To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–16

Research article

An investigation of the impact of various geographical scales for the specification of spatial dependence

Su Yun Kang^{a,b,*}, James McGree^{a,b}, Peter Baade^{c,d,e}, and Kerrie Mengersen^{a,b}

^a*Mathematical Sciences School, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia;* ^b*CRC for Spatial Information, 204 Lygon Street, Carlton, Victoria 3053, Australia;* ^c*Viertel Centre for Research in Cancer Control, Cancer Council Queensland, Gregory Tce, Fortitude Valley, Australia;* ^d*School of Public Health, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia;* ^e*Griffith Health Institute, Griffith University, Australia*

(Received 00 Month 20XX; final version received 00 Month 20XX)

Ecological studies are based on characteristics of groups of individuals, which are common in various disciplines including epidemiology. It is of great interest for epidemiologists to study the geographical variation of a disease by accounting for the positive spatial dependence between neighbouring areas. However, the choice of scale of the spatial correlation requires much attention. In view of a lack of studies in this area, this study aims to investigate the impact of differing definitions of geographical scales using a multilevel model. We propose a new approach – the grid-based partitions and compare it with the popular census region approach. Unexplained geographical variation is accounted for via area-specific unstructured random effects and spatially structured random effects specified as an intrinsic conditional autoregressive process. Using grid-based modelling of random effects in contrast to the census region approach, we illustrate conditions where improvements are observed in the estimation of the linear predictor, random effects, parameters, and the identification of the distribution of residual risk and the aggregate risk in a study region. The study has found that grid-based modelling is a valuable approach for spatially sparse data while the SLA-based and grid-based approaches perform equally well for spatially dense data.

Keywords: Bayesian hierarchical models; ecological fallacy; grid-based partitions; integrated nested Laplace approximation; intrinsic conditional autoregression; spatial epidemiology

Classification codes: 62H11; 91B72

1. Introduction

Ecological studies are common in many disciplines, including environmental epidemiology [41, 51], social science [16, 18, 19], political science [30, 31], and geography [32, 42]. Ecological analyses refer to study and inference based on characteristics of groups or aggregates of individuals [23, 35]. These aggregates are typically geographically defined by census regions, where the population information are readily available.

Ecological analyses are commonly based on group level data to make inferences about the collective set of individuals within the groups due to confidentiality issues for accessing individual level disease data [22, 28]. Ecological fallacy [48, 52] is committed when a relationship observed at the group level is assumed to exist at the individual level. Some other issues that arise in ecological analyses are as follows. First, the analysis of aggregated data may suffer from changes in geographical boundaries over time. Second, a problem

*Corresponding author. Email: s7.kang@qut.edu.au

concerning aggregated data is the modifiable areal unit problem (MAUP), which is defined as sensitivity of statistical results to the definition of geographical units over which data are collected [43]. For instance, various datasets may exhibit different spatial patterns when viewed at one spatial scale compared to another, which is known as a ‘scale’ effect [45].

To overcome the issue of ecological fallacy and to increase the statistical capacity to study small-scale geographical variation, this study employs both group and individual level characteristics to study their effect on individual outcomes in a multilevel (hierarchical) modelling framework [28, 58, 60, 61]. The individual level data control for the bias while the ecological data may provide additional efficiency gains in the estimation of the parameters of interest [24]. Multilevel models are flexible as they allow the formation of complex nested and crossed structures, and also dependencies in data through random effects [61]. However, the complex nature of the multilevel models often results in an intractable likelihood function. A solution to facilitate inference in such cases is the application of a Bayesian hierarchical multilevel modelling framework accompanied by computational algorithms such as Markov chain Monte Carlo (MCMC) techniques [15, 21]. See [36, 40, 47, 53] for a range of Bayesian applications in complex epidemiological and biomedical problems.

In ecological studies, the modelling of spatial dependence is often based on the belief that individuals within areas are more likely to be similar and areas that are geographically closer exhibit more similar residual relative risks. Ignoring positive spatial dependence between the residuals of neighbouring areas may underestimate standard errors on parameter estimates, and cause confounding by location if the exposure of interest has a spatial structure [57]. Given the acknowledged MAUP, better understanding of the choice of the scale of the spatial correlation is vital. One of the most common approaches in defining a neighbourhood structure is to take areas i and j to be neighbours if they share a common boundary. [59] argues that this approach is not attractive unless all regions are of similar size and arranged in a regular pattern. Other possible neighbourhood structure definitions could be some known function of the distance between centroids of areas [14]; and both a non-spatial and a spatial random effect assigned as an intrinsic conditional autoregressive (ICAR) prior [9]. See [37] for more alternative specifications within the CAR class.

The aim of this paper is to investigate the impact of differing definitions of geographical areas on the estimation of individual disease risks using a multilevel model that combines both group and individual level characteristics in a Bayesian framework. We propose to model area-specific random effects using regular lattices which are far smaller than the census regions to allow for better specification and identification of spatial effect. This research endeavours to examine the difference in model outcomes using a range of geographical scales for spatially sparse data and spatially dense data.

The paper is organized as follows. In Section 2, we provide details about the case study data, spatially sparse data and spatially dense data for analyses, and an exploratory data analysis. Section 3 outlines the methods and modelling approach employed in this study including model, likelihood, priors, posteriors, computation and model performance evaluation. The results of model fitting for spatially sparse data and spatially dense data are reported in Sections 4 and 5, respectively. Section 6 summarizes the results from both studies and finally, Section 7 discusses some of the issues and provides an overall conclusion.

2. Motivating Example

To motivate our research, we investigate the relationship between area-disadvantage and individual characteristics, and the risk of being diagnosed as having advanced breast can-

cer, using breast cancer data in Queensland, Australia. [5] employed multilevel analytical methods to investigate the links between geographic remoteness, area disadvantage, individual level factors and advanced breast cancer among women aged 30–79 years who were living in Queensland during 1997–2006. The study was designed to investigate the extent of large-area geographical variation in advanced breast cancer risk after adjusting for the characteristics of individuals within those area. In contrast, one of the ultimate aims of our study is to explore various spatial scales in capturing the geographical variation effectively. We consider two scenarios for spatial clustering of the data, described below.

2.1 Case study data

Ethics and Data Custodian approval was obtained from Queensland Health for this study (HREC/09/QHC/25). The breast cancer data were extracted from the population-based Queensland Cancer Registry (QCR). The data consist of all women diagnosed as having invasive breast cancer in Queensland between 1 January 1996 and 31 December 2009 (inclusive). We obtained data for each case on individual-level characteristics (i.e., year of diagnosis, age, indigenous status, occupation, and marital status) as well as measures of geographic remoteness and area disadvantage defined at the census region level, namely Statistical Local Areas (SLAs). Each cancer case also contains the longitude and latitude co-ordinates of the patient’s residential address as well as the SLA variable, which denotes the spatial entity where a patient resides. The co-ordinate information is also used to assign patients to their respective grid cells after partitioning the study region into regular grid cells. Both the SLA and grid information are used in modelling the spatial random effects. We follow the definition of advanced breast cancer in [5]. With respect to the area level covariates, since the information was only measured at the SLA level, we thus assigned the scores to the patients based on the SLAs in which they resided. The patients’ characteristics of interest at time of diagnosis are described in Table 1.

2.2 Study 1: Spatially sparse data

In Study 1, we focus on breast cancer cases in a study region with latitude ranges from -28 to -27.29315 and longitude ranges from 151.91364 to 153.18721 . The study region extends from Brisbane, the capital and most populous city in Queensland, to Toowoomba, the most populous non-capital inland city in Queensland, Australia. The cancer cases are spatially sparse with of a mixture of dense and scarce points.

