054	0.030	0.029	0.029	0.014	0.021	0.046

Table 2: Estimated regression coefficients and variance parameters for the proposed smooth-indiCAR method using simulation scenario (ii).

True value	β_2 -2.00	β_{32} -1.50	β_{33} 0.15	β_{34} 0.50	β_{35} 0.20	γ 0.20	σ 0.40	λ
λ	Estimated coefficient							
0.00	-1.999	-1.499	0.151	0.500	0.199	0.197	0.379	0.021
0.25	-1.999	-1.494	0.154	0.504	0.203	0.197	0.371	0.207
0.50	-2.000	-1.495	0.156	0.506	0.208	0.199	0.372	0.403
0.75	-2.004	-1.511	0.153	0.503	0.203	0.199	0.371	0.619
0.99	-2.004	-1.499	0.154	0.503	0.205	0.200	0.391	0.901
	Empirical standard error							
0.00	0.040	0.107	0.059	0.058	0.059	0.026	0.045	0.039
0.25	0.040	0.104	0.063	0.060	0.059	0.025	0.048	0.141
0.50	0.039	0.110	0.063	0.058	0.060	0.021	0.043	0.167
0.75	0.040	0.110	0.062	0.058	0.061	0.021	0.039	0.178
0.99	0.038	0.107	0.062	0.059	0.060	0.018	0.035	0.108
	Average of the simulated standard error							
0.00	0.040	0.106	0.061	0.058	0.059	0.025	0.040	0.070
0.25	0.040	0.107	0.062	0.059	0.060	0.022	0.041	0.127
0.50	0.040	0.108	0.062	0.059	0.060	0.021	0.037	0.161
0.75	0.040	0.108	0.062	0.059	0.060	0.020	0.033	0.154
0.99	0.040	0.108	0.062	0.059	0.060	0.019	0.032	0.067

Table 3: Comparison of estimated regression coefficients and variance parameters of smooth-indiCAR and indiCAR using neutropenia data.

	smooth-indiCAR		**indiCAR**	
Regression coefficients	Estimates	Std. Error	Estimates	Std. Error
Intercept	non-identifiable		-1.493	0.047
Sex				
Female	Ref		Ref	
Male	-0.091	0.020	-0.043	0.020
Age	see Figure 3		-0.027	0.001
Year of Diagnosis				
2001	Ref		Ref	
2002	0.016	0.038	0.021	0.038
2003	0.083	0.038	0.083	0.038
2004	0.023	0.038	0.019	0.038
2005	0.097	0.037	0.095	0.037
2006	0.038	0.038	0.038	0.038
2007	0.029	0.038	-0.022	0.038
2008	0.000	0.038	-0.004	0.038
2009	-0.313	0.042	-0.315	0.042
ARIA				
Major Cities	Ref		Ref	
Inner Regional Australia	-0.024	0.047	-0.006	0.022
Outer Regional Australia	-0.150	0.068	-0.118	0.037
Remote/ Very remote Australia	-0.244	0.163	-0.192	0.142
Cancer Type				
Breast Cancer	Ref		Ref	
Lung cancer	0.259	0.038	0.240	0.037
Colon & rectum cancer	-0.428	0.040	-0.463	0.040
Haematological Malignancy	1.579	0.029	1.497	0.029
Other cancer	-0.934	0.032	-0.986	0.031
No. of major comorbidities				
0	Ref		Ref	
1	0.424	0.026	0.413	0.026
2	0.680	0.026	0.682	0.026
3	0.625	0.036	0.589	0.036
4+	0.623	0.035	0.594	0.035
SEIFA				
Most disadvantaged	Ref		Ref	
2	-0.082	0.044	-0.089	0.043
3	-0.070	0.041	-0.078	0.038
4	-0.125	0.047	-0.134	0.045
Least disadvantaged	-0.129	0.056	-0.144	0.053
Variance parameter				
σ	0.203	0.022	0.209	0.022
λ	0.992	0.012	0.992	0.012

Table 4: Comparison of estimated time (in seconds) when data are generated with varying spatial random effect parameter, $\boldsymbol{\lambda}$ and sample sizes.

