Journal of the Korean Data & Information Science Society
2016, **27**(5), 1225–1239

http://dx.doi.org/10.7465/jkdi.2016.27.5.1225
한국데이터정보과학회지

Modeling pediatric tumor risks in Florida with conditional autoregressive structures and identifying hot-spots[†]

Bit Kim[1] · Chae Young Lim[2]

[12]Department of Statistics, Seoul National University

Received 13 July 2016, revised 4 September 2016, accepted 8 September 2016

Abstract

We investigate pediatric tumor incidence data collected by the Florida Association for Pediatric Tumor program using various models commonly used in disease mapping analysis. Particularly, we consider Poisson normal models with various conditional autoregressive structure for spatial dependence, a zero-inflated component to capture excess zero counts and a spatio-temporal model to capture spatial and temporal dependence, together. We found that intrinsic conditional autoregressive model provides the smallest Deviance Information Criterion (DIC) among the models when only spatial dependence is considered. On the other hand, adding an autoregressive structure over time decreases DIC over the model without time dependence component. We adopt weighted ranks squared error loss to identify high risk regions which provides similar results with other researchers who have worked on the same data set (e.g. Zhang *et al.*, 2014; Wang and Rodriguez, 2014). Our results, thus, provide additional statistical support on those identified high risk regions discovered by the other researchers.

Keywords: Bayesian methods, conditional autoregressive model, disease mapping, pediatric tumor risk model, spatio-temporal model.

1. introduction

The Florida Association for Pediatric Tumor Programs (FAPTP) collected pediatric cancer cases for 11 years to monitor and manage those cases occurred in Florida. The data from year 2000 to 2010 are summarized based on zip code tabulation with age, sex and race for each case.

Several researchers have worked on the same data using different statistical methods to identify regions with potentially high risk. The FAPTP wants identified high risk regions to be statistically strongly supported which could be achieved by getting common findings from independent researchers who have analyzed the data with same aim but using different statistical methods.

[†] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B4008237).

[1] Ph.D. Program, Department of Statistics, Seoul National University, Seoul 08826, Korea.
[2] Corresponding author: Professor, Department of Statistics, Seoul National University, Seoul 08826, Korea. E-mail: limc@stats.snu.ac.kr

These analysis attempts are published in *Statistics and Public Policy*, a newly opened journal of American Statistical Association as the collection on Analyses of Florida Pediatric Cancer Data in 2014. Five articles are included in the issue. Waller (2015) discussed these five articles on various perspectives including existence of sustained clusters over space-time, impacts of covariates such as race and insights on potential risk factors.

Heaton (2014) investigated spatial and demographic factors in regions with rapid changes on cancer incidence rates. The data were modeled by marked point process (Diggle (2003)) using available covariates to find out if there is any pattern. The intrinsic autoregressive structure was considered for the spatial heterogeneity in an intensity function. High cancer occurrence was detected by assigning rapid increasing boundaries which were obtained from the estimated intensity function. The race as a covariate strongly impacts on high risk and younger population (under 19 years old) is also recognized for being exposed to the high risk.

Amin *et al.* (2014) employed the software SaTScan to identify surveillance clusters based on hypothesis testing for three common pediatric cancers (brain tumors, leukemia and lymphoma). They also conducted nonparametric permutation tests for each cancer in spatio-temporal analysis as well as multivariate analysis for three cancer types together. The results showed a significant spatial cluster in south Florida for each cancer type. Two significant spatial clusters appear in south and central north regions when all cancer type were considered.

Wang and Rodriguez (2014) adopted the penalized generalized linear model to find out a cluster. The data were fitted with Poisson regression to log-linear model having a conditionally autoregressive structure, which is commonly used in disease mapping. They estimated disease risks using log-likelihood which is regularized by the fused Lasso penalty. As a result, the map with age-race adjusted risks represented the highest cancer risks in some areas of southeast Florida. Also, there are several small-sized high risk regions in Northern Florida.

Lawson and Rotejanaprasert (2014) applied the standard Poisson convolution (SPC) model to pediatric brain cancer data and considered zero-inflation with a factored intercept. They mapped exceedance probability of relative risk greater than 1, 0 or 2, or residuals greater than 0 for each region with their SPC model and the model by Besag, York, and Mollié (1991). High exceedance probability areas are dispersed across the state but west of Orlando which is located in central Florida has some concentration of high exceedence probability.

Zhang *et al.* (2014) applied a Bayesian hierarchical model with spatial clustering prior for log age-adjusted rates which were related to covariates differently by each cluster. Markov Chain Monte Carlo (MCMC) was implemented to obtain cluster configuration. Resulting maps contain areas clustered with risk for each year. High-risk areas were identified in southwest, northeast and northwest Florida.

The above mentioned articles have applied various statistical methods to the pediatric cancer data in Florida by the FAPTP to detect a contributing factor affecting high cancer risks and their clustered areas. In this paper, we also investigated the FAPTP data to find out contributing covariate information and identify high risk areas. We started from basic Poisson normal models with conditional autoregressive (CAR) structures, where we investigated three different CAR models. Then, we investigated a zero-inflated Poisson normal model to see if we need a zero-inflated component since there are relatively many areas with zero counts. Lawson and Rotejanaprasert (2014) also considered a model that can accom-

modate excess zero count in data but their model is different from a zero-inflated model in that zero count information in their model is incorporated in the Poisson mean while a zero-inflated model consider a mixture of a Poisson model and zero mass.

Finally, we considered a spatio-temporal Poisson normal model with a CAR structure for spatial dependence and an autoregressive error structure for time dependence. A Bayesian MCMC method was used for inference and we compare models by Deviance Information Criterion (DIC). Then, we consider a weighted ranks squared error loss (WRSEL) proposed by Wright *et al.* (2003) to identify high relative risk regions as a posterior analysis.

The rest of the paper is organized as follows: Section 2 introduces various Poisson normal models that we are considering and Bayesian approaches for inference. Section 3 provides description of the FAPTP data and the analysis results followed by discussion in Section 4.

2. Models and methods

The data we are going to analyze are disease incidence counts over regions and years, which have many zero counts as well. To investigate spatial and spatio-temporal structure of the data, we consider various Poisson-normal models with CAR structures. Specifically, we consider Poisson-CAR models in which an normal error term has CAR covariance structures to give spatial dependency among regions, a zero inflated Poisson-CAR model to take into account excess zero counts and a Poisson-CAR model with autoregressive structure over time to model time dependence.