To evaluate the impact of modelling the spatial effect at different spatial scales, we consider the partitions at the SLA level and the grid level by discretizing the study region using grids 10×10 , 20×20 , 30×30 , 50×50 and 100×100 (see Table 2). The adjacency matrix for the SLAs is calculated using the program `GeoDa` [3] using first order queen definition of adjacency, where the SLAs are considered to be neighbours if they share a common border or vertex. On the other hand, the `cell2nb` function in the `spdep` R package [11] is used to generate a list of neighbours for the grid cells, by applying a queen definition of neighbourhood, where two grid cells are termed neighbours if they share a common edge or vertex.

2.3 Study 2: Spatially dense data

In Study 2, we focus on breast cancer cases in Brisbane only as we are interested in studying disease distribution of a smaller geographical area. The latitude in the study region of Brisbane ranges from -27.63009 to -27.29315 while the longitude ranges from 152.91364 to 153.18721 . The cancer cases in Brisbane exhibit spatially dense pattern as

Brisbane is a highly populated region. In this scenario, we consider spatial scales at the SLA level and the grid levels 10×10 , 20×20 , 40×40 , 50×50 and 100×100 (see Table 3).

2.4 Exploratory data analysis

Here we conduct some exploratory analysis to assess the spatial covariance structure at different scales. Figures 1 and 2 contain the variograms [13] at various scales for Study 1 and Study 2. We include a fitted Gaussian variogram model [44] to illustrate its suitability as in the analysis that follows we employ a Gaussian process with a Gaussian covariance structure [27].

As shown in Figure 1, at the SLA level, the empirical variogram displays inconsistent spatial structure at larger lags and the Gaussian variogram produces a poor fit. At the regular grid levels, the empirical variogram displays more typical and realistic covariance structures. The Gaussian variogram model seems to fit reasonably well at grid 10×10 , 20×20 , and 30×30 at which the last two scales fit better than the first scale. At grids 50×50 and 100×100 , there is possibly cyclicity in the variogram that could be due to lack of data as there are large number of grid cells with zero counts. The exploratory analysis presented here has implied that modelling the spatial data using a Gaussian covariance structure is adequate at most spatial scales in Study 1 except for the SLA level which has irregularly-shaped regions and the grid 100×100 which has excessively fine grid cells and little spatial structure as shown in the empirical variogram.

With regard to the variogram obtained for Study 2 in Figure 2, fitting a Gaussian variogram model at the SLA level and grids 10×10 and 20×20 seems to be adequate, with questionable benefit of the spatial model at the grid 40×40 . The variograms at grids 50×50 and 100×100 suggest that there is a lack of spatial structure at these two scales. Again, these variograms support the choice of a Gaussian covariance structure to model the spatial random effects in Study 2.

We note that [55] cautions against the use of a Gaussian process with a Gaussian covariance structure as they may be overly smooth. See [27] for the connection between the Gaussian covariance function used in linear Gaussian process models and the Gaussian covariance structure of linear Gaussian Markov random field models. Though for our purposes and disease mapping in general, the use of an ICAR model has been motivated by the ability to smooth out random effects, as in [4] we are interested in smoothing out random variation unrelated with underlying risk.

3. Methods and modelling

3.1 Overall model

Using the motivating dataset described above, we are interested in modelling the individual risk of advanced breast cancer using a multilevel model [28, 58, 60, 61]. Let y_{hij} denotes the Bernoulli outcome of advanced breast cancer for the j -th individual in the i -th grid in the h -th SLA and p_{hij} denotes the individual risk of advanced breast cancer for the j -th individual in the i -th grid in the h -th SLA. The aggregate risk in the h -th region is $p_h = \frac{1}{n_h} \sum_{i=1}^{n_h} p_{hi}$. In the spirit of [29], the individual risk is modelled via the logistic regression model with inclusion of individual level and area level covariates. Several studies have developed methods for incorporating individual level information on the exposure of interest [10, 20, 25, 26, 46, 62].

In this study, we model the spatial and non-spatial random effects on one of these two

partitions: the SLA level and the grid level. At the SLA level, the model is as follows:

$$\text{logit}(p_{hij}) = \mu + u_h + v_h + \sum_r \alpha_r x_{hij}^{(r)} + \sum_p \beta_p z_h^{(p)},$$

where μ denotes the intercept, u_h and v_h denote unstructured and spatially-structured random effects respectively [9], that are both modelled on SLAs, α_r denotes the covariate effect of the r -th individual level covariate, β_p denotes the covariate effect of the p -th SLA level covariate, x_{hij} denotes individual level covariates, and z_h denotes SLA level covariates. Here, $\mathbf{u} = (u_1, u_2, \dots, u_h)$ is a set of area-specific random effects that are assumed to be independent, while $\mathbf{v} = (v_1, v_2, \dots, v_h)$ is a set of area-specific spatially structured random effects that are believed to play a role in the structuring of the hierarchy. When the random effects are modelled at a grid level, u_h is replaced by u_i , the unstructured random effects modelled on grids and v_h is replaced by v_i , the spatially-structured random effects modelled on grids.

The likelihood for the individual level model is a Bernoulli distribution,

$$\mathcal{L}(p_{hij}|y_{hij}) = \prod_{h,i,j} (p_{hij})^{y_{hij}} (1 - p_{hij})^{1-y_{hij}}.$$

To complete the Bayesian specification, we assign prior distributions to the unknown parameters $\{\mathbf{u}, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$. We model $u_i \sim \text{Normal}(0, 1/\tau_u)$, representing exchangeable random baseline risks for each grid. The precision parameter τ_u of component $\mathbf{u}$, is assigned a gamma distribution $\tau_u \sim \text{Gamma}(a_u, b_u)$ [7]. The spatial component $\mathbf{v}$ is specified such that spatial structure is induced, via conditional autoregression [8]. The full conditional for v_i , $i = 1, \dots, I$ depends solely on the random effects of neighbouring areas. Let W_{ij} be an indicator function which takes on the value one if grid cells i and j are neighbours that share a common boundary, and zero otherwise; W_{ii} is set equal to zero. We note that other specifications for $\mathbf{W}$ are also possible, for example $\mathbf{W}$ could be a function of distance to the centroid of the area, of the size of the area, or the length of common boundary. The intrinsic conditional autoregression (ICAR) specification [24] is given by

$$v_i | \mathbf{v}_{-i} \sim \text{Normal} \left(\frac{1}{m_i} \sum_{i \sim j} W_{ij} v_j, \frac{1}{m_i \tau_v} \right),$$

where $m_i = \sum_{i \sim j} W_{ij}$ is the number of neighbours for grid cell i . As the induced precision matrix is not positive definite, the conditional specification above does not yield a proper joint distribution for $\mathbf{v}$. The density can, however, be expressed via a pairwise difference distribution as

$$\pi(\mathbf{v} | \mathbf{W}, \tau_v) \propto \tau_v^{(N-1)/2} \exp \left\{ -\frac{\tau_v}{2} \sum_{i < j} W_{ij} (v_i - v_j)^2 \right\},$$

where N denotes the total number of grid cells and $\mathbf{W}$ denotes the $I \times I$ matrix of adjacency indicators representing the neighbourhood structure. The unknown precision parameter τ_v determines the strength of dependence between the parameters v_i and v_j , and is assigned a gamma distribution $\tau_v \sim \text{Gamma}(a_v, b_v)$ [7]. For the elements of the regression components $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_r)$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$, we choose univariate normal priors with mean zero and variance σ_α^2 and σ_β^2 , respectively.

