Analyzing 2000–2010 Childhood Age-Adjusted Cancer Rates in Florida: A Spatial Clustering Approach

Zhen ZHANG, Chae Young LIM, and Tapabrata MAITI

In this work, we apply a Bayesian hierarchical model that uses spatial clustering techniques to data from the Florida Association of Pediatric Tumor Programs (FAPTP) for the period 2000–2010. The goal is to determine whether there are statistically significant childhood cancer clusters at the Zip Code Tabulation Area (ZCTA) level of geography. The model provides estimates of the uncertainty associated with the clustering configurations, which is typically lacking in classical analyses of large datasets where a unique clustering representation can be insufficient. The model also allows covariate adjustment for known risk factors, bringing further relevant information, and it produces clusters that are spatially contiguous, enabling simple interpretation. The output clustering map is able to capture such patterns as the high-risk area that appear in the Southwest, Northeast, and Northwest Florida, which is consistent with the previous studies, but with finer details and deeper insight into year-specific features. New findings from the latest data, from 2008 to 2010, were also obtained and investigated. Our post-hoc validation of the clusters provides evidence for concluding that areas of elevated risk exist.

KEY WORDS: Bayesian hierarchical models; Epidemiology; Marginal likelihood; Pediatric cancers; Reversible jump MCMC

1. CHILDHOOD CANCER RATES IN FLORIDA

In the state of Florida, elevated rates of pediatric cancer in some regions have been previously reported (see Amin et al. 2010), which led us to investigate possible risk factors associated with such regions. Using data on pediatric cancer counts, together with additional demographic and geographic information available for Zip Code Tabulation Areas (ZCTAs) in the state of Florida from 2000 to 2010, we seek to identify regions that have higher risk of pediatric cancer in case further public health investigation should be performed to determine causal factors.

Before applying advanced statistical methods to identify ZCTAs with high risk, we first explore the dataset. To avoid possible confounding due to age, we calculate the age-adjusted rates using indirect standardization with the internal standard population (see, e.g., Curtin and Klein 1995). More specifically, let \( d_{ij} \) be the observed count of reported cases for site \( i \), time \( t \), and age group \( j \), where \( i = 1, \ldots, N \) and \( j = 1, \ldots, J \) represents the ZCTAs, time \( t = 1, \ldots, T \) and \( T = 11 \) corresponds to the year (2000 to 2010), and \( J = 4 \) represents four age groups: [0, 4], [5, 9], [10, 14], and [15, 19]. We let \( p_{ij} \) be the corresponding population size, and calculate the crude rate for the standard population \( M_t = \sum_{i,j} d_{ij} / \sum_{i,j} p_{ij} \) for each year \( t \). The incidence rates for each age group were pooled across the entire state to serve as the standard \( m_{ij} \) (see, e.g., Kulldorff et al. 1997). The indirect age-adjusted rate \( R_{it} \) per 100,000 population at site \( i \) and time \( t \) is calculated as

\[
R_{it} = \frac{M_t \sum_{j} d_{ij}}{\sum_{j} p_{ij} m_{ij}} \times 100,000
\]

and we investigate the log-transformed age-adjusted rate, \( Y_{it} = \log(R_{it}) \). We add a reasonably small positive value to the calculated rates \( R_{it} \) that are 0, to avoid the logarithm of zero. Note if the adjustment is too small, it can potentially introduce extreme clusters, which can deter our goal to identify hot spots. We empirically choose 0.5 by comparing the raw and transformed data distributions, and it well preserve the patterns, in particular for the large observations, without introducing a cluster of small values. Other choices such as 0.1 lead to similar patterns of data distribution, and conclusions of hot spots.

The age-adjusted cancer rates can also be calculated by choosing \( m_{ij} = \sum_{j} d_{ij} / \sum_{j} p_{ij} \) as the age-specific rates for indirect standardization, or using direct standardization as the exact formula provided by the Surveillance, Epidemiology, and End Results (SEER) program (seercancer.gov) of the U.S. National Cancer Institute (NCI) (2010). However, both calculations introduce a larger discrepancy between zero and nonzero rates that is more sensitive to the small-amount adjustment for the log-transformation, and hence carry an extra burden of modeling the zero inflation. Other normalization and variance stabilization techniques such as Freeman–Tukey transformation (see, e.g., Cressie and Chan 1989) can also be considered. In previous studies, such as Amin et al. (2010), it is also common to calculate the cancer rates adjusted not only for age, but also for gender and race information. Alternatively, we explicitly explore the effects of sex and race by considering both the factors as covariates in a regression model for the \( Y_{it} \). This allows us to further explore the potential regional heterogeneity in these effects, so that we can cluster the regions based on not only the mean level of \( Y_{it} \), but also the regional effects of the covariates. In this study, we consider the percentage of white residents \( X_{it1} \) and the percentage of male residents \( X_{it2} \) at each ZCTA \( i = 1, 2, \ldots, N \) and each year \( t = 1, 2, \ldots, T \) as covariates.