We take Bayesian Markov Chain Monte Carlo approach for inference of the introduced models. Thus, we briefly introduce prior specifications of hyper-parameters, Deviance Information Criterion for model comparisons and weighted ranks squared error loss to estimate relative risks which will be used to identify hot-spots as well.

2.1. Poisson-CAR models

Let Y_i be the disease counts in the region i. To introduce Poisson-CAR models, we first consider the following hierarchical model structure: for $i = 1, \cdots, N$,

$$\begin{aligned} Y_i \,|\, \theta_i &\sim \text{Poisson}(E_i\, \theta_i), \\ \log(\theta_i) &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{Q}^{-1}), \end{aligned} \tag{2.1}$$

where E_i denotes the expected number of cases at risk in the region i. θ_i is relative risk for each region. We use a log link function to incorporate covariate information and spatial dependency. $\boldsymbol{x}_i$ denotes an appropriate covariate vector for the model setting. $\mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ denotes N dimensional multivariate normal distribution with a mean vector, $\boldsymbol{\mu}$, and a precision matrix, $\boldsymbol{Q}$. A CAR structure is given in $\boldsymbol{Q}$. There are a number of different CAR variants and we consider the following three CAR models.

2.1.1. Intrinsic CAR (ICAR) model

Besag (1974) proposed the following CAR model for the spatial dependence structure when the data are observed on regions.

$$\boldsymbol{Q} = \frac{1}{\sigma^2}(\boldsymbol{M}^{-1} - \boldsymbol{W}), \tag{2.2}$$

where $\boldsymbol{W}$ is an $N \times N$ adjacency matrix whose (i_1, i_2) entry has value 1 if i_1 and i_2 are neighbor or 0 if they are not. Diagonal entries of $\boldsymbol{W}$ are zero. Thus, $\boldsymbol{W}$ has neighborhood structure of regions of interest. $\boldsymbol{M}^{-1}$ is the $N \times N$ diagonal matrix whose diagonal entries are row sums of $\boldsymbol{W}$. This structure has improper density function since $\boldsymbol{M}^{-1} - \boldsymbol{W}$ has rank $N-1$ (c.f. $(\boldsymbol{M}^{-1} - \boldsymbol{W})\mathbf{1} = \mathbf{0}$). Although the density is improper, the posterior distributions can be well defined when it is used in a inner layer of hierarchy.

2.1.2. Proper CAR model I

ICAR has no parameter to control spatial dependence and induce improper distribution. This impropriety problem can be resolved by modifying the spatial precision matrix, $\boldsymbol{Q}$. The one approach is to assume $\boldsymbol{Q} = \frac{1}{\sigma^2}\boldsymbol{D}(\lambda)$, where $\boldsymbol{D}(\lambda) = \lambda(\boldsymbol{M}^{-1} - \boldsymbol{W}) + (1-\lambda)\boldsymbol{I}$ with $0 \leq \lambda < 1$, which was proposed by Leroux *et al.* (2000). This CAR model includes an independent case which corresponds to the case with $\lambda = 0$ and is getting close to the ICAR model which corresponds to the case with $\lambda = 1$. We denote this approach a proper CAR model I.

2.1.3. Proper CAR model II

The other approach is to assume $\boldsymbol{Q} = \frac{1}{\sigma^2}(\boldsymbol{M}^{-1} - \gamma\boldsymbol{W})$. When γ tends to 1, this model becomes the ICAR model. When γ tends to 0, the spatial dependency vanishes and ϵ_i is distributed as an independent Normal distribution. Since (i, i) entry of the diagonal matrix $\boldsymbol{M}^{-1}$ is the number of neighbors in the i^{th} region, ϵ_i with $\gamma = 0$ is not identical compared to the Proper CAR I and has less variance as the region i has more neighbors. This implies that it is not completely ignoring spatial information because each variance depends on the number of neighbors. When the region i has many neighbors, it gets more information from its neighbors and uncertainty becomes weak. On the other hand, necessary condition for standing Proper CAR II is that $\gamma \in (\lambda_{min}^{-1}, \lambda_{max}^{-1})$, where λ_{min} and λ_{max} are minimum and maximum eigenvalues of $\boldsymbol{M}^{1/2}\boldsymbol{W}\boldsymbol{M}^{1/2}$, respectively (Banerjee *et al.*, 2013). Also note that the range for γ guarantees to include 0 since $\text{tr}(\boldsymbol{M}^{1/2}\boldsymbol{W}\boldsymbol{M}^{1/2}) = 0$.

2.2. Zero-inflated Poisson CAR model

When there are many regions with zero counts, Poisson-normal model may not be enough to capture excess zero counts. One approach to model such characteristics is a zero-inflated Poisson (ZIP) model. Since the data have spatial information, we consider zero-inflated Poisson CAR model, where, in particular, we assume a ICAR model for spatial dependence structure. The ZIP regression with the ICAR covariance model takes the following form:

$$\begin{aligned} Y_i \,|\, \pi_i, \theta_i &= \begin{cases} 0, & \text{with probability } \pi_i \\ \text{Poisson}(E_i\,\theta_i), & \text{with probability } 1-\pi_i, \end{cases} \\ \log(\theta_i) &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{z}_i^\top \boldsymbol{\alpha} + \eta_i. \end{aligned} \tag{2.3}$$

Here $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are appropriate covariate vectors for relative risks and zero probabilities, respectively. We assume a ICAR model introduced in 2.1.1 for spatial covariance structure

of ϵ_i. Since zero counts may be also spatially related, we consider a ICAR model for spatial covariance structure of η_i as well. That is, we assume

$$\begin{aligned} \boldsymbol{\epsilon} &\sim \mathcal{N}_N\left(0, \sigma_\epsilon^2(\boldsymbol{M}^{-1} - \boldsymbol{W})^{-1}\right), \\ \boldsymbol{\eta} &\sim \mathcal{N}_N\left(0, \sigma_\eta^2(\boldsymbol{M}^{-1} - \boldsymbol{W})^{-1}\right), \end{aligned} \tag{2.4}$$

where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_N)^\top$. The other notations were introduced in 2.1.1.

2.3. Spatio-temporal Poisson CAR model with an autoregressive error in time

In the previous sections, we have considered the models for the spatial count data. When the spatial count data are observed over time, which is the case for our real data application, it is natural to consider time dependency in the model as well. There are various ways to model such time dependency but we will consider the following model to take care of both spatial and temporal dependency.