		Time to convergence (in seconds)					
		$\boldsymbol{\lambda} = 0.0$			$\boldsymbol{\lambda} = 0.75$		
Number of groups	Total sample	Smooth-indiCAR (T1)	GAMM with random intercept (T2)	Relative time T1/T2	Smooth-indiCAR (T3)	GAMM with random intercept (T4)	Relative time T3/T4
100	20000	22.4	14.8	1.5	22.3	15.8	1.4
	50000	53.0	43.2	1.2	52.7	42.6	1.2
	100000	98.5	116.1	0.8	106.2	122.4	0.9
	200000	203.7	521.1	0.4	122.8	372.4	0.3
400	20000	35.6	13.9	2.6	24.6	13.4	1.8
	50000	45.9	35.2	1.3	44.3	35.0	1.3
	100000	80.8	76.3	0.1	76.9	74.5	1.0
	200000	211.8	185.7	1.1	215.1	182.9	1.2

Appendix *(Supplementary Materials)*

Appendix A: Implementation of PQL in Step 2

The PQL estimation procedure is an iterative approach where at each step one must define a working dependent variable and a weight matrix which are then updated at each iteration and solved via Fisher scoring (5; 21). The detailed procedure has been illustrated elsewhere (5; 21).

The GLM adjusted dependent variable ($\boldsymbol{Y}_{c-pseudo}$) at the group level is calculated as

$$\boldsymbol{Y}_{c-pseudo} = \hat{\boldsymbol{\eta}}_c + (\boldsymbol{Y}_c - \hat{\mu}_c)\frac{d\hat{\boldsymbol{\eta}}_c}{d\hat{\mu}_c}. \tag{13}$$

Here, $\boldsymbol{\eta}_c = g(\mu_c) = \mathrm{O}_1 + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{b}$ and $\mathrm{O}_1 = \boldsymbol{Z}^{\mathrm{T}} \log\{\exp(\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{B}^{\mathrm{T}}(T)\widehat{\boldsymbol{\nu}})\}$ is an offset vector with dimension $M \times 1$. The Poisson link $g(\mu_c) = \log \mu_c$ and variance function $V(\mu_c) = \mu_c$ are used. The covariance matrix of $\boldsymbol{Y}_{c-pseudo}$ is then approximated by

$$\widehat{V}_c = \widehat{W}_c^{-1} + \widehat{D}, \tag{14}$$

where $\hat{D}$ is the covariance matrix of the random effects, $\boldsymbol{b}$, evaluated at the current estimate for the variance parameters, and $\hat{W}_c$ is the $M \times M$ diagonal matrix with diagonal terms $w = \hat{\mu}_c$. Updated estimates of the fixed effect vector $\boldsymbol{\gamma}$ and random effect vector $\boldsymbol{b}$ are then obtained from the solution of the following mixed model equations:

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{U}^{\mathrm{T}}\widehat{V}_c^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}\widehat{V}_c^{-1}(\boldsymbol{Y}_{c-pseudo} - \mathrm{O}_1), \tag{15}$$

and

$$\widehat{\boldsymbol{b}} = \hat{D}\widehat{V}_c^{-1}(\boldsymbol{Y}_{c-pseudo} - \mathrm{O}_1 - \boldsymbol{U}\widehat{\boldsymbol{\gamma}}). \tag{16}$$

The updated estimates of the variance parameters, λ and σ are obtained by a Newton-Raphson iterative procedure as follows:

$$\begin{pmatrix} \widehat{\sigma} \\ \widehat{\lambda} \end{pmatrix}^{updated} = \begin{pmatrix} \widehat{\sigma} \\ \widehat{\lambda} \end{pmatrix}^{old} + \boldsymbol{I}^{-1}\boldsymbol{S}. \tag{17}$$

where $\boldsymbol{S}$ is the score vector and $\boldsymbol{I}$ is the expected information matrix based on REML likelihood for $\boldsymbol{Y}_{c-pseudo}$. Letting $\boldsymbol{\theta}^{\mathrm{T}} = (\sigma, \lambda)$, the expression for the element of score vector and information matrix, are then given by