The combination of the ICAR prior for $\mathbf{v}$ and a set of independent random effects $\mathbf{u}$ is known as the convolution prior/model proposed by [9]. We assume that the random effects

consist of two parts where the first part ($\mathbf{v}$) is spatially interdependent between areas and the second part ($\mathbf{u}$) captures unstructured heterogeneity. This is a reasonable assumption in practice as the convolution model is able to compromise between the nonspatial exchangeable effects and the spatially structured effects. Its use is well preceded, see for instance, [33, 34] and [64].

Up to a constant of proportionality, the product of the full likelihood and the independent prior distributions for the unknown parameters ($\pi(\mathbf{u})$, $\pi(\mathbf{v})$, $\pi(\boldsymbol{\alpha})$, $\pi(\boldsymbol{\beta})$, $\pi(\tau_u)$, $\pi(\tau_v)$) constitutes the joint posterior distribution of the model parameters. That is,

$$\pi(\mathbf{u}, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau_u, \tau_v | y_{hij}) \propto \pi(y_{hij} | \{\mathbf{u}, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau_u, \tau_v\}) \pi(\mathbf{u}) \pi(\mathbf{v}) \pi(\boldsymbol{\alpha}) \pi(\boldsymbol{\beta}) \pi(\tau_u) \pi(\tau_v).$$

3.2 Computation

Since it is not possible to compute the above specification of the posterior distribution of the parameters directly; a popular approach to facilitate computation is MCMC. However, by building a MCMC scheme for this model, we found that the computational cost increases proportionally to the number of individuals included in the model as well as the number of grid cells. In light of this computational burden, we adopt the integrated nested Laplace approximation (INLA) approach proposed by [50], which performs approximate Bayesian inference for latent Gaussian models. INLA is able to return accurate parameter estimates in relatively short computational time. In addition, a model choice criterion termed the deviance information criterion (DIC), and predictive measures including logarithmic score and probability integral transform (PIT) are provided. We note that INLA methodology requires the full conditional density for the latent field to be ‘near’-Gaussian, which is usually achieved either by replications or smoothing. As this study uses a spatial model (we are interested in spatial smoothing) that falls under the category of latent Gaussian models [39, 50], it is reasonable to employ INLA computation. Furthermore, [49] detail the validity of the approximate Bayesian inference for hierarchical GMRF models (the ICAR model is one of the GMRF-based models).

The latent Gaussian models can be defined as a Bayesian hierarchical model with three levels. The first level is the observational equation $\mathbf{y} | \mathbf{x} \sim \pi(\mathbf{y} | \mathbf{x})$, where $\mathbf{y}$ denotes the observations. Here, we formulate the distributional assumption for the observations dependent on latent components where in our case $\mathbf{y}$ follows a Bernoulli distribution. Given the parameters of the observation model, the observations are assumed to be conditionally independent, as in [49]. Secondly, we assign an a priori model for the unknown parameters and specify the corresponding Gaussian Markov random field (GMRF). The second level can be written as $\mathbf{x} | \boldsymbol{\theta} \sim N(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{Q}(\boldsymbol{\theta})^{-1})$, where $\mathbf{x}$ contains all components of the latent Gaussian field. At the last level, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ denotes the prior distribution for the parameters. $\mathbf{Q}(\boldsymbol{\theta})$ is the precision matrix of the Gaussian random vector $\mathbf{x}$, which is sparse. The posterior density can be written as

$$\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \pi(\mathbf{x} | \boldsymbol{\theta}) \prod_{i \in I} \pi(y_i | x_i, \boldsymbol{\theta}).$$

By default a flat improper prior for the intercept $\boldsymbol{\mu}$ is assumed in R-INLA. The components of the latent field, $\mathbf{x} = \{\boldsymbol{\mu}, \mathbf{u}, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ are assigned a zero-mean Gaussian distribution with precision matrix $\mathbf{Q}(\boldsymbol{\theta})$, resulting in a latent Gaussian field. The latent field is controlled by a few hyperparameters $\boldsymbol{\theta} = \{\tau_u, \tau_v\}$. These hyperparameters are assigned different distributions, as previously described.

The main goal is to estimate the desired posterior marginals

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \quad (1)$$

while the posterior marginals of $\boldsymbol{\theta}$ are approximated by

$$\pi(\theta_i|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-i}. \quad (2)$$

where $\boldsymbol{\theta}_{-i}$ denotes all elements in $\boldsymbol{\theta}$ except for θ_i . Nested approximations and numerical integration are used to integrate out $\boldsymbol{\theta}$ in order to estimate Equations (1) and (2). The Laplace approximation [56] to the posterior of hyperparameters can be written as

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})},$$

where $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation to the full conditional of $\mathbf{x}$ and $\mathbf{x}^*(\boldsymbol{\theta})$ is the mode of the Gaussian approximation for each $\mathbf{x}$. Posterior marginals for the latent variables $\mathbf{x} = \{\boldsymbol{\mu}, \mathbf{u}, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ and the hyperparameters $\boldsymbol{\theta} = \{\tau_u, \tau_v\}$ are both computed via numerical integration. The posterior marginals can be used to compute summary statistics of interest, such as posterior means, variances or quantiles. We refer the reader to [50] and [12] for more details on INLA computation and applications.

In terms of prior specification, the precision parameters of the unstructured random effect and spatial effect, τ_u and τ_v , are both assigned gamma priors with parameters (1, 0.001) to impose the same level of spatial smoothing on the spatial field for each model throughout the study. Normal priors with mean zero and variance $\sigma_\alpha^2 = \sigma_\beta^2 = 100$ are chosen for the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We carried out sensitivity analyses to assess the impact of various choices of prior distributions on the models and found that the influence of priors is negligible based on minimal changes in the DIC.

3.3 Performance evaluation

In regard to model selection, DIC was used to select the most parsimonious model after penalizing for model complexity. A smaller DIC indicates a better fit of the model. As suggested by [54], DIC should not be used as an absolute measure of the ‘best’ model, but rather a method for screening alternative formulations in order to provide an indication of the relative fit of a set of candidate models. They argued that candidate models receiving DIC within 1 – 2 of the ‘best’ deserve consideration, while 3 – 7 have considerably less support. We present the candidate models with DIC within 5 of the smallest DIC in the following section.

In order to assess the predictive performance of these models, the logarithmic score (LS) for each model is computed [17]. Each model is assigned a numerical score based on the predictive distribution using the cross-validated scoring rules. For discrete observations Y_{hij} , the LS is defined as

$$\text{LS} = -\log(\pi_{y_{hij}}),$$

where $\pi_{y_{hij}} = \text{Prob}(Y_{hij} = y_{hij}|\mathbf{y}_{-hij})$ denotes the cross-validated predictive probability mass at the observed event. A smaller LS indicates a better predictive power of the model.

In order to investigate how the estimation of parameters changes across different spatial scales, the standard deviation of the posterior estimates of the estimated linear predictor,

random effects, and parameter estimates are used to provide guidance for each model. Furthermore, to assist in comparison across various spatial scales, the image plots of the spatial random effect and the aggregate risk at various scales are also presented in Section 4. The resulting disease maps can be used to study the geographical distribution of disease burden and inform public health resource allocation.