Let Y_{it} be the disease count in the region i at time t for $i = 1, \cdots, N$ and $t = 1, \cdots, T$.

$$\begin{aligned} Y_{it} \,|\, \theta_{it} &\sim \text{Poisson}(E_{it}\,\theta_{it}), \\ \log(\theta_{it}) &= \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \epsilon_{it}, \\ \boldsymbol{\epsilon}_t \,|\, \boldsymbol{\epsilon}_{t-1} &\sim \mathcal{N}_N\left(\rho\,\boldsymbol{\epsilon}_{t-1}, \sigma^2 \boldsymbol{D}^{-1}(\lambda)\right), \quad t > 1, \\ \boldsymbol{\epsilon}_1 &\sim \mathcal{N}_N\left(0, \sigma^2 \boldsymbol{D}^{-1}(\lambda)\right), \\ \boldsymbol{D}(\lambda) &= \lambda(\boldsymbol{M}^{-1} - \boldsymbol{W}) + (1-\lambda)\boldsymbol{I}, \end{aligned} \tag{2.5}$$

where $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \cdots, \epsilon_{Nt})^\top$ and ρ is a parameter for temporal autocorrelation with $|\rho| < 1$, i.e. we consider AR(1) model for temporal dependence structure. When $\rho = 0$, $\boldsymbol{\epsilon}_t$ becomes independent of each time t. A larger value of ρ implies stronger temporal autocorrelation. This model is considered in Rushworth *et al.* (2014).

Note that a spatial dependence structure in zero-inflated models and spatio-temporal models can be modeled in various ways. For example, any of the models introduced in 2.1 can be used.

2.4. Bayesian inference

2.4.1. Prior and posterior distributions

To proceed with Bayesian MCMC methods for the inference, we need specifications on prior distributions. We tried to use conjugate priors wherever possible for computational simplicity with larger variances to make them vague priors. The prior distributions for the parameters of the models introduced in the earlier sections are

$$\begin{aligned} \alpha_j, \beta_j &\sim \mathcal{N}(\mu_0, 1/\sigma_0^2), \\ \gamma &\sim Uniform(\lambda_{\min}^{-1}, \lambda_{\max}^{-1}), \\ \rho &\sim Uniform(-1, 1), \\ \lambda &\sim Uniform(0, 1), \\ 1/\sigma^2, 1/\sigma_\epsilon^2, 1/\sigma_\eta^2 &\sim Gamma(a_0, b_0), \end{aligned} \tag{2.6}$$

where $Gamma(a, b)$ denotes Gamma distribution with shape a and scale b. α_j and β_j are jth regression coefficients for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. For hyper-parameters, we set $\mu_0 = 0$, $\sigma_0 = 0.1$, $a_0 = 0.1$ and $b_0 = 0.1$ so that variance of prior distributions are relatively large. This setting enables us to draw MCMC samples of parameters easily since posterior distributions for α_j, β_j, $1/\sigma^2$, $1/\sigma_\epsilon^2$, and $1/\sigma_\eta^2$ are known distributions (Normal and Gamma). λ and γ can be sampled by using the inverse CDF method since the posterior distributions are not known distribution. For updating θ_i, we use Metropolis-Hasting (MH) algorithm with an appropriate proposal density. Updating π_i for the ZIP-ICAR model can be done in a similar way.

2.4.2. Deviance Information Criterion (DIC)

As we are investigating several models, we consider model comparison to see which model fits the data better. For comparison among models, we consider Deviance Information Criterion (DIC), which is a useful measure in Bayesian (hierarchical) model selection when MCMC samples are obtained. Define deviance as $\mathcal{D}(\Theta) = -2\log(p(\boldsymbol{Y}|\Theta)) + c$, where $\boldsymbol{Y}$ denotes the observed data, $p(\boldsymbol{Y}|\Theta)$ is the corresponding likelihood given parameters Θ and c is an arbitrary constant term which will be canceled out in further calculation. Spiegelhalter *et al.* (2002) proposed DIC as $\bar{\mathcal{D}}(\Theta) + p_{\mathcal{D}}$, where $\bar{\mathcal{D}}(\Theta) = \mathbf{E}_\Theta[\mathcal{D}(\Theta)]$ and $p_{\mathcal{D}}$ is defined as $\bar{\mathcal{D}}(\Theta) - \mathcal{D}(\bar{\Theta})$, where $\bar{\Theta}$ is the posterior mean of the parameters. $p_{\mathcal{D}} \geq 0$ can be induced directly by Jensen's inequality if $\mathcal{D}(\Theta)$ is convex.

If the model complexity increases, $p_{\mathcal{D}}$ also does but $\bar{\mathcal{D}}(\Theta)$ decreases. Smaller $\bar{\mathcal{D}}(\Theta)$ is thought to be a good fit and $p_{\mathcal{D}}$ gives penalty to the model complexity. Thus, a model having smaller DIC value is recommended because it can offer a good performance with moderate complexity. We are able to compare previously introduced models through DIC because of MCMC samples. Denote the r^{th} MCMC sample by $\Theta^{(r)}$ and suppose that R MCMC samples are obtained. Then, estimates for the posterior mean of deviance and the deviance of posterior mean for parameters are given as

$$\begin{aligned} \widehat{\bar{\mathcal{D}}(\Theta)} &= \frac{1}{R}\sum_{r=1}^{R}\mathcal{D}(\Theta^{(r)}), \\ \mathcal{D}(\hat{\bar{\Theta}}) &= \mathcal{D}\left(\frac{1}{R}\sum_{r=1}^{R}\Theta^{(r)}\right) \end{aligned} \tag{2.7}$$

and the DIC can be estimated from the above estimates.

2.4.3. Estimating relative risks

It is common to use posterior means by minimizing squared error loss in point estimation of parameters in Bayesian methods. The histogram of posterior means over entire regions has less variation than a sample drawn from posterior distribution. Louis (1984) stated posterior means have been shown to be underdispersed in this regard. In disease mapping analysis, we are interested in estimating relative risks or log relative risks and finding high risk regions by comparing estimated relative risks over regions. Such high risk regions are so-called hot-spots. Using posterior means as estimates can underestimate relative risks, which is not desirable when finding hot-spots. Wright *et al.* (2003) introduced weighted ranks squared error loss (WRSEL) to address the underdispersion issue.