$$\begin{aligned}
\boldsymbol{S}_\sigma &= \frac{1}{2}(\boldsymbol{Y}_{c-pseudo} - \boldsymbol{U}\widehat{\boldsymbol{\gamma}} - \mathrm{O}_1)^{\mathrm{T}} P \frac{\delta V_c}{\delta \sigma} P(\boldsymbol{Y}_{c-pseudo} - \boldsymbol{U}\widehat{\boldsymbol{\gamma}} - \mathrm{O}_1) - \frac{1}{2} tr\left(P\frac{\delta V_c}{\delta \sigma}\right) \\
\boldsymbol{S}_\lambda &= \frac{1}{2}(\boldsymbol{Y}_{c-pseudo} - \boldsymbol{U}\widehat{\boldsymbol{\gamma}} - \mathrm{O}_1)^{\mathrm{T}} P \frac{\delta V_c}{\delta \lambda} P(\boldsymbol{Y}_{c-pseudo} - \boldsymbol{U}\widehat{\boldsymbol{\gamma}} - \mathrm{O}_1) - \frac{1}{2} tr\left(P\frac{\delta V_c}{\delta \lambda}\right)
\end{aligned}$$

and

$$I_{\sigma\lambda} = -\frac{1}{2} tr\left(P\frac{\delta V_c}{\delta \sigma} P \frac{\delta V_c}{\delta \lambda}\right),$$

where, $P = V_c^{-1} - V_c^{-1}\boldsymbol{U}(\boldsymbol{U}^{\mathrm{T}}V_c^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}^{\mathrm{T}}V_c^{-1}$. The derivatives of V_c with respect to σ and λ are:

$$\begin{aligned}
\frac{\delta V_c}{\delta \sigma} &= 2\sigma \boldsymbol{R}_\lambda^{-1} \\
\frac{\delta V_c}{\delta \lambda} &= \sigma^2 \boldsymbol{R}_\lambda^{-1}(\boldsymbol{R} - \boldsymbol{I})\boldsymbol{R}_\lambda^{-1},
\end{aligned}$$

where $\boldsymbol{R}_\lambda = (1-\lambda)\boldsymbol{I} + \lambda\boldsymbol{R}$ and $\boldsymbol{R}$ is the intrinsic autoregression matrix determined by neighbourhood structure . The typical element of $\boldsymbol{R}$ is given by

$$\boldsymbol{R}_{ij} = \begin{cases} n_i, & i = j \\ -I\{i \sim j\} & i \neq j. \end{cases}$$

Here, n_i is the number of neighbours of region i, and $I\{i \sim j\}$ is an indicator function indicating whether regions i and j are neighbours.

Repeated iterations of equations (13)-(17) are carried out, leading to reliable estimates of the region specific fixed effect and random effect parameters. Convergence is achieved when the changes in parameter estimates are less than a prespecified tolerance level (less than $1e-3$, in the simulation study reported). Approximate standard errors for λ and σ are obtained from the above information matrix in the usual way.

Appendix B: Additional Table

Table 5: Descriptive analysis of neutropenia data

Variables	Neutropenia n (%)	Total
Age (mean± sd)	59.6 (14.3)	64.9 (14.5)
Sex		
Female	6,363 (5.0)	127,519
Male	6,298 (4.1)	152,104
Year of Diagnosis		
2001	1,343 (4.9)	27,356
2002	1,411 (5.0)	28,451
2003	1,503 (5.1)	29,560
2004	1,478 (4.8)	30,970
2005	1,596 (5.1)	31,533
2006	1,452 (4.6)	31,865
2007	1,453 (4.5)	32,603
2008	1,405 (4.2)	33,343
2009	1,020 (3.0)	33,942
ARIA		
Major Cities	9,199 (4.9)	189,322
Inner Regional Australia	2,638 (3.9)	67,086
Outer Regional Australia	774 (3.6)	21,664
Remote or Very remote Australia	50 (3.2)	1,551
Cancer Type		
Breast Cancer	2,059 (5.3)	38,620
Lung cancer	1,401 (6.2)	22,744
Colon & rectum cancer	1,011 (3.0)	34,018
Haematological Malignancy	5,134 (25.0)	20,518
Other cancer	3,056 (1.9)	163,723
No. of major comorbidities		
0	6,072 (3.7)	163,645
1	2,228 (4.9)	45,817
2	2,315 (6.7)	34,670
3	976 (5.7)	17,264
4+	1,070 (5.9)	18,227
SEIFA		
Most disadvantaged	1,388 (4.6)	30,302
2	1,750 (4.1)	42,558
3	3546 (4.5)	78,006
4	2800 (4.6)	60,880
Least disadvantaged	3177 (4.7)	67,877