Figure 1 shows the log-transformed age-adjusted pediatric cancer rates, \( Y_{it} \), of the \( N = 983 \) ZCTAs in Florida over the \( T = 11 \) years in (a), and the map of the ZCTAs in (b), with 13...
ZCTAs highlighted that correspond to the upper 0.13% extreme observations of all $Y_{it}$’s. The threshold 0.13% is empirically chosen by investigating the histogram as our initial exploration of high-risk areas. We note that the data are unevenly distributed, particularly in Southern Florida. Densely clustered ZCTAs are seen on the east and west coast of South Florida, while the ZCTAs in the middle are less dense, with larger areas but possibly smaller population compared to the urbanized coasts.

From our preliminary exploration in Figure 1, it is plausible that the potential hot spots with high risk tend to appear in different regions and years. For instance, the three highest peaks in Figure 1(a) correspond to ZCTA 34101 from the Southwest Florida in 2000, ZCTA 34445 from the Northeast Florida in 2007, and ZCTA 32530 from the Northwest Florida in 2009, respectively. However, for the remaining years, these areas have relatively smaller log-rates. Such areas may appear as transient hot spots, and a clustering method with temporal smoothing can mask such patterns and result in estimates that are oversmoothed due to the regression-toward-the-mean phenomenon. On the other hand, the local similarity is remarkable. For instance, there is a group of ZCTAs that have extreme observations of the log-rates in Northeast Florida, particularly for 2005, and it is of interest to identify groups of ZCTAs with elevated risk and geographic proximity, rather than identifying a single ZCTA that has extreme values that may be due to chance. It is also reasonable to assume that adjacent ZCTAs share similar childhood cancer risk due to their similar environmental and socio-economic conditions. After the investigation, we conduct a year-by-year analysis to identify spatial clusters with similar levels of cancer risk, and shall suppress $Y_{it}$ into $Y_i$ in the following sections for model description. Furthermore, cancer rates are associated with demographic variables such as gender and race, so our analysis includes that information.

2. SPATIAL CLUSTERING MODEL

2.1 Literature Review

The Bayesian approaches for randomly partitioning the study region and searching nonoverlapping, spatially contiguous clusters, have been widely recognized and fully developed with applications in disease mapping and risk assessment. Popular choices in the literature are using Voronoi Tessellations (see, e.g., Knorr-Held and Raßer 2000; Kim, Mallick, and Holmes 2005; Feng et al. in press), or Bayesian Treed Partitions (see, e.g., Gramacy and Lee 2008; Konomi, Sang, and Mallick 2014). The latter choice commonly provides regularly shaped clusters such as rectangles or bands such as in the examples of Gramacy and Lee (2008) and Konomi, Sang, and Mallick (2014), unless ad hoc splitting rules are specified for the treed partition or the study domain is topologically transformed. The tessellation approach, however, can provide more generally shaped polygons, which better describe the disease maps, and it further allows the flexibility of choosing certain distance measures that are suitable for the study domain. We therefore consider the tessellation approach for searching regions with high risk of disease in the following context.

Our response variable is the log-rate $Y_i$ and our model includes an intercept and two covariates, percentage of male residents, and percentage of white residents, for the $i = 1, 2, \ldots, N$ ZCTAs in Florida. A spatial clustering configuration $\pi = (d, G_d)$ under tessellation approach is defined by two components: $d$ is the number of clusters, and $G_d = \{g_1, \ldots, g_d\}$ are the $d$ cluster centers, which is a subset of the $N$ sites that determine the cluster membership of all sites according to the minimal distance criterion. That is, the clustering label of the $i$th site is assigned to $k$ if the $k$th cluster center $g_k$ is the first center in $G_d$ that has the smallest distance from the site $i$.