Let denote the vector of relative risks by $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)^\top$ and its estimate is denoted by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_N)^\top$. Let $\boldsymbol{c} = (c_1, \cdots, c_N)^\top$ be a weight vector that corresponds to the ranks. Then, the WRSEL is defined as

$$W(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\,;\, \boldsymbol{c}) := \sum_{k=1}^{N} \sum_{j=1}^{N} c_j I_{\{\theta_k = \theta_{(j)}\}} \left(\theta_k - \hat{\theta}_k\right)^2, \tag{2.8}$$

where $\theta_{(j)}$ denotes the jth order statistic in $\boldsymbol{\theta}$. This expression implies that a different weight is applied to each loss depending on it's rank. The Bayes estimate of $\boldsymbol{\theta}$ using WRSEL is

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbf{E}\left[W(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}; \boldsymbol{c}) \;\middle|\; \boldsymbol{Y} \right]. \tag{2.9}$$

Thus, the estimated relative risk, $\hat{\theta}_k$, for $k = 1, \cdots, N$ is obtained by (2.9) and takes the following form:

$$\hat{\theta}_k = \frac{\sum_{j=1}^{N} \mathbf{E}\left(\theta_k | \theta_k = \theta_{(j)}, \boldsymbol{Y}\right) c_j \, \mathbf{P}\left(\theta_k = \theta_{(j)} | \boldsymbol{Y}\right)}{\sum_{j=1}^{N} c_j \, \mathbf{P}\left(\theta_k = \theta_{(j)} | \boldsymbol{Y}\right)}.$$

Expectations and probabilities can be evaluated from MCMC samples, since they represent the posterior distribution itself.

The choice of the weight vector, $\boldsymbol{c}$, can be appropriately determined depending on the purpose. When $\boldsymbol{c} = a(1, \cdots, 1)^\top$ for arbitrary $a > 0$, WRSEL becomes exactly same as the conventional squared error loss and estimates for relative risks are posterior means. If the purpose is finding hot-spots, the choice of $\boldsymbol{c}$ can be extreme such as $c_j = 0$ for $j = 1, \cdots, N - m$ and $c_j = 1$ for $j = N - m + 1, \cdots, N$. This gives extreme weights on last m components in order statistics; therefore, potentially high risk regions can be captured by ordering estimates from such an extreme weight vector.

3. Analysis on pediatric tumor data in Florida

3.1. Pediatric tumor data in Florida

The data consist of pediatric cancer case counts from the Florida Association for Pediatric Tumor Programs (FAPTP). They are surveyed yearly by zip code tabulation areas (ZCTAs) from 2000-2010. Counts of the cases were collected over four age groups: 0-4 years, 5-9 years, 10-14 years and 15-19 years for each zip code.

Corresponding population data are obtained from the 2000 and 2010 census from U.S. Census Bureau which provides population of each age group for ZCTAs in 2000 and 2010. Population data for the years 2001-2009 were obtained by using linear interpolation because census has taken only in every 10 years. Some zip codes (about 7.5%) are only available in 2000 but not 2010 and vice versa. In these cases, it can not be interpolated. So, we use either only 2000 year data or only 2010 year data for the population as a constant over 11 years.

Finally we consider 983 zip code areas with corresponding case counts and population. In addition, we consider some demographic information (e.g. proportion of white population) for each zip code as a covariate. Each ZIP code areas are indexed to $i = 1, \cdots, 983$ $(= N)$, age groups are indexed to $j = 1, \cdots, 4$ and $t = 1, \cdots, 11$ $(= J)$ indicates years. Notice that the observed count data have about 60∼65% ZCTAs with zero counts in each year.

For Poisson-CAR models, we need to calculate expected number of cases at risk as an offset in each area at time t. If the cases and population are separated into age groups denoted by $j = 1, \cdots, J$, expected number of cases at risk at time t is given by $E_{ijt} = \frac{\sum_{i=1}^{N} d_{ijt}}{\sum_{i=1}^{N} p_{ijt}} \times p_{ijt}$. By adding up for js, we get age-adjusted rates (ADR) as E_{it} for each area i and at time t, i.e. $E_{it} = \sum_{j=1}^{J} E_{ijt}$. Here, d_{ijt} and p_{ijt} denote pediatric cancer counts and population on region i, jth age group and in year t, respectively.

Available covariates for $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are demographic information such as proportion of white population (race) and proportion of male population (sex). However, for most regions over the years, the proportions of male population are around 0.5 so we did not consider sex covariate because of a possible identifiability issue with an intercept term.

3.2. Results

We fit models described in Section 2 using the FAPTP data. For the ZIP Poisson ICAR model, log and logit link functions to include covariates takes the following form:

$$\begin{aligned} \log(\theta_i) &= \beta_0 + \text{WHITE}_i\, \beta_1 + \epsilon_i \\ \text{logit}(\pi_i) &= \alpha_0 + \text{WHITE}_i\, \alpha_1 + \eta_i, \end{aligned} \tag{3.1}$$

where WHITE_i denotes proportion of white population in the ith region.

We run Bayesian MCMC for the considering models per each year separately except the spatio-temporal model in subsection 2.3. We ran 3 parallel chains with 10000 iterations per each chain, discarded the first 5000 samples as burn-in and chose the samples by considering thinning interval of size 10. Finally, we get 1500 MCMC samples for each model and time period.

Table 3.1 DIC values of year 2000-2010 for proper CAR I, proper CAR II, ICAR and ZIP with ICAR. The values in parentheses are p_D

year	CAR I	CAR II	ICAR	ZIP.ICAR
2000	1675.6 (16.4)	1703.6 (108.9)	1684.5 (13.5)	1689.5 (91.3)
2001	1663.0 (6.29)	1679.3 (94.3)	1660.3 (10.6)	1674.7 (87.0)
2002	1725.5 (6.32)	1734.8 (102.3)	1717.2 (9.33)	1729.7 (91.0)
2003	1726.3 (28.8)	1718.0 (137.2)	1750.8 (4.37)	1752.7 (102.1)
2004	1676.6 (4.06)	1673.5 (115.1)	1658.3 (24.7)	1657.4 (98.2)
2005	1740.5 (8.93)	1733.6 (116.9)	1737.4 (16.0)	1744.3 (92.4)
2006	1807.4 (50.0)	1815.9 (118.7)	1814.4 (11.1)	1819.2 (101.2)
2007	1799.5 (4.99)	1810.3 (107.1)	1791.2 (7.54)	1808.4 (96.1)
2008	1802.7 (19.7)	1817.2 (115.7)	1808.8 (4.32)	1811.5 (99.3)
2009	1804.8 (6.63)	1807.0 (111.3)	1804.4 (3.65)	1810.6 (101.1)
2010	1731.3 (3.63)	1736.8 (107.4)	1715.5 (11.6)	1722.4 (94.3)

Table 3.1 shows DIC values calculated from 1500 MCMC samples for each model and year. For the comparison among Poisson CAR models, ICAR has smaller DIC values over the years overall. Secondly, proper CAR I has moderate performance but proper CAR II has relatively large values of DIC except the year 2003. On the other hand, ICAR model which has the best fit among three Poisson CAR models still offers smaller DIC values compared to DIC values from the ZIP Poisson ICAR model with exception of the year 2004.