Table 6: Estimated regression coefficients and variance parameters for the proposed smooth-indiCAR method using NSW as geography and 20000 sample sizes.

True value	β_2 -2.00	β_{32} -1.50	β_{33} 0.15	β_{34} 0.50	β_{35} 0.20	γ 0.20	σ 0.40	λ
λ	Estimated coefficient							
0.00	-1.988	-1.498	0.146	0.492	0.199	0.205	0.356	0.002
0.25	-1.985	-1.490	0.147	0.496	0.205	0.206	0.369	0.220
0.50	-1.986	-1.488	0.151	0.497	0.204	0.195	0.355	0.405
0.75	-1.992	-1.496	0.113	0.443	0.183	0.201	0.366	0.625
0.99	-1.991	-1.467	0.124	0.445	0.187	0.211	0.378	0.852
	Empirical standard error							
0.00	0.022	0.057	0.032	0.031	0.031	0.018	0.015	0.016
0.25	0.022	0.059	0.033	0.032	0.032	0.015	0.017	0.049
0.50	0.022	0.058	0.033	0.031	0.032	0.014	0.018	0.079
0.75	0.022	0.058	0.033	0.031	0.031	0.013	0.020	0.101
0.99	0.025	0.064	0.036	0.034	0.035	0.013	0.022	0.079
	Average of the simulated standard error							
0.00	0.020	0.056	0.029	0.036	0.031	0.011	0.018	0.006
0.25	0.020	0.062	0.030	0.032	0.031	0.009	0.017	0.047
0.50	0.020	0.057	0.030	0.035	0.031	0.009	0.015	0.070
0.75	0.021	0.033	0.018	0.017	0.018	0.010	0.014	0.061
0.99	0.022	0.049	0.025	0.026	0.027	0.010	0.016	0.056

Appendix C: R Codes

Data simulation

```
library(MASS)
library(plyr)
library(reshape)
library(mgcv)
DataSimulation<-function(nGroup=400,sigma=0.4,lambda=0.75, beta0=0, beta1=1,
beta2=-2.0, beta32=-1.5,beta33=0.15,beta34=0.5,beta35=0.2, gamma=0.2)
{
#        nGroup: Total number of group
#        lambda: Spatial range parameters
#        sigma: variance parameter
#        beta0-beta3: coefficients of individual level covariates
#        gamma: coefficient of group level covariates
#        Randomly generate number of individuals within a group
repeat{
indivGroup<-as.vector(rmultinom(n=1, size=20000,prob=runif(400,0.05)))
if (all(indivGroup)>=1){break}
}
totalSample<-sum(indivGroup)
#generate covariate values
# Generate x2 as a binary variable to represent sex in the data
pr =runif(totalSample,0.45,0.55)
x2<-rbinom(totalSample,1,pr)
# Generate smooth covariate, f1
x1<-sort(runif(totalSample))
f <-function(x) 1/(1+x) - 2*exp(-20*(x-1.0)^2)
f1<-f(x1)
# Generate categorical variable to represent age group.
x3<-sample(1:5,totalSample, replace=TRUE, prob=c(0.06,0.09,0.19,0.25,0.25))
#generate group level covariate
z2<-rnorm(nGroup)
x4<-rep(z2,indivGroup)
# generate postcode ID's
ID<-rep(1:nGroup,indivGroup)
# GroupData
GroupData<-as.data.frame(cbind(1:nGroup,z2))
names(GroupData)<-c("ID","x4")
#Generating spatial random effect in a lattice grid
#### Set up a square lattice region
x.easting <- 1:20
x.northing <- 1:20
Grid <- expand.grid(x.easting, x.northing)
n <- nrow(Grid)
#### set up distance and neighbourhood matrices
distance <-array(0, c(n,n))
W <-array(0, c(n,n))
for(i in 1:n)
{
for(j in 1:n)
{
temp <- (Grid[i,1] - Grid[j,1])^2 + (Grid[i,2] - Grid[j,2])^2
distance[i,j] <- sqrt(temp)
if(temp==1)  W[i,j] <- 1
}
}
R <- -W
diag(R) <- as.numeric(apply(W, 1, sum))
Sigma.inv<-1/sigma^2*(lambda*R + (1-lambda)*diag(rep(1,n)))
Sigma<-solve(Sigma.inv)
phi <- mvrnorm(n=1, mu=rep(0,n), Sigma=Sigma)
phi_long<-rep(phi,indivGroup)
mu <- exp(beta0 + beta1*f1+beta2*x2+beta32*I(x3==2)+beta33*I(x3==3)+
beta34*I(x3==4)+beta35*I(x3==5)+gamma*x4+phi_long)
#generate Y-values
y <- rpois(totalSample, lambda=mu)
#data set
data <- data.frame(ID,y=y, x1,x2,x3,f1)
output<-list(data=data,R=R,W=W,GroupCovData=GroupData)
output
}
```