It is evident that such spatial clustering configurations can strongly depend on the selection of distance between sites. We choose the ordinary Euclidean distance based on the spatial coordinates of the centroids of ZCTAs, which typically does not result in tied distances. Other geo-spatial distances can be also considered when the adjacency information of each ZCTA, represented as polygons for areal data, is available. Note that the number of clusters $d$ is bounded above when we specify a minimal cluster size to ensure stable estimates of coefficients for the regressors.

2.2 Model Description

To identify spatial clusters of the data, we consider a hierarchical model. First, we place a prior distribution on the clustering configuration, $\pi(\pi)$. Given the clustering configuration $\pi$ with $d$ clusters, $C_1, \ldots, C_d$, the covariate $X_i$ affects the response $Y_i$.

Figure 1. Visualization of the (a) calculated log-transformed age-adjusted cancer rates with upper 0.13% extreme observations, and (b) 983 ZCTAs in the map of Florida with a few highlighted areas that correspond to ZCTAs in the left panel.
differently for each cluster. More specifically, we have
\[
\sigma_i \sim \pi(\sigma),
\]
for \(i \in C_r\),
\[
\lambda_k \sim \text{igamma}(a_{\lambda_k}, b_{\lambda_k}),
\]
\[
\sigma^2_r \sim \text{igamma}(a_{\sigma^2_r}, b_{\sigma^2_r}),
\]
\[
\beta^k_r \sim N(0, \lambda_k \sigma^2_r),
\]
\[
Y_i \sim N(\lambda_r, \beta_r, \sigma^2_r),
\]
where we assume the sites within the same cluster share both the regression parameters \(\beta\), and the variation \(\sigma^2\), which are both varying across clusters. In the model specification above, \(\lambda_k\) is an overall signal-to-noise ratio, or detectability for the \(k\)th effect \(\beta^k_r\) for all \(r = 1, 2, \ldots, d\). We update this scaling parameter by assuming the conjugate inverse Gamma prior with shape \(a_{\lambda}\) and scale \(b_{\lambda}\), chosen such that the specified density is rather dispersed to allow the estimate of \(\lambda\) to be more data driven. One common choice is \(a_\lambda = 1\) and \(b_\lambda = 0.1\) to achieve this. We similarly choose \(a_{\sigma^2}\) and \(b_{\sigma^2}\) to yield a dispersed prior density for the cluster-specific variation \(\sigma^2\).

Let \(\Theta\) be the collection of all possible clustering configurations \(\sigma\). For the distribution of a spatial clustering configuration on \(\Theta\), we consider the prior model \(\pi(\sigma) = \pi(G_d \mid d) \pi(d \mid \kappa) \pi(\kappa)\), with \(\pi(d \mid \kappa) \propto (1 - \kappa)^{d-1}\) for \(d = 1, \ldots, N_0\) and \(\kappa \in (0, 1)\). \(N_0\) is the upper bound of \(d\), which is less than or equal to \(N\). Here, the hyperparameter \(\kappa\) controls the penalty of model complexity introduced by a large \(d\). The prior for \(d\) is a power decay law under this elicitation, since when \(\kappa\) is larger, \(d\) is more likely to be small, while when \(\kappa\) is larger, \(d\) is more likely to be small, while when \(\kappa\) is fixed at a tiny value, \(d\) is almost uniformly distributed over \(\{1, 2, \ldots, N_0\}\), indicating weak prior information. Inspired by the more traditional approach for model selection, we reparameterize \(\kappa\) as an exponential decay: \(\pi(d \mid \kappa) \propto \exp(-\tan(\pi\kappa/2)d)\) for \(\kappa \in (0, 1)\). For instance, when \(\tan(\pi\kappa/2) = 0\) is fixed to be \(0\) of parameters under \(\sigma\) \(\times a_\sigma^2/2\), it is closely related to the well-known Akaike information criterion (AIC) if \(a_\sigma = 2\), or Bayesian information criterion (BIC) if \(a_\sigma = \log(N)\). It is evident that the value of \(\kappa\) can affect the posterior aspect of the number of clusters \(d\). We therefore obtain the posterior estimate of \(\kappa\) from the data by eliciting another hierarchy \(\pi(\kappa) = \text{Uniform}(0, 1)\) that is flat over its support when no prior information is available, otherwise one can specify a general Beta density prior on \((0, 1)\) to incorporate any prior belief on \(d\). Note that since \(d\) is bounded above by \(N_0\), it requires \(\sum_{d=1}^{N_0} \pi(d \mid \kappa) = 1\), which yields
\[
\pi(d \mid \kappa) = \frac{f(k)^{d-1}(1 - f(k))}{1 - f(k)^{N_0}} (1)
\]
where \(f(\kappa) = \exp(-\tan(\pi\kappa/2))\). Next, conditional on \(d\), we also consider the uniform prior for centers \(\pi(G_d \mid d) = (N - d)!/N!\) where the \((N)!d!\) possible \(G_d\)s receive equal probability of being the cluster centers.