At first, we suspected that the zero-inflated model would be preferable since the data contain over 60% of zip code areas with zero counts almost every year. However, results

show that the Poisson model with spatial covariance structure is enough to fit the FAPTP data according to DIC. ZIP Poisson ICAR model might fit better than simpler models but the increased accuracy of fit, $\bar{\mathcal{D}}(\Theta)$, seems less than the quantity of increased model complexity, $p_{\mathcal{D}}$, in DIC. Thus, Poisson ICAR model is rather better than the mixture model to deal with such a vague zero-inflation in the view of model complexity. We also tried to consider proper CAR structure in the ZIP setting. However, due to the increased number of parameters to estimate, we found the MCMC samples are rather unstable so that we did not report them in the paper.

Table 3.2 Estimated posterior means and 95% credible interval for the parameters of the Poisson ICAR model from MCMC samples

Estimated parameters of the ICAR model						
	β_0 (Intercept)		β_1 (WHITE)		σ	
Year	mean	95% interval	mean	95% interval	mean	95% interval
2000	-0.4179	(-0.8181, 0.0250)	0.4739	(-0.0015, 0.9521)	4.3685	(2.2622, 7.8567)
2001	-0.7851	(-1.1939, -0.3935)	0.9913	(0.4971, 1.4856)	5.4623	(3.2951, 10.2062)
2002	-0.6655	(-1.0709, -0.2667)	0.8213	(0.3502, 1.3247)	6.7729	(3.5245, 12.0386)
2003	-0.0411	(-0.3395, 0.2712)	0.0532	(-0.3458, 0.4442)	10.5999	(6.0412, 20.000)
2004	-0.3459	(-0.7331, 0.0398)	0.3957	(-0.0937, 0.8817)	2.7493	(1.8270, 6.4685)
2005	-0.1783	(-0.5319, 0.1560)	0.1968	(-0.2408, 0.6379)	4.3274	(2.5327, 6.6667)
2006	-0.2945	(-0.6088, 0.0290)	0.3464	(-0.0764, 0.7646)	5.7166	(3.9684, 9.1287)
2007	-0.1467	(-0.4681, 0.1780)	0.1447	(-0.2790, 0.5735)	7.2357	(4.6127, 12.3091)
2008	-0.1683	(-0.5124, 0.1331)	0.1949	(-0.1868, 0.6257)	11.7851	(6.7729, 16.4399)
2009	-0.1941	(-0.5140, 0.1200)	0.2679	(-0.1357, 0.6747)	13.2453	(7.4329, 21.3201)
2010	-0.1317	(-0.4769, 0.1684)	0.1170	(-0.2684, 0.5483)	5.3683	(3.4483, 8.1111)

In summary, we found that the Poisson ICAR model fits the data better compared to the other models based on DIC from the comparison among four models in Table 3.1. So, we focus on the Poisson ICAR model. Table 3.2 shows the estimated posterior means and 95% credible intervals of the parameters from MCMC samples of the Poisson ICAR model. Parameters are β_0, β_1 and σ. As we consider the model for each year separately, they are evaluated yearly. Their posterior means and medians are close to each other. Hence we only present posterior means.

From the estimates of β_0 and β_1 with a log link function, relative risk θ_i tends to be less than 1. This leads to the incidence rate less than the expected rate which makes sense given there are many regions with zero counts. Estimation for the standard deviation varies more over years, which may indicate that we need to consider temporal dependence together in the model.

Figure 3.1 shows estimated relative risk, θ_i using posterior means on the map of Florida for each year from the Poisson ICAR model. Six different color gradation levels in Figure 3.1 are assigned to six interval ranges created by five different percentiles: 42, 60, 73, 85 and 95 % from the entire posterior means over years. High risk areas according to the Poisson ICAR model appears in north Florida for the years 2000, 2001, 2003, 2004 and 2005. South Florida areas show low risks compared to the north Florida; However, they possess slightly high risks in 2000 and make a huge cluster in 2002 and 2006. The one thing to pay attention is the coastal areas of south Florida. It has high risk even if overall south areas have relatively low risks for the years 2001, 2004, 2005 and 2008. On the other hand, high risk regions are sporadically located in 2007-2009. North east to central areas of Florida are estimated having higher risks than the other areas in 2010.

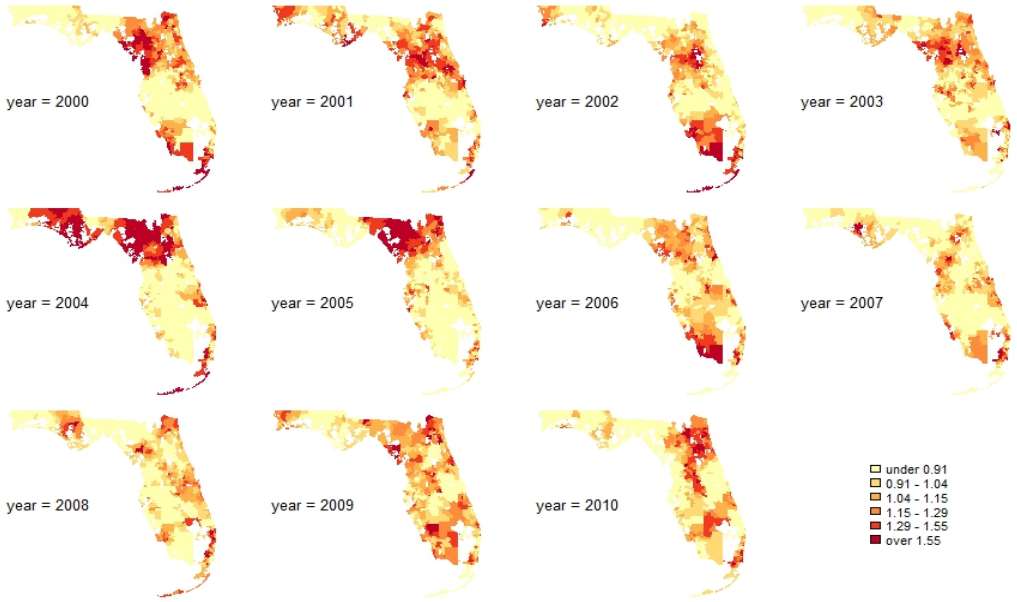

Figure 3.1 Map with estimated θ_i as posterior means from the Poisson ICAR model for 2000-2010 in Florida