Implementation of smooth-indiCAR

```
set.seed(12345)
# Obtain individuallevel data, group level data and adjacency based weight matrices
# Note: In real data analysis these datasets and matrices might come from different
# sources, hence needs to be read in memory using read.table () or similar function
DataSim<-DataSimulation()
# Individual level outcome and covariate data
Data<-DataSim$data
# Obtain Grouplevel data based on simulation
GroupCovData<-DataSim$GroupCovData
# Spatial correlation matrix
R <- DataSim$R
#Use of generalized linear model with individual level covariates
# Fit individual level outcome with individual level covariates
fit_ind<-gam(y~s(x1,k=10)+x2+as.factor(x3), data=Data,family="poisson")
beta <- coef(fit_ind)
#Calculate the fitted value
Data$Predict<-exp(predict(fit_ind))
#Aggregate data Over postcode
AreaData<-Data[,c("y","ID","Predict")]
aggdata <-aggregate(. ~ ID, data = AreaData, sum)
nGroup<-nrow(aggdata)
GroupData<-merge(GroupCovData,aggdata,by="ID")
names(GroupData)<-c("ID","x4","totalY","totalePredict")
#Fitting PQL model
gamma.iter<-NULL
theta.iter<-NULL
beta.iter<-NULL
#Set initial values
gamma.hat<-0
b.rand<-rep(0,nGroup)
sigma.hat <-0.5
lambda.hat <-0.5
theta.hat<-c(sigma.hat,lambda.hat)
betaCombined<-c(beta,gamma.hat)
X_group <-as.matrix(GroupData$x4)
Y<-as.matrix(GroupData$totalY)
OffSet<-as.matrix(log(GroupData$totalePredict))
repeat{
repeat{
# Estimate covariance matrices
R.lambda <- lambda.hat*R + (1-lambda.hat)*diag(rep(1,nGroup))
R.lambda.inv<-solve(R.lambda)
D.hat.inv<-1/(sigma.hat^2)*R.lambda
D.hat<-solve(D.hat.inv)
# Calculate the PQL elements
repeat{
eta.est<-OffSet+X_group%*%gamma.hat+b.rand
mu.est<-exp(eta.est)
zz.est<-eta.est+(Y-mu.est)/mu.est
W.hat<-diag(as.vector(mu.est))
W.hat.inv<-diag(as.vector(1/mu.est))
#Estimate covariance matrix of Y
V.hat<-W.hat.inv+D.hat
V.hat.inv<-solve(V.hat)
#Estimate the fixed and random effect
gammaUpdate<-solve(t(X_group)%*%V.hat.inv%*%X_group)%*%
(t(X_group)%*%V.hat.inv%*%(zz.est-OffSet))
b.rand.update<-D.hat%*%V.hat.inv%*%(zz.est-OffSet-X_group%*%gammaUpdate)
diff<-abs(gammaUpdate-gamma.hat)
diff.rand<-abs(b.rand.update-b.rand)
gamma.iter<-rbind(gamma.iter,gammaUpdate)
if (all(diff< 1e-5) & all(diff.rand< 1e-3)){break}
gamma.hat<-gammaUpdate
b.rand<-b.rand.update
names(gamma.hat)<-c("gamma")
}