The full details of the model likelihood are described in Appendix A, and the implementation via the Markov chain Monte Carlo (MCMC) technique is described in Appendix B, where we embed the reversible jump MCMC (Green 1995) algorithm into Gibbs sampling scheme, for stochastically searching clustering configuration \(\sigma\) based on its integrated likelihood, with \(\beta^s\) and \(\sigma^2_s\)s proposed from their full conditional densities and hence marginalized out, that is, we restrict our attention on model search by spontaneously proposing the data-driven parameters. This is a different approach from Knorr-Held and Raßer (2000) and Feng et al. (in press) with extra efforts on tailoring the full likelihood to different proposal densities for the parameters during the model search.

The proposed clustering model is similar to the one proposed by Kim, Mallick, and Holmes (2005), which also uses Voronoi tessellations and piecewise Gaussian processes. However, there are several notable differences. First, our approach does not involve modeling the spatial dependence within each cluster under \(\sigma\), for two main reasons: (i) Certain ZCTA can have extremely high risk of disease, which largely exceeds that of the surrounding ZCTAs, and hence forms a singleton cluster. The spatial dependence within such singleton clusters is not quite meaningful and can cause unstable estimates. The spatial smoothing effects can be prone to clusters with more member ZCTAs rather than singleton clusters, which however may manifest the disease patterns and (ii) While Appendix A shows the marginal likelihood after integrating \(\beta^s\) and \(\sigma^2_s\)s out has a convenient form, the marginal model likelihood that further integrates the spatial parameters out is known to be intractable, and is numerically evaluated in Kim, Mallick, and Holmes (2005). Consequently, the computation cost aggregates over the MCMC iterations, which requires a large number of total runs for searching \(\sigma\)s in a massive candidate set particularly for large \(N\).

Second, we allow the adaptive learning of the overall effects of covariates by assuming random \(\lambda_{\kappa}\)s that are updated through MCMC iterations. We further allow the adaptive learning of the penalty on the number of clusters from the data by sampling \(\kappa\). Empirical results suggest the adaptive learning processes can potentially improve the mixing comparing to the choice by Kim, Mallick, and Holmes (2005), and the posterior behaviors of \(\lambda_{\kappa}\)s and \(\kappa\) are closely related to the number of clusters and can manifest the complexity of the data structure. With diffuse prior on \(\lambda_{\kappa}\)s and \(\kappa\), we seek for data-driven penalties for model complexity, which is relevant as the introduction of extra clusters and hence more \(\beta^s\)s can easily yield better model fit particularly when the within-cluster correlation is intendedly dropped.

Third, we explicitly update \(\beta^s\)s and \(\sigma^2_s\)s, which are not nuisance parameters in this study, with an extra Gibbs step to draw their posterior samples based on which our proposed post hoc procedure is executed for identifying hot spots.

Finally, as described in Appendix C, we employ a model averaging procedure proposed in Dass et al. (in press) to obtain a central clustering configuration \(\tilde{\sigma}_C\) from the full posterior samples of \(\sigma\), which produces clusters with more general shapes comparing to the commonly used posterior modal sample \(\tilde{\sigma}_M\). See Appendix C for more detailed discussions.

3. SPATIAL CLUSTERING OF CHILDHOOD CANCER RATES IN FLORIDA

We fit the spatial clustering model to the age-adjusted childhood cancer log-rates in Florida for each of the 11 years, with the regression parameters that change between different clusters. The summary of posterior distribution of the number of clusters \(d\) with the posterior modal estimates \(\hat{d}\) for each year
is shown in Table 1, while the summary of model parameters for measuring the uncertainty of clustering information and data complexity is given in Appendix C. It is seen that the estimated number of clusters and configurations are quite distinct across years, which further justifies this year-by-year study.