We now turn into the spatio-temporal model to capture both spatial and temporal dependence of the data. We consider three different parameter settings in the spatio-temporal model. One is for fixed $\lambda = 1$ and fixed $\rho = 0$, which corresponds to ICAR for spatial dependence structure and no time dependency. The another is for fixed $\rho = 0$ only which corresponds to the proper CAR I for spatial dependence structure and no time dependency. The last case corresponds to the proper CAR I for spatial dependence structure with AR(1) time dependence structure.

Table 3.3 Estimated posterior means and 95% credible intervals for the parameters of the spatio-temporal CAR with AR(1) model. Either λ or ρ are set to the given values or estimated through the Bayesian inference.

	Estimated parameters of the spatio-temporal CAR with AR(1)					
	case 1: $\lambda = 1$, $\rho = 0$		case 2: $\rho = 0$		case 3: no fixed parameters	
	mean	95% interval	mean	95% interval	mean	95% interval
β_0	-0.2796	(-0.3865, -0.1802)	-0.2672	(-0.3765, -0.1659)	-0.2929	(-0.4218, -0.1645)
β_1	0.3353	(0.2039, 0.4683)	0.3118	(0.1808, 0.4549)	0.3110	(0.1389, 0.4747)
σ	5.6900	(4.2333, 7.1247)	3.1909	(2.6939, 3.5991)	2.9919	(2.4390, 3.4879)
λ	1	-	0.9507	(0.8897, 0.9870)	0.6371	(0.3941, 0.8740)
ρ	0	-	0	-	0.7968	(0.7186, 0.8539)

Table 3.3 shows that the estimated posterior means for β_0 and β_1 are similar for all cases and so are 95% credible intervals. On the other hand, the posterior estimates of σ are quite different. The estimated σ is larger in the case 1 compared to the case 2, which may implies that a flexible spatial dependence structure (case 2) was able to capture variability of the data more. The estimated σ in case 3 is smaller than the other two cases. However, σ in case 3 is not the error variance of the model but the variance of the innovations in the AR(1) structure for the error. Thus we can not directly compare this quantity to the σ in cases 1 and 2 which is error variance of the model.

The estimated spatial dependence parameter λ in case 2 is close to 1 while the estimated λ in case 3 is smaller than the one in case 2. This could be because some of variability of the data was captured by the temporal dependence structure in the model. The estimated temporal autocorrelation parameter, ρ, is away from zero and 95% credible interval is well above the zero. This implies that we should consider a temporal dependence structure in the model. DIC values for each case are 19143.807 (case 1), 19143.684 (case 2) and 19064.405 which also suggests that the data supports the spatial temporal model (case 3) over non-temporal models (cases 1 and 2)

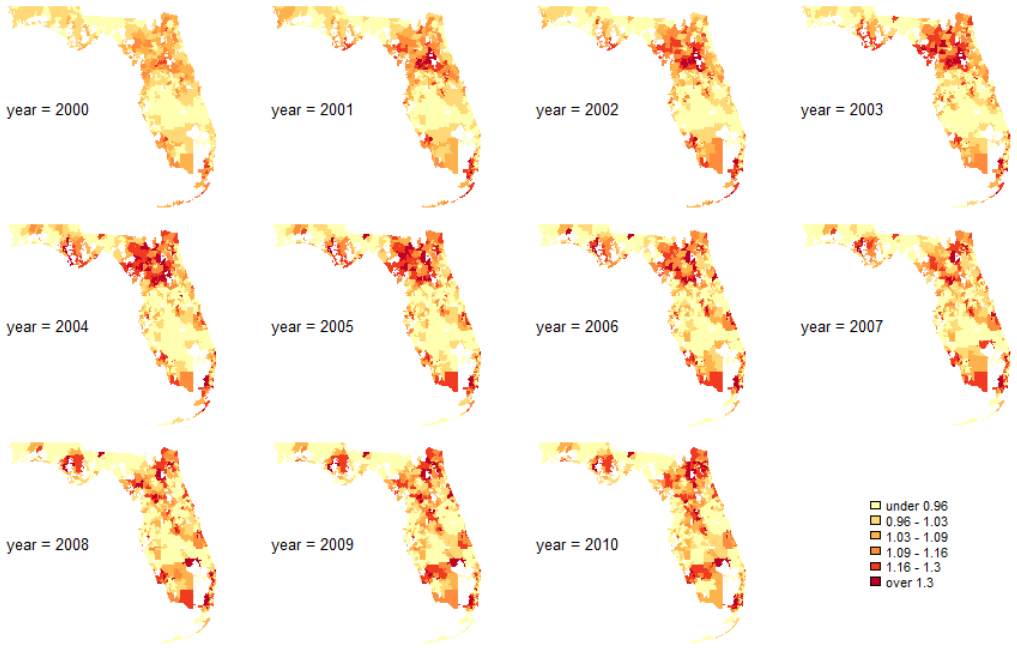

Figure 3.2 Relative risk θ_i obtained from posterior mean of the spatio-temporal CAR with AR(1) model during 2000-2010 in Florida

Figure 3.2 shows the relative risks, θ_i estimated by posterior means from the case 3 which is spatio-temporal Poisson CAR with AR(1) model. It is clearly seen that the temporal autocorrelation largely affects the estimated θ_i compared to non-temporal model in Figure 3.1. As we expected from a relatively large positive value of estimated temporal autocorrelation parameter, ρ, the estimated relative risk map for one year is somewhat similar from the one for the previous year. Also, it appears that high risk regions in each year are rather sporadic which is reflected by reduced λ estimates compared to case 2. The relative risk maps are overall more smoothed over regions compared to the maps in Figure 3.1.