4. Results for Study 1: Spatially sparse data

Based on the DIC and LS described in Section 3.3, the five best models for fitting the breast cancer data are given below. These models are considered because their DIC values are within five from the smallest DIC value. Thus, we discuss these five models only for the rest of the section.

$$\text{Model 1} = \begin{cases} \text{logit}(p_{hij}) = \mu + \mathbf{u} + \text{YEAR}_{hij} + \text{MARITAL}_{hij} + \text{IRSD}_h, \\ \text{logit}(p_{hij}) = \mu + \mathbf{u} + \mathbf{v} + \text{YEAR}_{hij} + \text{MARITAL}_{hij} + \text{IRSD}_h. \end{cases}$$

$$\text{Model 2} = \begin{cases} \text{logit}(p_{hij}) = \mu + \mathbf{u} + \text{YEAR}_{hij} + \text{IRSD}_h, \\ \text{logit}(p_{hij}) = \mu + \mathbf{u} + \mathbf{v} + \text{YEAR}_{hij} + \text{IRSD}_h. \end{cases}$$

$$\text{Model 3} = \begin{cases} \text{logit}(p_{hij}) = \mu + \mathbf{u} + \text{MARITAL}_{hij} + \text{IRSD}_h, \\ \text{logit}(p_{hij}) = \mu + \mathbf{u} + \mathbf{v} + \text{MARITAL}_{hij} + \text{IRSD}_h. \end{cases}$$

$$\text{Model 4} = \begin{cases} \text{logit}(p_{hij}) = \mu + \mathbf{u} + \text{YEAR}_{hij}, \\ \text{logit}(p_{hij}) = \mu + \mathbf{u} + \mathbf{v} + \text{YEAR}_{hij}. \end{cases}$$

$$\text{Model 5} = \begin{cases} \text{logit}(p_{hij}) = \mu + \mathbf{u} + \text{IRSD}_h, \\ \text{logit}(p_{hij}) = \mu + \mathbf{u} + \mathbf{v} + \text{IRSD}_h. \end{cases}$$

4.1 Model selection - DIC and LS

The DIC and LS for Models 1 to 5 are presented in Figure 3. The smaller the scores for a model, the more favoured the model. Model 2 and Model 5 are observed to have the smallest DIC and LS. However, we note that the difference compared to the other models is only small, which suggests that these five models are comparable based on these criteria. It is also found that the DIC and LS for the models with random effects modelled at the SLA level are marginally smaller than that of the grid levels. It is observed that both scores are largest at the grid 10×10 and gradually decrease as the grid cell size becomes increasingly fine. In other words, model fit and predictive performance of the models improve as the grid cell size becomes smaller. When the grid cell size is fine enough, the model fit and predictive performance of a model are similar to that at the SLA level. Comparison of both scores at the SLA level and the grid 100×100 level for all five models are given in Table 4. We note that the percentage of difference between both spatial scales are relatively small. We will consider other model comparison criteria before selecting the best fitted model.

4.2 Estimation for linear predictor and random effects

The performance of the models is further examined using box plots of the standard deviation of the estimated linear predictor. The linear predictor refers to $\text{logit}(p_{hij})$ which include both the fixed and random effects in the model. The precision of estimation of each model is improved at the grid level compared to the SLA level as 10% – 20% smaller standard deviations are generally observed. Although there is no substantial difference across different grid levels, the standard deviation is slightly larger at the grid 100×100 . The box plots of the width of the 90% credible intervals of the estimated linear predictor appear to be similar to the patterns observed for the standard deviation. The width indicates the difference of the 5th percentile and the 95th percentile of the credible interval. Larger values indicate a larger spread of the posterior distribution. Based on the standard deviation and the width of the 90% credible intervals, Model 5 has better precision in estimation compared to the other models. Nevertheless, the performance of Model 4 is comparable to that of Model 5. The box plots of the standard deviation of the estimated linear predictor for Model 5 are illustrated in Figure 4(a).

In order to compare the estimation of the random effects at various spatial scales, we present the box plots of the standard deviation of the estimated unstructured random effect (σ_u) and spatially structured random effects (σ_v) of Model 5 in Figure 4(b) and Figure 4(c), respectively. We note that all other models show very similar patterns in terms of the standard deviation of the random effect. The precision of the estimates of $\mathbf{u}$ and $\mathbf{v}$ at the grid levels is noticeably larger than at the SLA level. A larger σ_u and σ_v at the SLA level may suggest that the covariates explain less variation in the SLA-based models than the grid-based models. The grid 10×10 appears to have the smallest standard deviation for all models which is possibly due to the small number of grid cells. We note that its difference with other grid levels is marginal, especially with the grid 30×30 . It is also observed that the increase in the number of grid cells in a model gradually increases the standard deviation of the random effects. For instance, at the grid 100×100 , many grid cells do not contain any information, which may induce a lot of unexplained variability in the model. Guided by these results, we suggest that grids 10×10 and 100×100 may not be the most appropriate scales due to the fact that the former has larger DIC and LS than all other scales while the latter exhibits marginally larger standard deviations for the estimated linear predictor and random effects. On the other hand, the grid 30×30 appears to be a reasonable choice to model the random effects in this study.

4.3 Parameter estimates

The posterior distributions of the regression parameter estimates for each model were examined by comparing the posterior mean, standard deviation, and the width of the 90% credible interval of the estimated parameters. We present the results for Model 5 only as this model appeared to be the best-fitted model amongst all models based on their predictive performance. For all five models, the posterior mean of the parameters differs marginally across different spatial scales. The difference is most likely contributed by the changes in spatial dependence across various spatial scales. It is also observed that the standard deviation of the posterior estimates are generally smaller at the grid levels compared to at the SLA level (see Figure 5). Furthermore, the standard deviation reduces slightly as the grid cell sizes become increasingly small. We note that the standard deviation for the model with unstructured random effects only is slightly smaller than the model with both unstructured and spatially structured random effects. Therefore, we suggest inspection of the plots of the posterior mean of the spatial effect to decide if it is necessary to retain the spatial component in modelling. We note that the width of the 90% credible interval of the posterior estimates display the same pattern as the

standard deviation of the posterior estimates due to the fact that the posterior estimates are Gaussian distributed. Posterior probabilities ($P(\beta_p < 0)$) for the regression parameters for Model 5 is presented in Table 5. It is shown that across the various spatial scales, there is only marginal difference in the posterior probabilities which suggests that the changing spatial scales have a relatively small effect on the posterior inference of the regression parameters. The results of the parameter estimates appear to agree with the recommendation made earlier on choosing the grid 30×30 for this dataset.

4.4 *Plots of spatial effect and aggregate risk*

Based on the previously discussed results, we present the plots of the posterior mean of the spatial effect for Model 5 in Figure 6. The distribution of the spatial random effect at the SLA level is observed to be different to those at the grid levels. When inference is made at the SLA level which contains large geographical units, only crude estimates of the spatial effect are obtained. As the grid cell becomes increasingly small, the spatial effect is better captured and identified as a more localized distribution can be seen. The plots at grids 30×30 and finer clearly show that the excess risk exists in the middle right part (darker spot) of the map. It is therefore apparent that the spatial effect at a smaller spatial scale informs the presence of excess risk on a map more effectively than at a larger scale.

Plots of the aggregate risk predicted from Model 5 at various spatial scales are illustrated in Figure 7. The plots provide information on the risk of being diagnosed with advanced breast cancer for each region at various spatial resolutions. The aggregate risk at the SLA level appears to be different from those at the grid levels. This can be explained by the fact that the former aggregates the individual risks to a larger scale and assumes the risk to be constant within a region, hence crude estimation of the risks. A spatial scale that is too fine (such as grid 100×100) leads to small counts in a grid cell, more uncertainty, and less robustness in the spatial model; and may produce imprecise estimates. It is therefore necessary to have sufficient counts in the grid cells.