For each year, we extract the central clustering configuration \( \hat{\mathcal{C}}_d = \{ \hat{C}_1, \ldots, \hat{C}_d \} \), which integrates the clustering information over \( b = 1, 2, \ldots, B \) posterior samples. The detailed description of the procedure is given in Appendix C. For each cluster \( \hat{C}_r \) and posterior sample \( b \), we estimate the cluster mean age-adjusted log-rates \( \hat{\mu}_r^{(b)} \) by averaging the predicted values \( \hat{Y}_i^{(b)} = X_i \hat{\beta}(i)^{(b)} \) over member sites \( i \in \hat{C}_r \), where \( \hat{\beta}(i)^{(b)} \) denotes the regression parameter value assigned to site \( i \), as determined by its membership in \( \sigma(i)^{(b)} \) in the \( b \)th posterior sample. We then obtain the labels of each cluster \( \hat{C}_r \) by sorting the posterior mean \( \sum_{b=1}^{B} \hat{\mu}_r^{(b)}/B \) in descending order. Consequently, a cluster \( \hat{C}_r \) with a smaller label is considered to be more risky in terms of pediatric cancer. We calculate the posterior mean with 95% credible interval for the cluster means of age-adjusted rates, using the back-transformed posterior samples \( \exp(\hat{\mu}_r^{(b)})'s \), and plot them versus the cluster labels in Figure 2, with different scales for the y-axis for each year. The posterior mean can be viewed as the estimated number of cases per 100,000 children for each cluster under the way of calculating age-adjusted rates in Section 1. We also plot the map of the extracted central configuration \( \hat{\mathcal{C}}_d \) with cluster labels, which is shown in Figure 3, for all 11 years. We also show the several high-risk ZCTAs from Figure 1, the preliminary investigation, in the last panel of Figure 3 for comparisons. Note that the cluster labels and colors are consistent for both Figure 2 and 3, such that one can match the estimated cluster mean with the spatial locations.

To further identify potential hot spots out of the high-risk clusters in each year, we compare the estimated cluster mean of the cluster \( \hat{C}_r \) versus its surrounding clusters, as shown in Figure 3. Note that the choice of Euclidean distance between the coordinates of centroids for the spatial clustering model tend to yield convex, sphere-shaped clusters. Consequently, a concave cluster could be represented by several convex clusters arranged in a concave shape, and this would need joint investigation. Therefore, we define the probability of being a hot spot (poh) for a set of central clusters \( \mathcal{S} \), based on the its neighbor set \( \mathcal{R} \) that is comprised of all surrounding clusters with labels larger than that of the members in \( \mathcal{S} \), to be

\[
poh(\mathcal{S}) = \frac{1}{B} \sum_{b=1}^{B} \left( \max_{\hat{C}_r \in \mathcal{S}} \{\hat{\mu}_r^{(b)}\} > \max_{\hat{C}_r \in \mathcal{R}} \{\hat{\mu}_r^{(b)}\} \right).
\]

Clusters with a high value of poh(\( \mathcal{S} \)) can potentially be the hot spots for further investigation. Based on this criterion and the spatial information from Figure 3, we scan potential hot spots from the spatial clustering model for each year as follows:

1. In 2000, the high-risk ZCTA 34101 from Southwest Florida, as shown in Figure 1(a), forms a singleton cluster with the highest risk of the pediatric cancer and poh(\( \{1\} \)) = 1. Figure 2 indicates it has much higher log-rates than the remaining clusters. The other high-risk cluster is centered on ZCTA 32658 in North Florida, as shown in Figure 1(a), with a total of 12 member ZCTAs and poh(\( \{2\} \)) = 0.875. Moreover, we notice that cluster 3 is considered as a surrounding cluster for calculating poh(\( \{2\} \)) from Figure 3. By combining the two high-risk clusters, we have poh(\( \{2, 3\} \)) = 0.96, which is even larger. Hence, there is a strong evidence of a hot spot in year 2000 that consists of both cluster 2 and 3 in North Florida. The remaining sets of clusters have poh values dramatically dropped. For instance, cluster 4 also has higher risk, however poh(\( \{4\} \)) = 0.624 when compared to the surrounding clusters \( \{6, 9, 10\} \) with larger cluster labels.