We now consider estimating the relative risks, θ_i by minimizing WRSEL given in (2.8) for each model to focus on identifying hot-spots. We set $r = 10$, i.e. we give extreme weights to the last 10 components of the weight vector $\boldsymbol{c}$ which corresponds to the largest 10 θ_is. Because each estimate is evaluated from MCMC samples, it is possible for some θ_i not to be within top 10 ranks in the whole samples. We give 0 for such case in the estimation procedure. Recall

that estimates for relative risks having moderate to lower values are distorted quantitatively when we consider WRSEL because the weights are concentrated on large extreme cases. This approach, however, enables us to look up potentially high risk regions and to compare between models through ranking high risks regions. We show 25 largest estimated relative risks, θ_i using WRSEL for ICAR model and spatio-temporal Poisson CAR with AR(1) model in Figures 3.3 and 3.4. We assign 5 different color gradation levels from the red color to the yellow for each 5 ranks in the Figures. Areas under 25-rank are painted light yellow to represent a base map.

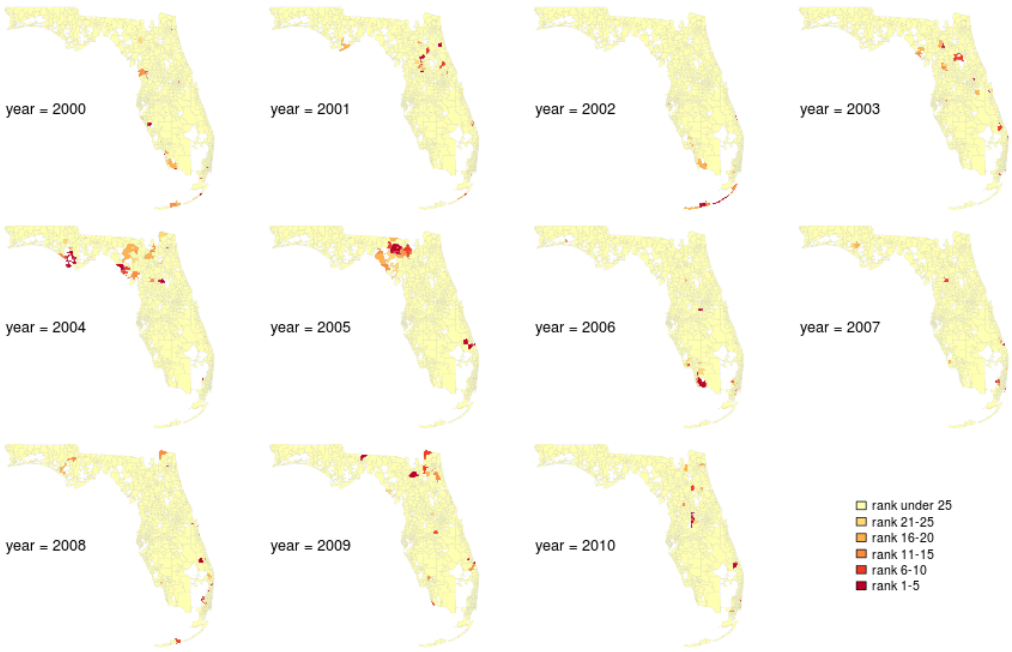

Figure 3.3 Map of the ICAR model by coloring highest 25 ranks of θ_i obtained from estimates minimizing WRSEL for 2000-2010 in Florida

Figure 3.3 shows a map of top 25 ranked regions for estimates of relative risks under WRSEL from the Poisson ICAR model for 2000-2010 in Florida. Recall that we fit the model for each year separately. Coastal areas of South Florida forms hot-spot in 2002. North Florida areas tend to be grouped for the years 2001, 2004 and 2009. Northwest Florida areas form a hot-spot group in 2004. North central regions make a large cluster and southeast Florida has a small cluster in 2005. Hot-spots are mainly occurred in the south Florida for the year 2006. They are also sparsely formed for 2007-2010 including 2003. It seems that hot-spots are more likely to be located shorelines or a little more inside. Hot-spot areas for each year support results of the spatial clustering configuration appeared in Zhang *et al.* (2014). Zhang *et al.* (2014) identified potential hot-spots by mapping clustering configuration in each year. In 2002, a cluster having the second highest risk and hot-spots in Figure 3.3 are mapped in the same regions (coastal areas of south Florida) as in Zhang *et al.* (2014). In 2003, their high risk clusters are appeared in north central Florida and our result well reflects their finding.

Especially for the year 2004, both maps pick the highest risk region in northwest Florida near the ocean. In 2006, hot-spot areas in Figure 3.3 are concentrated on south Florida and it appears as the second highest risk cluster in their clustering analysis.

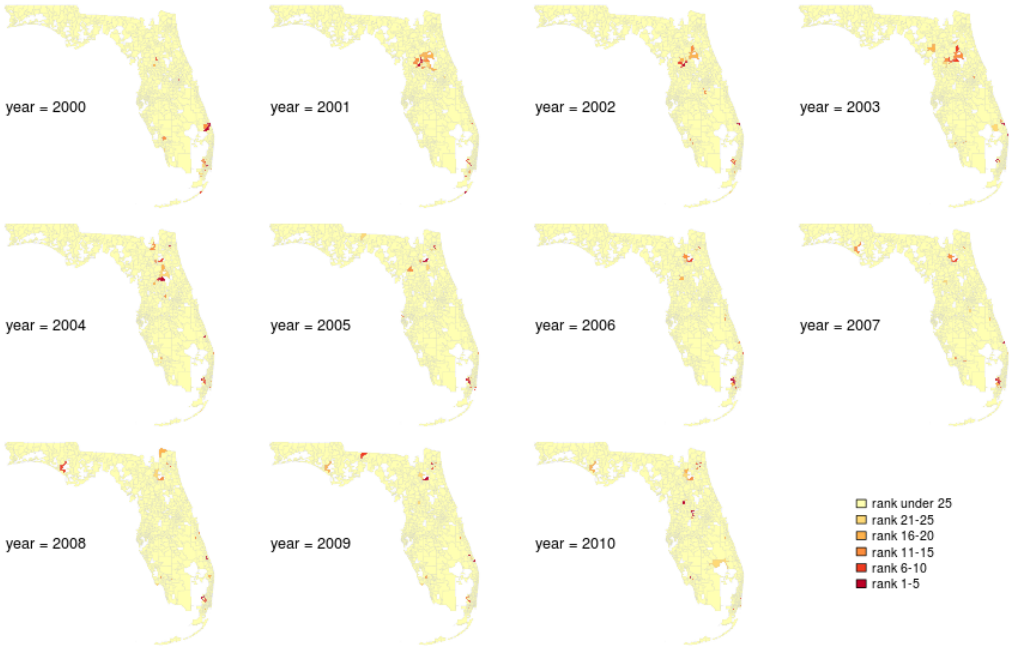

Figure 3.4 Map of the spatio-temporal CAR with AR(1) model by coloring highest 25 ranks of θ_i obtained from estimates minimizing WRSEL for 2000-2010 in Florida