Figure 6 presents the posterior mean of the spatial effect, which is the excess risk. The excess risk is also known as residual disease risk which is unaccounted for in a model after adjusting for individual and area level covariates. Figure 7, on the other hand, presents the aggregate risk at various spatial scales. The aggregate risk is obtained by taking the average of the disease risk of the individuals in each geographical partition. Based on Figures 6 and 7, it is evident that the SLA level is inferior to the grid-based partitions in that estimation can only be made at a coarse scale. Among the various grid sizes, the grid 50×50 appears to produce the smoothest plot for the spatial effect which displays the presence of excess risk and a risk map that shows a more localized distribution of aggregate risk. However, the plots produced by the grid 30×30 also appear to adequately serve the same purposes.

5. Results for Study 2: Spatially dense data

The four best models for fitting the breast cancer data in Study 2 are chosen based on the DIC, where their DIC values are within five from the smallest DIC value. We do not present the results for Study 2 in great detail but highlight those that are significantly different from those observed in Study 1. The four models are: Model 1 (INDIG + YEAR + IRSD); Model 2 (INDIG + IRSD); Model 3 (INDIG + YEAR); and Model 4 (INDIG).

Based on the DIC and LS presented (Figure 8), it is apparent that the model fit and predictive performance at the grid levels are comparable to those at the SLA level. How-

ever, there is no clear improvement in the scores as the grid cell size becomes increasingly small, despite slight variations that are observed. It is also noted that the scores at the grid 10×10 are larger than the scores at the other scales as this geographical partition has the largest area size amongst all scales. Model 3 and Model 4 have similar LS that are smaller than the other models, which implies that the predictive performance of these two models are comparable. In terms of DIC, Model 3 and Model 4 are also favoured compared to the other models.

Based on the box plots of the standard deviation of the estimated linear predictor, the precision of estimation of each model is very similar at both the grid levels and the SLA level. Nevertheless, there are no significant changes in precision of estimation across various sizes of the grid cell. Model 4 appears to have the highest precision in estimation compared to the other models. In terms of the estimation of the parameters of interest, the results are similar to those observed in Study 1, in that the posterior mean of the parameters differs across different spatial scales. The standard deviation and the width of the 90% credible interval of the posterior estimates fluctuate at various spatial scales. There is no improvement seen at the grid levels compared to the SLA level.

The results of Study 2 suggest that the inclusion of the spatial component $\mathbf{v}$ does not necessarily improve the model fit and predictive performance. The models with random effects $\mathbf{u} + \mathbf{v}$ are found to have slightly higher DIC and LS than the models with $\mathbf{u}$. This also applies to the standard deviation of the estimated linear predictor for every model. Based on these results, we suggest that for spatially dense data, the effect of the spatial component is minimal due to the lack of inhomogeneity across the geographical regions. Thus, the spatial component $\mathbf{v}$ may be removed from the models.

6. Summary of results from both studies

The results from Study 1 and Study 2 are summarized in Table 6.

7. Discussion

In this article, we investigate the impact of the choice of spatial scale for modelling individual disease outcomes via Bayesian hierarchical spatial modelling by combining individual and aggregate information in the context of epidemiology. The hierarchical model allows for unmeasured covariates and for potential errors in the observed data using the area-specific random effects [47]. We examine the modelling outcomes for the scenarios where the random effects are modelled based on regular grid partitions as opposed to the census region level. The intrinsic conditional autoregressive (ICAR) specification [8, 24] of spatially structured effects for study regions that contain highly irregular lattices, such as our SLAs, has often been criticized [59]. It is more natural and reasonable to implement the ICAR specification for regular lattices where the geographical areas are of similar size and arranged in a regular pattern [59], such as the grid-based partitions proposed in this study. As seen, the region of Brisbane that constitutes the spatially dense data is included in the study region for the spatially sparse data. This is due to the fact that a mixture of dense and scarce points is required to form the spatially sparse data. This decision has not resulted in bias for the results of the spatially sparse data as the spatial smoothing effect induced by the ICAR model only takes into account neighbouring SLAs or grid cells. The spatially dense region does not contribute to the smoothing effect of the region with scarce points since these two regions are far apart.

In the analyses presented in Study 1 (spatially sparse data), the DIC and LS for the SLA-based models were slightly smaller than the grid-based models. However, we acknowledge

that DIC is not the definitive criterion in choosing the best model but rather provides a list of candidate models. Indeed, despite having slightly higher DIC, the grid-based modelling of random effects has the distinct advantage that the estimation of linear predictor and random effects is better than the SLA-based modelling. In addition, the estimation of the regression parameters using the grid-based approach is increasingly precise as the grid cell size reduces. The performance of the models at the grid level was superior to that of the SLA-level even when there are fewer grid cells (100 grid cells) than SLAs (308 regions). This provides evidence that the modelling of random effects based on regular lattices (grid partitions) is a valuable approach. We note that, however, it is necessary to have a sufficiently large sample size in order to discretize the study region into finer grid cells, as it was observed that the parameter estimation of some models was worse for grid cells finer than 30×30 . A further study has to be carried out to determine the sample size for an optimum spatial scale. Based on the results from Study 1, we recommend partitioning of the study region into fine grid cells for spatially sparse data by selecting the most appropriate spatial scale from a range of choices.

On the other hand, based on the results from Study 2 (spatially dense data), we note that the model fit (based on DIC), predictive performance (based on LS), and precision of estimation at the grid levels are similar to those at the SLA level. In other words, there is no apparent advantage in partitioning the study region into finer scales for spatially dense data. Modelling of the random effects at the census regions level (SLA) appears to be sufficient. This could be explained by the fact that with a lot of data, there is an intrinsic averaging effect, so that homogeneity assumption is better fulfilled within each SLA. The SLAs thus borrow strength effectively from neighbouring SLAs and produce good fit. Nevertheless, if the distributions of the excess risk and the aggregate risk are of interest, it is evident that the SLA level is inferior to the grid-based partitions as the estimation can only be made at a coarse scale which is not as specific and relevant to the local population.

As seen, choosing an appropriate spatial scale is important, particularly for spatially sparse data. Based on the results of the study, it is recommended to repeat the spatial analyses at multiple spatial scales in order to determine a suitable scale. The choice of spatial scale may change according to various inferential aims of interest. For instance, based on the plot of the posterior mean of the spatial effect (Figure 6), grids 30×30 and finer are favoured as more localized excess risk is identified compared to the coarser spatial scales. In contrast, the plot of aggregate risk (Figure 7) does not support the grid 100×100 because the fine grid cells lead to insufficient information in each grid cell. It is therefore important to consider model fit, predictive performance and parameter estimation of the models. Although the analyses conducted in this study were comprehensive, they were deliberately confined to investigating the impact of SLA and grid-based partitions for a real disease phenomenon. Thus, they do not allow us to make more definitive statements about the choice of grid size or spatial scale in general. This remains an open topic for further research.

The breast cancer data in this study consist of routinely-collected health information obtained from Queensland Cancer Registry (QCR). The data contain the population of women diagnosed with female breast cancer who lived within grid partitions or within ecologic units (SLAs), which makes the multilevel model particularly attractive. However, in some instances, multilevel studies could face a major practical limitation where it is expensive or ethically difficult to obtain routinely collected individual data [22]. Clearly, the modelling of spatial effects at a smaller grid scale displays the distribution of residual risk and the aggregate risk of a disease at a more localized scale compared to the census region level. By capturing the excess and aggregate risks at a finer scale, the resulting disease maps can be used to study the geographical variation of the disease more accurately. Reliable

information about current and predicted areas of high disease risk is necessary to address the social and financial burden of the disease in a cost-effective manner. Knowledge of the variability in residual spatial risk will greatly assist in disease surveillance. We note that, similarly to SLAs that are purely administrative regions, the grid partitions are artificial boundaries and so do not reflect the access to health services or sense of community. However, the researcher has the power to manipulate the spatial resolution so as to obtain an appropriate grid cell size. If the data aggregation is to be undertaken on the grid partitions, one of the challenges is to obtain population data for the grids.