2. In 2001, cluster 1 in Northwest Florida around ZCTA 32334 is detected as a highly potential hot spot with poh(\( \{1\} \)) = 0.988. Another potential hot spot in West Florida consists of both cluster 2, which is centering around the high-risk ZCTA 34201, and cluster 4 around ZCTA 33763, with poh(\( \{2, 4\} \)) = 0.996 for the joint set. Note that cluster 2 itself is less qualified as a hot spot, with a smaller poh(\( \{2\} \)) = 0.855, even though ZCTA

| \( \mathcal{S} \) | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| \{1\} | \text{0.246} | \text{0.349} | \text{0.297} | \text{0.167} | \text{0.323} | \text{0.521} | \text{0.004} | \text{0.219} | \text{0.595} | \text{0.001} | \text{0.014} |
| \{2\} | \text{0.118} | \text{0.096} | \text{0.048} | \text{0.194} | \text{0.113} | \text{0.161} | \text{0.171} | \text{0.011} | \text{0.007} | \text{0.001} | \text{0.001} |
| \{3\} | \text{0.006} | \text{0.048} | \text{0.048} | \text{0.194} | \text{0.113} | \text{0.161} | \text{0.171} | \text{0.001} | \text{0.001} | \text{0.001} | \text{0.001} |
| \{4\} | \text{0.002} | \text{0.001} | \text{0.001} | \text{0.001} | \text{0.001} | \text{0.001} | \text{0.001} | \text{0.001} | \text{0.001} | \text{0.001} | \text{0.001} |
| \{5\} | \text{0.171} | \text{0.037} | \text{0.037} | \text{0.122} | \text{0.063} | \text{0.053} | \text{0.014} | \text{0.014} | \text{0.014} | \text{0.014} | \text{0.014} |

NOTE: A bold number indicates the highest posterior probability.
34201 has higher risk. This example suggests that the observations of certain high-risk areas can be smoothed by those of surrounding areas with moderately high risk.

3. In 2002, cluster 1 consists of 45 ZCTAs in North Florida including the three high-risk ZCTAs 32639, 32139, and 34797 in Figure 1(a). It can be potentially identified as a hot spot with a high \( \text{poh}(11) = 0.940 \). The remaining sets of clusters have relatively smaller values of poh.

4. In 2003, the cancer risk is overall lower than in the remaining years. ZCTAs in cluster 1 that center around ZCTA 34442, as shown in Figure 3, have higher risk than the remaining areas, with the largest \( \text{poh}(11) = 0.907 \). Cluster 4, near the west coast of South Florida around ZCTA 34224, forms another potential hot spot with \( \text{poh}(4) = 0.93 \).

5. In 2004, the cluster with highest risk has \( \text{poh}(11) = 0.871 \) and it consists of 21 member ZCTAs. The high-risk ZCTA 34739 from the preliminary scan is contained in cluster 3, with \( \text{poh}(2,3) = 0.765 \).

6. In 2005, cluster 1 with 17 ZCTAs around the high-risk ZCTA 32190, and cluster 2 with 7 ZCTAs around ZCTA 32181, form a hot spot near north central Florida with \( \text{poh}(1,2) = 0.948 \). Cluster 3 along the west coast of South Florida, which includes the high-risk ZCTA 33921, has \( \text{poh}(3) = 0.899 \), while cluster 4 along the east coast of South Florida with 44 ZCTAs around ZCTA 33128, has \( \text{poh}(4) = 0.971 \).

7. In 2006, the regions in Southern Florida represented by combining clusters 2 and 3 have significantly higher risk than the Central Florida region represented by cluster 6, with \( \text{poh}(2,3) = 0.999 \). This is consistent with the findings by Amin et al. (2010). We also find the Northern Florida area represented by clusters 1 and 4 with relatively high value of \( \text{poh}(2,3) = 0.885 \).

8. In 2007, cluster 1, consisting of 4 ZCTAs around the high-risk ZCTA 34445 from the initial scan, can be potentially identified as a hot spot with a high \( \text{poh}(11) = 0.907 \). Cluster 4, near the west coast of South Florida around ZCTA 34224, forms another potential hot spot with \( \text{poh}(4) = 0.93 \).

9. In 2008, cluster 1 in North Central Florida, containing 10 ZCTAs around the high-risk ZCTA 32664 in Figure 1(a), has a large potential of being a hot spot with \( \text{poh}(11) = 0.92 \). Another hot spot, cluster 3 at the northeast corner around ZCTA 32207, has high \( \text{poh}(3) = 0.984 \). In Southern Florida, the east and west coast areas represented by cluster 4 and 7 have higher risk than the central regions represented by cluster 5 and 12, with \( \text{poh}(4) = 0.932 \) and \( \text{poh}(7) = 0.987 \), respectively.

10. In 2009, cluster 1 in Northwest Florida, consisting of three adjacent ZCTAs around the high-risk ZCTA 32530 in Figure 1(a), can be potentially identified as a hot spot with \( \text{poh}(11) = 0.949 \). Cluster 2 also has a relatively high value with \( \text{poh}(2) = 0.81 \). Another potential hot spot that is comprised of cluster 3 and 4 around the west coast of North Central Florida has \( \text{poh}(3,4) = 0.955 \).