Figure 3.4 shows a map of top 25 ranked regions for estimates of relative risks under WRSEL from the spatio-temporal Poisson CAR with AR(1) model for 2000-2010 in Florida. It is easily seen from the Figures 3.3 and 3.4 that how the temporal parameter makes the difference compared to the Poisson ICAR model. A cluster in the southeast areas for the year 2000 vanishes as year goes on but is more frequently turned up for the Poisson ICAR model (Figure 3.3). Regions in further south from that cluster, which are located near Miami forms a cluster all years. Wang and Rodriguez (2014) also found that near Miami areas form a highly concentrated cluster when they map with covariate-adjusted relative risks. North Florida areas are clustered largely in 2001-2003. It becomes smaller and moves gradually over years after 2004. This phenomenon was also captured by Wang and Rodriguez (2014) by having high but not the highest in large areas of north Florida. Amin *et al.* (2014) also found that their approach for each cancer type among brain cancer, leukemia and lymphoma makes a cluster in south Florida and includes Miami in common. Results of all pediatric cancer types by Amin *et al.* (2014) give the main significant regions as a cluster in southwest Florida which is slightly different from our outcomes but the cluster size is big enough to include Miami. The significant secondary regions are located in north central Florida which are mapped with orange colors for the year 2001-2004 in Figure 3.4.

4. Discussion

We investigated several spatial and spatio-temporal models typically used in disease mapping intensively for pediatric tumor incidence data in Florida during 2000-2010. They are based on a generalized linear model with Poisson regression under a Bayesian hierarchical setting. An alternative inferential approach for such models is h-likelihood (e.g. Lee and Nelder, 1996, 2001; An *et al.*, 2015).

The ICAR model tends to provide smaller DIC than proper CAR models when only spatial dependence structure was considered. The zero-inflated Poisson model was not strongly supported over the ICAR model in terms of the DIC value which suggests the zero-inflated structure is weak for pediatric tumor incidence data in Florida. If cancer counts are further divided by cancer types, zero-inflation would be more prominent. Then, one can consider a multivariate zero-inflated model such as Kim (2013) who considered a zero inflated bivariate negative binomial regression model. Estimated variances sharply changes over years when the model is fitted yearly, which suggests to look at an extended model that allows varying variance by years.

Since the data have a temporal component, we also considered spatio-temporal models with a CAR structure for the spatial component and a AR(1) structure for the time component. An estimated temporal autocorrelation parameter suggests that we need to incorporate temporal dependence in the model. Estimates of relative risks by WRSEL could capture hot-spot areas which is similar to the results obtained by the other researchers (Amin *et al.*, 2014; Wang and Rodriguez, 2014; Zhang *et al.*, 2014). While hot-spot regions by the Poisson ICAR model under WRSEL substantially vary over the years, those by spatio-temporal model do not. Also, the model considering temporal dependence provides more similar results to the other researchers who investigated the same dataset. We like to point out that the approaches we have taken to analyze the pediatric tumor incidence data in Florida are relatively simple compared to the other researchers who analyze the same data but we produced similar findings.

Acknowledgment

The authors thank Dr. Raid Amin and the Florida Association of Pediatric Tumor Programs (FAPTP) for providing them with the data. Also, the authors thank Dr. Amin for his valuable comments while preparing this article.

References

An, D., Han, J., Yoon, T., Kim, C. and Noh, M. (2015). Small area estimations for disease mapping by using spatial model, *Journal of the Korean Data & Information Science Society*, **26**, 101-109.

Amin, R. W., Hendryx, M., Shull, M. and Bohnert, A. (2014). A cluster analysis of pediatric cancer incidence rates in Florida: 2000-2010, *Statistics and Public Policy*, **1**, 69-77.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2013). *Hierarchical modeling and analysis for spatial data*, 2nd Ed., CRC Press, Boca Raton.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, **36**, 192-236.

Besag, J., York, J. C. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

Diggle, P. J. (2003). *Statistical analysis of spatial point patterns*, 2nd Ed., Arnold, London.

Heaton, M. J. (2014). Wombling analysis of childhood tumor rates in Florida. *Statistics and Public Policy*, **1**, 60-67.

Kim, D. (2013). A simple zero in ated bivariate negative binomial regression model with different dispersion parameters. *Journal of the Korean Data & Information Science Society*, **24**, 895-900.

Lawson, A. B. and Rotejanaprasert, C. (2014). Childhood brain cancer in Florida: a Bayesian clustering approach. *Statistics and Public Policy*, **1**, 99-107.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society B*, **58**, 619-678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalized linear models: A synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987-1006.

Leroux, B. G., Lei, X. and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, **116**, 179-191.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, **79**, 393-398.

Rushworth, A., Lee, D. and Mitchell R. (2014). A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology*, **10**, 29-38.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, **64**, 583-639.

Waller, L. A. (2015). Discussion: Statistical cluster detection, epidemiologic interpretation, and public health policy. *Statistics and Public Policy*, **2**, 4-11.

Wang, H. and Rodriguez, A. (2014). Identifying pediatric cancer clusters in Florida using loglinear models and generalized Lasso penalties. *Statistics and Public Policy*, **1**, 86-96.

Wright, D. L., Stern, H. S. and Cressie, N. (2003). Loss functions for estimation of extrema with an application to disease mapping. *The Canadian Journal of Statistics*, **31**, 251-266.

Zhang, Z., Lim, C. and Maiti, T. (2014). Analyzing 2000-2010 childhood age-adjusted cancer rates in Florida: A spatial clustering approach. *Statistics and Public Policy*, **1**, 120-128.