Throughout the analyses reported here, we have used the ICAR specification for the latent spatial component. This choice has been made mainly for its broad usage in Bayesian disease mapping to account for adjacency-based spatial correlation effects in areal data where small area geographical variation is of interest [6]. It is appealing in that it incorporates local smoothing via the consideration of each area's neighbours [24]. This prior can be used regardless of how the study region is discretized, e.g., into the SLAs or grid partitions. It is also straightforward to implement the ICAR approach in the R-INLA program, where different adjacency matrices can be specified according to the various spatial scales.

There are many alternative specifications of the latent spatial process [63]. For example, another approach implemented by the INLA algorithm [50] is the GMRF approximation to the Matérn field via stochastic partial differential equations (SPDE) [38]. This computational approach avoids the use of a grid and allows for continuous modelling of latent spatial effects. This offers several computational advantages. In the context of a geostatistical model, using the SPDE, [38] construct a mesh based on a triangulation of the sampling points instead of a regular grid. The SPDE is used to find a GMRF, with local neighbourhood and corresponding sparse precision matrix that best represents the Matérn field. However, the triangulation mesh constructed for the observed points is of irregular shape and size. We did not consider the SPDE approach in our work because it is not straightforward to implement it on the SLA partition. We note that it is important to be able to determine the shape and size of the geographical areas in this study as one of the main aims is to investigate whether the SLA partition is a reasonable choice of scale to model the latent spatial effect (residual disease risk) as opposed to the grid partitions, given that the SLA approach (administrative districts) is a popular choice in studying small area geographical variation.

In summary, by investigating two scenarios for spatial clustering of the data (spatially sparse data and spatially dense data), we have demonstrated the efficiency gains in estimation using regular grids in spatially sparse data. For both scenarios, the plots of posterior mean of the spatial effect and the plots of aggregate risk at smaller geographical units (grids) were favoured compared to large geographical units (SLAs). The presented analyses have suggested that the grid-based partition for the modelling of random effects is able to improve model estimation and inference in contrast to the SLA-based modelling in the context of spatially sparse data which has a mixture of dense and scarce points; whereas for spatially dense data, the SLA-based and grid-based approaches perform equally well.

Acknowledgements

The work was supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme. Peter Baade is supported by an NHMRC Career Development Fellowship (#1005334). The authors would like to thank Cancer Council Queensland for supplying the data for analysis, Susanna Cramb for assisting with the data and two anonymous reviewers for their helpful comments.

References

- [1] P. Adhikari, *Socio-economic indexes for areas: Introduction, use and future directions*, Canberra: Australian Bureau of Statistics (2006).
- [2] AIHW, *Rural, regional and remote health: A guide to remoteness classifications*, Canberra: Australian Institute of Health and Welfare AIHW Cat. No. PHE 53 (2004).
- [3] L. Anselin, I. Syabri, and Y. Kho, *GeoDa: An introduction to spatial data analysis*, Geogr. Anal. 38 (2006), pp. 5–22.
- [4] R. Assunção and E. Krainski, *Neighborhood dependence in Bayesian spatial models*, Biometrical Journal 51 (2009), pp. 851–869.
- [5] P.D. Baade, G. Turrell, and J.F. Aitken, *Geographic remoteness, area-level socio-economic disadvantage and advanced breast cancer: A cross-sectional, multilevel study*, J. Epidemiol. Community Health 65 (2011), pp. 1037–43.
- [6] S. Banerjee, B.P. Carlin, and A.E. Gelfand, *Hierarchical modeling and analysis for spatial data*, Vol. 101, Chapman & Hall, 2004.
- [7] L. Bernardinelli, D. Clayton, and C. Montomoli, *Bayesian estimates of disease maps: How important are priors?*, Stat. Med. 14 (1995), pp. 2411–2431.
- [8] J. Besag and C. Kooperberg, *On conditional and intrinsic autoregressions*, Biometrika 82 (1995), pp. 733–46.
- [9] J. Besag, J. York, and A. Mollié, *Bayesian image restoration, with two applications in spatial statistics*, Ann. Inst. Statist. Math. 43 (1991), pp. 1–59.
- [10] N. Best, S. Cockings, J. Bennett, J. Wakefield, and P. Elliott, *Ecological regression analysis of environmental benzene exposure and childhood leukaemia: Sensitivity to data inaccuracies, geographical scale and ecological bias*, J. Roy. Statist. Soc. Ser. A 164 (2001), pp. 155–74.
- [11] R. Bivand, L. Anselin, O. Berke, A. Bernat, M. Carvalho, Y. Chun, C.F. Dormann, S. Dray, R. Halbersma, N. Lewin-Koh, *et al.*, *spdep: Spatial dependence: Weighting schemes, statistics and models*, R package version 0.5-56, URL <http://cran.r-project.org/web/packages/spdep/index.html> (2011).
- [12] M. Blangiardo, M. Cameletti, G. Baio, and H. Rue, *Spatial and spatio-temporal models with R-INLA*, Spat. Spatiotemporal Epidemiol. 4 (2013), pp. 33–49.
- [13] N. Cressie, *Statistics for Spatial Data*, John Wiley and Sons, Inc, 1993.
- [14] N. Cressie and N.H. Chan, *Spatial modeling of regional variables*, J. Amer. Statist. Assoc. 84 (1989), pp. 393–401.
- [15] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in practice: Interdisciplinary statistics*, Vol. 2, Chapman & Hall/CRC, 1995.
- [16] A.N. Glynn and J. Wakefield, *Ecological inference in the social sciences*, Stat. Med. 7 (2010), pp. 307–22.
- [17] T. Gneiting and A.E. Raftery, *Strictly proper scoring rules, prediction, and estimation*, J. Amer. Statist. Assoc. 102 (2007), pp. 359–78.
- [18] L.A. Goodman, *Ecological regressions and behaviour of individuals*, Amer. Sociol. Rev. 18 (1953), pp. 663–64.
- [19] L.A. Goodman, *Some alternatives to ecological correlation*, Amer. J. Sociol. 64 (1959), pp. 610–25.
- [20] P. Goovaerts, *Combining area-based and individual-level data in the geostatistical mapping of late-stage cancer incidence*, Spat. Spatiotemporal Epidemiol. 1 (2009), pp. 61–71.
- [21] P.J. Green, *A primer on Markov chain Monte Carlo*, in *Complex Stochastic Systems*, O.E. Barndorff-Nielsen, D.R. Cox, and K. Kluppelberg, eds., Chapman & Hall: London, UK, 2001, pp. 1–62.
- [22] S. Greenland, *Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects*, Int. J. Epidemiol. 30 (2001), pp. 1343–50.
- [23] S. Greenland, *A review of multilevel theory for ecologic analyses*, Stat. Med. 21 (2002), pp. 389–95.
- [24] S. Haneuse and J. Wakefield, *Geographic-based ecological correlation studies using supplemental case-control data*, Stat. Med. 27 (2008), pp. 864–87.
- [25] S.J. Haneuse and J.C. Wakefield, *Hierarchical models for combining ecological and case-control data*, Biometrics 63 (2007), pp. 128–36.
- [26] S.J. Haneuse and J.C. Wakefield, *The combination of ecological and case-control data*, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (2008), pp. 73–93.
- [27] M. Haran, *Gaussian random field models for spatial data*, in *Handbook of Markov Chain Monte Carlo*, S.P. Brooks, A. Gelman, G.L. Jones, and X.L. Meng, eds., Boca Raton: CRC Press, 2011, pp. 449–478.
- [28] C. Jackson, N. Best, and S. Richardson, *Improving ecological inference using individual-level data*, Stat. Med. 25 (2006), pp. 2136–59.
- [29] C. Jackson, N. Best, and S. Richardson, *Hierarchical related regression for combining aggregate and*