# Extract score and observed information matrix
P<-V.hat.inv-(V.hat.inv%*%X_group%*%solve(t(X_group)%*%V.hat.inv%*%X_group)%*%
t(X_group)%*%V.hat.inv)
dV.sigma<-2*sigma.hat*R.lambda.inv
dV.lambda<--1*sigma.hat^2*R.lambda.inv%*%(R-diag(rep(1,nGroup)))%*%R.lambda.inv
#Score vector
```

```
score.sigma<-0.5*(t(zz.est-OffSet-X_group%*%gammaUpdate)%*%P)%*%dV.sigma%*%
(P%*%(zz.est-OffSet-X_group%*%gammaUpdate))-0.5*sum(diag(P%*%dV.sigma))
score.lambda<-0.5*(t(zz.est-OffSet-X_group%*%gammaUpdate)%*%P)%*%dV.lambda%*%
(P%*%(zz.est-OffSet-X_group%*%gammaUpdate))-0.5*sum(diag(P%*%dV.lambda))
score<-c(score.sigma,score.lambda)
exp.inf11<-0.5*sum(diag(P%*%dV.sigma%*%P%*%dV.sigma))
exp.inf12<-0.5*sum(diag(P%*%dV.sigma%*%P%*%dV.lambda))
exp.inf21<-0.5*sum(diag(P%*%dV.lambda%*%P%*%dV.sigma))
exp.inf22<-0.5*sum(diag(P%*%dV.lambda%*%P%*%dV.lambda))
exp.infor<-matrix(c(exp.inf11,exp.inf12,exp.inf21,exp.inf22),ncol=2)
thetaUpdate<-theta.hat+solve(exp.infor)%*%score
if (thetaUpdate[2]>1) {thetaUpdate[2]<-0.99999}
if (thetaUpdate[2]<0) {thetaUpdate[2]<-0.00001}
if (thetaUpdate[1]<0) {thetaUpdate[1]<-0.00001}
sigma.hat<-thetaUpdate[1]
lambda.hat<-thetaUpdate[2]
diff.theta<-abs(thetaUpdate-theta.hat)
theta.iter<-rbind(theta.iter,as.vector(thetaUpdate))
if (all(diff.theta< 1e-5)){break}
theta.hat<-thetaUpdate
names(theta.hat)<-c("sigma","lambda")
}
# Repeat individual level data fitting
GroupData$Predict_group<-X_group%*%gamma.hat+b.rand
combinedData<-merge(Data,GroupData[,c("ID","Predict_group")],by=c("ID"))
fit_ind<-gam(y~s(x1,k=10)+x2+as.factor(x3),offset=Predict_group,
data=combinedData,family="poisson")
#Calculate the fitted value
betaUpdate<-coef(fit_ind)
#Calculate the fitted value
combinedData$Predict<-exp(predict(fit_ind))
AreaData<-combinedData[,c("y","ID","Predict")]
aggdata <-aggregate(. ~ ID, data = AreaData, sum)
GroupData<-merge(GroupCovData,aggdata,by="ID")
names(GroupData)<-c("ID","x4","totalY","totalePredict")
Y<-as.matrix(GroupData$totalY)
OffSet<-as.matrix(log(GroupData$totalePredict))
betaCombined.Update<-c(betaUpdate,gamma.hat)
diff.est<-abs(betaCombined.Update-betaCombined)
beta.iter<-rbind(beta.iter,betaCombined.Update)
if (all(diff.est< 1e-5)) {break}
betaCombined<-betaCombined.Update
# End of Our method
}
#Estimation of standard error
# Extract design matrix
G<-gam(y~s(x1,k=10)+x2+as.factor(x3),offset=Predict_group, fit=FALSE,
data=combinedData,family="poisson")
X_bar<-G$X
#Extract smoothing vector and parameter