11. In 2010, the three clusters with the highest risks can be all identified as hot spots in Northern Florida with relatively large \( \text{poh}(11) = 0.937 \), \( \text{poh}(2) = 0.951 \), and \( \text{poh}(3) = 0.93 \). The high-risk ZCAT 32061 from initial
Figure 3. Map of the central clustering configuration with clustering labels assigned by descending order of the posterior mean of the cluster mean estimates $\hat{\mu}_r$ for the 11 years.

Finally, by jointly investigating the distributions of hot spots and clustering label assignment according to risks from 2000 to 2010 in Figure 3, we conclude that, overall, the hot spots and high-risk clusters are more likely to be observed in North Florida than in South Florida. In particular, we have recognized hot spots and the high-risk clusters around North Florida and North Central Florida for the years 2000, 2002, 2003, 2005, 2006, 2007, 2008, and 2010. In South Florida, the coastal areas have generally higher risk than the inland regions, and the gap is particularly visible for year 2005 and 2008.

4. CONCLUSION

We have applied a Bayesian hierarchical model that uses a spatial clustering method to the age-adjusted childhood cancer rates in Florida from 2000 to 2010. The model is able to capture clustering and identify high-risk regions. The Bayesian estimates provide further insight into the uncertainty associated with the clustering configuration suggested by this dataset, and give statistical validation of high-risk clusters. Our main findings are consistent with the previous study in Amin et al. (2010), which had a study period from 2000 to 2007 and concluded that there were two major clusters: the Southern Florida cluster (2006–2007) and the North Central Florida cluster (2001–2004), based on scan statistics. Nevertheless, our cluster analysis and post hoc validation by different years produce more detailed maps. For instance, the centroids of the North Central Florida clusters switch from different regions in North Central Florida over time in Figure 3. Also, in year 2007, we found the west coast areas (cluster 4) form a hot spot and have significantly higher risk compared to the other regions in the Southern Florida (cluster 7). Our approach also captures the clustering pattern for the new data from 2008 to 2010, which are consistent with the exploratory data analysis. All extreme values have entered the high-risk clusters, which are recognized as highly potential hot spots. In our study, the regional covariates effects for race and gender turn out to be nonsignificant on the log-transformed age-adjusted rates.

APPENDIX A: FULL DESCRIPTION OF THE MODEL LIKELIHOOD

From the data perspective, we write the model likelihood at the cluster level by grouping the sites in each cluster, that is,
\( Y_r = (Y_{r1}, \ldots, Y_{rn_r})^T \) and \( X_r = (X_{r1}^T, \ldots, X_{rn_r}^T)^T \) for the \( r \)th cluster \( C_r = \{i_1, \ldots, i_{n_r}\} \) with \( n_r \) member sites, as follows:

\[
Y_r \sim N(X_r \beta_r, \sigma_r^2 I_{n_r}). \tag{A.1}
\]

When clustering is of primary interest, we note that compared to (A.1), we obtain the nuisance parameters \( \beta_r \) with the prior \( \beta_r^2 \sim N(0, \lambda_r \sigma_r^2) \), so the likelihood (A.1) becomes

\[
Y_r \sim N(0, \sigma_r^2(I_{n_r} + X_r \Lambda X_r^T)), \tag{A.2}
\]

where \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d) \), and a larger variance is introduced in (A.2) compared to (A.1). We then integrate out the nuisance parameters \( \sigma_r^2 \) from (A.2) with the prior \( \sigma_r^2 \sim \text{igamma}(a_r, b_r) \), so the likelihood becomes

\[
Y_r \sim T(0, \frac{b_r}{a_r}(I_{n_r} + X_r \Lambda X_r^T), 2a_r), \tag{A.3}
\]