- individual data in studies of socio-economic disease risk factors*, J. Roy. Statist. Soc. Ser. A 171 (2008), pp. 159–78.
- [30] R. Johnston and C. Pattie, *Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting*, Political Analysis 8 (2000), pp. 333–45.
 - [31] G. King, *A Solution to the Ecological Inference Problem*, Princeton University Press, Princeton, New Jersey, 1997.
 - [32] G. King, *Geography, statistics, and ecological inference*, Ann. Assoc. Am. Geogr. 90 (2004), pp. 601–06.
 - [33] L. Knorr-Held, *Bayesian modelling of inseparable space-time variation in disease risk*, Statistics in Medicine 19 (2000), pp. 2555–2567.
 - [34] L. Knorr-Held and J. Besag, *Modelling risk from a disease in time and space*, Statistics in Medicine 17 (1998), pp. 2045–2060.
 - [35] L.I. Langbein and A.J. Lichtman, *Ecological Inference*, Series/No. 07-010, Sage: Thousand Oaks, CA, 1978.
 - [36] A.B. Lawson and H.R. Song, *Bayesian hierarchical modeling of the dynamics of spatio-temporal influenza season outbreaks*, Spat. Spatiotemporal Epidemiol. 1 (2010), pp. 187–95.
 - [37] D. Lee, *A comparison of conditional autoregressive models used in Bayesian disease mapping*, Spat. Spatiotemporal Epidemiol. 2 (2011), pp. 79–89.
 - [38] F. Lindgren, H. Rue, and J. Lindström, *An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach*, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (2011), pp. 423–498.
 - [39] S. Martino and H. Rue, *Case studies in Bayesian computation using INLA*, in *Complex Data Modeling and Computationally Intensive Statistical Methods*, P. Mantovan and P. Secchi, eds., Springer-Verlag, Milan, Italy, 2010, pp. 99–114.
 - [40] X. Min, D. Sun, Z. He, and M. Schootman, *A Bayesian hierarchical model of nontraumatic lower-extremity amputation rates*, Spat. Spatiotemporal Epidemiol. 1 (2010), pp. 169–76.
 - [41] H. Morgenstern, *Ecologic studies in epidemiology: Concepts, principles, and methods*, Annu. Rev. Publ. Health 16 (1995), pp. 61–81.
 - [42] S. Openshaw, *Concepts and Techniques in Modern Geography Number 38: The Modifiable Areal Unit Problem*, Geo Books, Norwich, 1984.
 - [43] S. Openshaw and P. Taylor, *A million or 80 correlation coefficient: Three experiments on the modifiable areal unit problem*, in *Statistical Applications in the Spatial Sciences*, N. Wrigley, ed., London, UK: Pion Press, 1979, pp. 127–44.
 - [44] E.J. Pebesma, *Multivariable geostatistics in S: the gstat package*, Computers & Geosciences 30 (2004), pp. 683–691.
 - [45] J.N. Perry, A.M. Liebhold, M.S. Rosenberg, J. Dungan, M. Miriti, A. Jakomulska, and S. Citron-Pousty, *Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data*, Ecography 25 (2002), pp. 578–600.
 - [46] R.L. Prentice and L. Sheppard, *Aggregate data studies of disease risk factors*, Biometrika 82 (1995), pp. 113–25.
 - [47] S. Richardson and N. Best, *Bayesian hierarchical models in ecological studies of health–environment effects*, Environmetrics 14 (2003), pp. 129–47.
 - [48] W.S. Robinson, *Ecological correlations and the behavior of individuals*, Int. J. Epidemiol. 38 (2009), pp. 337–41.
 - [49] H. Rue and S. Martino, *Approximate Bayesian inference for hierarchical Gaussian Markov random field models*, Journal of Statistical Planning and Inference 137 (2007), pp. 3177–3192.
 - [50] H. Rue, S. Martino, and N. Chopin, *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations*, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2009), pp. 319–92.
 - [51] R. Salway and J. Wakefield, *Sources of bias in ecological studies of non-rare events*, Environ. Ecol. Stat. 12 (2005), pp. 321–47.
 - [52] H.C. Selvin, *Durkheim’s suicide and problems of empirical research*, Amer. J. Sociol. 63 (1958), pp. 607–19.
 - [53] H.R. Song, A. Lawson, R.B. D’Agostino, and A.D. Liese, *Modeling type 1 and type 2 diabetes mellitus incidence in youth: An application of Bayesian hierarchical regression for sparse small area data*, Spat. Spatiotemporal Epidemiol. 2 (2011), pp. 23–33.
 - [54] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A.V.D. Linde, *Bayesian measures of model complexity and fit*, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (2002), pp. 583–639.
 - [55] M.L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Chapter 2, Springer Verlag, New York, 1999.
 - [56] L. Tierney and J.B. Kadane, *Accurate approximations for posterior moments and marginal densities*,

- J. Amer. Statist. Assoc. 81 (1986), pp. 82–86.
- [57] J. Wakefield, *Sensitivity analyses for ecological regression*, Biometrics 59 (2003), pp. 9–17.
- [58] J. Wakefield, *Ecological inference for 2×2 tables (with discussion)*, J. Roy. Statist. Soc. Ser. A 167 (2004), pp. 385–445.
- [59] J. Wakefield, *Disease mapping and spatial regression with count data*, Biostatistics 8 (2007), pp. 158–83.
- [60] J. Wakefield, *Ecologic studies revisited*, Annu. Rev. Publ. Health 29 (2008), pp. 75–90.
- [61] J. Wakefield, *Multi-level modelling, the ecologic fallacy, and hybrid study designs*, Int. J. Epidemiol. 38 (2009), pp. 330–36.
- [62] J. Wakefield and R. Salway, *A statistical framework for ecological and aggregate studies*, J. Roy. Statist. Soc. Ser. A 164 (2001), pp. 119–37.
- [63] J.C. Wakefield, N.G. Best, and L. Waller, *Bayesian approaches to disease mapping*, in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best, and D.J. Briggs, eds., Oxford: Oxford University Press, 2000, pp. 99–114.
- [64] L. Waller and B. Carlin, *Disease mapping*, in *Handbook of Spatial Statistics*, A.E. Gelfand, P.J. Diggle, M. Fuentes, and P. Guttorp, eds., Boca Raton, FL. CRC Press, 2010, pp. 217–244.

Table 1. The number of advanced and non-advanced breast cancer (BC) cases included in Study 1 and Study 2 by individual and areal characteristics.