lambda<-as.vector(fit_ind$sp)
K<-G$S[[1]]
S<-lambda*K
M<-nGroup
group.size<-as.vector(table(combinedData$ID))
group.cov.long<-data.matrix(X_group[rep(1:nrow(X_group), times = group.size), ])
cov.combined<-cbind(X_bar,group.cov.long)
fitted.combined<-cov.combined%*%betaCombined+rep(b.rand,group.size)
mu.combined<-as.vector(exp(fitted.combined))
Xcov.m<-X_bar*mu.combined
XTWX<-t(Xcov.m)%*%X_bar
groupID<-rep(c(1:M),group.size)
XTWZ<-t(rowsum(Xcov.m, groupID))
XTWZD<-XTWZ%*%D.hat
ZTWZ.vec<-aggregate(mu.combined,by=list(groupID),sum)
ZTWZ<-diag(ZTWZ.vec$x)
ZTWZD<-ZTWZ%*%D.hat
I.ZTWZD<-(diag(M)+ZTWZD)
I.ZTWZD.inv<-solve(I.ZTWZD)
a11<-XTWX-XTWZD%*%I.ZTWZD.inv%*%t(XTWZ)
XTWZU<-XTWZ%*%X_group
ZTWZU<-ZTWZ%*%X_group
a12<-XTWZU-XTWZD%*%I.ZTWZD.inv%*%ZTWZU
```

```
a21<-t(a12)
a22<-t(X_group)%*%ZTWZ%*%X_group-t(X_group)%*%ZTWZ%*%D.hat%*%
I.ZTWZD.inv%*%ZTWZ%*%X_group
Q.inv<-as.matrix(rbind(cbind(a11,a12),cbind(a21,a22)))
Q<-solve(Q.inv)
se.coef<-sqrt(diag(Q))
# Frequentist (se.Freq) and Bayesian standard error (se.Bayes)
S1<-matrix(0,ncol(Q.inv),ncol(Q.inv))
S1[7:15,7:15]<-S
Q_penalty.inv<-Q.inv+S1
Bayes.var<-solve(Q.inv+S1)
se.Bayes<-sqrt(diag(Bayes.var))
Freq.var<-Bayes.var%*%Q.inv%*%Bayes.var
se.Freq<-sqrt(diag(Freq.var))
#Calculate standard error for non-linear function
var.basis.freq<-Freq.var[7:15,7:15]
var.basis.Bayes<-Bayes.var[7:15,7:15]
basis<-G$X[,7:15]
eta.hat<-as.vector(betaCombined[7:15])
f.est<-mean(combinedData$f1)+as.vector(basis%*%eta.hat)
var.Freq.f<-rowSums((basis%*%var.basis.freq)*basis)
var.Bayes.f<-rowSums((basis%*%var.basis.Bayes)*basis)
seFreq.f<-sqrt(var.Freq.f)
seBayes.f<-sqrt(var.Bayes.f)
# Calculate standard error of spatial random effect parameters
se.theta<-sqrt(diag(solve(exp.infor)))
# Combine all the estimated parameters
# betaCombined: Parameters for fixed effect coefficients
#(both individual and group level) and spline coefficients
# theta.hat: Spatial random effect parameters
coef<-c(betaCombined,theta.hat)
se.coef.all<--c(se.coef,se.Freq,se.Bayes,se.theta)
f.hat<-f.est
seFreqf.hat<-seFreq.f
seBayesf.hat<-seBayes.f
#*********************************************#
#        Final Output                                          #
#*********************************************#
output<-list(coef=coef,se.coef=se.coef.all,f.hat=f.hat,
seBayes.f=seBayesf.hat,seFreq.f=seFreqf.hat)
output$coef
output$se.coef
```