		Study 1		Study 2	
		#advanced BC	#non-advanced BC	#advanced BC	#non-advanced BC
Age group (AGE)					
1	< 30	16	53	8	39
2	30 – 34	51	165	40	118
3	35 – 39	141	414	88	277
4	40 – 44	219	782	147	517
5	45 – 49	305	1189	187	834
6	50 – 54	306	1383	205	924
7	55 – 59	275	1405	191	979
8	60 – 64	196	1191	134	797
9	65 – 69	160	1028	123	712
10	70 – 74	122	903	87	629
11	75 – 79	155	733	115	526
12	80+	175	1078	128	831
	Sum	2121	10322	1452	7183
Occupation (OCCUP)					
1	Blue collar	122	435	78	277
2	White collar	409	1730	262	1241
3	Professional	536	2179	396	1600
4	Not in workforce	822	4281	540	2779
5	Unknown	232	1697	176	1286
	Sum	2121	10322	1452	7183
Marital status (MARITAL)					
1	Single	202	769	160	599
2	Married	1286	6096	856	4134
3	Widowed	378	1947	267	1433
4	Divorced	180	953	115	668
5	Separated	55	284	40	176
6	Not stated	21	273	14	173
	Sum	2121	10322	1452	7183
Indigenous status (INDIG)					
1	Non-indigenous	1973	8743	1345	6032
2	Indigenous	20	51	11	26
3	Not stated	128	1528	96	1125
	Sum	2121	10322	1452	7183
Year of diagnosis (YEAR)					
1	1996 – 1998	250	2009	177	1498
2	1999 – 2001	465	2119	338	1555
3	2002 – 2004	470	2189	327	1506
4	2005 – 2007	549	2318	362	1507
5	2008 – 2009	387	1687	248	1117
	Sum	2121	10322	1452	7183
Index of relative socio-economic disadvantage (IRSD)^a					
1	Quintile 1 (most disadvantaged)	150	565	87	283
2	Quintile 2	445	2068	84	414
3	Quintile 3	314	1549	275	1404
4	Quintile 4	541	2698	427	2109
5	Quintile 5 (least disadvantaged)	670	3441	579	2972
	Sum	2121	10322	1452	7183
Geographic remoteness (ARIA)^b					
1	Major city	1842	8774	NA ^c	NA
2	Inner regional	279	1547	NA	NA
3	Outer regional	0	0	NA	NA
4	Remote	0	1	NA	NA
5	Very remote	0	0	NA	NA
	Sum	2121	10322	1452	7183

^a Socio-economic disadvantage of the SLAs was measured using the Index of Relative Socio-economic Disadvantage (IRSD) calculated by Australian Bureau of Statistics [1].

^b ARIA+ classification [2] was used to categorize the remoteness of residence when diagnosed as having breast cancer.

^c ARIA is omitted from Study 2 as the study region (Brisbane) shows only very slight variation in geographic remoteness.

Table 2. The number of regions or grid cells included in Study 1 and the regions/cells without cases.

Spatial scale	#regions/cells in the study	#regions/cells with zero cases
SLA	308	102
10×10	100	11
20×20	400	142
30×30	900	472
50×50	2500	1706
100×100	10000	8293

Table 3. The number of regions or grid cells included in Study 2 and the regions/cells without cases.

Spatial scale	#regions/cells in the study	#regions/cells with zero cases
SLA	163	0
10×10	100	11
20×20	400	90
40×40	1600	623
50×50	2500	1129
100×100	10000	6948

Table 4. Comparison of the DIC and LS at the grid 100×100 level to the SLA level.

Random effects	Scores	Model 1	Model 2	Model 3	Model 4	Model 5
u	DIC	0.005 ^a	0.005	0.006	0.017	0.006
	LS	0.005	0.006	0.006	0.017	0.006
$u + v$	DIC	0.003	0.003	0.003	0.020	0.003
	LS	0.003	0.003	0.003	0.020	0.004

^a The values are in percentages and are calculated as follows:

$$\text{Percentage difference} = \left| \frac{\text{Score at the grid } 100 \times 100 - \text{score at the SLA level}}{\text{Average of both scores}} \right| \times 100\%.$$

Table 5. Posterior probability ($P(\beta_p < 0)$) of the regression parameters for Model 5 (IRSD).

Parameters	SLA		10×10		20×20		30×30		50×50		100×100	
	u	$u+v$	u	$u+v$	u	$u+v$	u	$u+v$	u	$u+v$	u	$u+v$
Intercept	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
IRSD.1	0.272	0.272	0.287	0.295	0.272	0.276	0.268	0.270	0.268	0.269	0.267	0.268
IRSD.2	0.313	0.307	0.285	0.280	0.270	0.260	0.269	0.259	0.270	0.259	0.270	0.259
IRSD.3	0.999	0.998	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
IRSD.4	0.885	0.877	0.885	0.881	0.881	0.875	0.886	0.880	0.887	0.882	0.888	0.882

Figure 1. Variogram at all spatial scales for Study 1.

Figure 2. Variogram at all spatial scales for Study 2.

Figure 3. The DIC and LS for Models 1 to 5. Two settings of random effects for each spatial scale are: u (right), $u + v$ (left). Model 2 and Model 5 have similar DIC values that are smaller than the other models. The LS gives similar results.

Figure 4. (a) Box plots of standard deviation of the estimated linear predictor ($\text{logit}(p_{hij})$) of Model 5; (b) box plots of standard deviation of the estimated unstructured random effects (σ_u) of Model 5; (c) box plots of standard deviation of the estimated spatially structured random effects (σ_v) of Model 5. For (a) and (b), two settings of random effects for each spatial scale are: u (right), $u + v$ (left). The precision of estimation of u and v at the grid levels is noticeably higher than at the SLA level based on the smaller standard deviation observed.

Figure 5. Standard deviation of the posterior of the parameter estimates of Model 5 (IRSD). Two settings of random effects for each spatial scale are: u (right), $u+v$ (left). A larger standard deviation indicates a larger spread of the posterior distribution. The standard deviation decreases gradually from the SLA level to the decreasing grid cell size.

Figure 6. Plots of the posterior mean of the spatial effect at various spatial scales for Model 5 (IRSD) with random effects $\mathbf{u} + \mathbf{v}$. At finer grid cells, more localized excess risk is identified.

Figure 7. Plots of the aggregate risk at various levels of aggregation for Model 5 (IRSD) with random effects $u + v$. A more localized distribution of aggregate risk is produced at the fine grid cells.

Figure 8. The DIC and LS for Models 1 to 4. Two settings of random effects for each spatial scale are: u (right), $u + v$ (left). Model 3 and Model 4 have similar LS that are smaller than the other models.

Table 6. Model performance at various spatial scales for Study 1 and Study 2

	Study 1: Spatially sparse data	Study 2: Spatially dense data
Model fit - DIC	The lowest DIC at the SLA level but improvements in DIC are seen as the grid cell size reduces	Similar DIC at both the grid levels and the SLA level; no improvement seen as the grid cell size reduces
Predictive performance - LS	The lowest score at the SLA level but improvements in score are seen as the grid cell size reduces	Similar score at both the grid levels and the SLA level; no improvement seen as the grid cell size reduces
Estimation for linear predictor and random effects - Standard deviation and width of the 90% credible interval	Poorest estimation at the SLA level; improvements in precision of estimation as the grid cell size reduces; estimation is worse at fine grid cell (100×100)	Precision of estimation is very similar at both the grid levels and the SLA level; no improvement seen as the grid cell size reduces
Estimation of the regression parameters - Standard deviation and width of the 90% credible interval	Poorest estimation at the SLA level; improvements in precision of estimation as the grid cell size reduces; for some models, the estimation is worse at small grid cells including 30×30 , 50×50 and 100×100	Precision of estimation fluctuates across various spatial scales; no improvement seen as the grid cell size reduces
Recommendation on choice of spatial scale	Grid partitions seem to perform better; investigate a range of spatial scales to select the appropriate scale	Performance at the SLA level and the grid levels are similar, there is no apparent advantage in a finer partition