where \( T(\mu, \Sigma, v) \) denotes a multivariate Student’s-T distribution with location vector \( \mu \), scale matrix \( \Sigma \), and degrees of freedom \( v \). It is immediately seen that the marginal likelihood for \( \sigma \) given the data and \( \lambda_r \)’s depends on the hyperparameters \( a_r \) and \( b_r \). Comparing (A.3) to (A.2), we note that \( \frac{b_r}{a_r} \) is an estimator of \( \sigma_r^2 \) that is between the prior mode \( \frac{b_r}{a_r+1} \) and prior mean \( \frac{b_r}{a_r} \) for the inverse Gamma density we specify, and it is well known that the multivariate Student’s-T distribution generally involves more uncertainty than the multivariate Normal density with the same parameters. Our empirical results suggest that the posterior behavior of the clustering configurations are robust across several reasonable choices of \( a_r \) and \( b_r \), in particular when the \( \lambda_r \)’s (and hence \( \Lambda \)) are updated throughout the MCMC iterations. Alternatively one can choose the noninformative Jeffery’s prior by setting \( a_r = b_r = 0 \), so the marginal likelihood (A.3) for \( Y_r \) is no longer multivariate Student’s-T distributed, but rather proportional to

\[
|I_p + \Lambda X_r X_r^T|^{-1/2} \times \Gamma(n_r/2)
\times (Y_r^T(I_{n_r} - X_r(X_r^T X_r + \Lambda^{-1})^{-1}X_r^T)Y_r/2)^{-n_r/2}. \tag{A.4}
\]

In this application, we stick to the common choice of hyperparameters \( a_r \) and \( b_r \) that define a rather dispersed prior density for \( \sigma_r^2 \), while we note the flexibility of eliciting different hyperparameters, such as consistent estimates from data with no clusters, or the noninformative priors.

From (A.3) it is immediately seen that for fixed \( \lambda = (\lambda_1, \ldots, \lambda_d)^T \) and \( \kappa \), the posterior propriety holds, that is, let \( Y = (Y_1, \ldots, Y_d) \), \( \beta = (\beta_1^T, \ldots, \beta_d^T)^T \), and \( \sigma = (\sigma_1, \ldots, \sigma_d)^T \), giving

\[
m(Y) = \sum_{\sigma \in \Theta} \int_{\beta} \int_{\sigma} \pi(\beta | \sigma) \pi(\sigma) \pi(\sigma) \times \pi(\sigma^2 | \sigma) \pi(\sigma \beta | \sigma, \sigma^2) d\beta d\sigma d\sigma^2 < \infty, \tag{A.5}
\]

based on the fact that \( \Theta \) is a finite set. It follows that the propriety for the full posterior distributions of \( \{\sigma, \beta, \sigma^2, \lambda, \kappa\} \) also holds since we specified the conjugate prior for \( \lambda \) and the uniform prior for \( \kappa \).

**APPENDIX B: IMPLEMENTATION OF THE SPATIAL CLUSTERING MODEL**

To fit the spatial clustering model, the set of parameters \( \{\sigma, \beta, \sigma^2, \lambda, \kappa\} \) is estimated under the Bayesian framework. Since our prior choices for the mean and scaling parameters are conjugate, we consider the Gibbs sampler by deriving the full conditional distributions. On the other hand, since the cardinality of the set \( \Theta \) for \( \sigma \) can be quite large due to large \( N \), and updating \( \sigma \) involves interchanging variable model dimensions, we shall adopt an embedded reversible jump MCMC algorithm of Green (1995) for updating \( \sigma \) at each Gibbs cycle that consists of following steps:

1. Update \( \sigma \): In this step, we consider updating the clustering configuration \( \sigma = (d, G_d) \), which determines the dimensionality of the parameters \( (\beta, \sigma) \) conditional on the remaining parameters \( \lambda \) and \( \kappa \). For notation simplicity, we suppress \( \lambda \) and \( \kappa \) in the conditioning part of all probability densities below. We construct the reversible jump MCMC step for updating the triplet \( (\sigma, \beta, \sigma) \) with potentially changing dimensions. From the current clustering configuration \( \sigma \) with associated parameters \( (\beta, \sigma) \), suppose we propose a new configuration \( \sigma^* \) with associated parameters \( (\beta^*, \sigma^*) \), we define an auxiliary \( U = \theta^* = (\beta^*, \sigma^*) \) and let \( \theta = (\beta, \sigma) = U^* \). The corresponding invertible map \( q : (\theta, U) \to (\theta^*, U^*) \) is one-to-one with the Jacobian to be unity. We first propose \( \sigma^* \) from a certain probability mass function \( H(\cdot) \) given the current clustering configuration \( \sigma \), then propose \( U \sim h(\cdot | \sigma, \theta, \sigma^*) \) for a certain proposal density function \( h(\cdot) \). Consequently, the Metropolis–Hasting ratio for accepting the proposed triplet is

\[
\frac{1}{H(\sigma | \sigma^*) \pi(\sigma^*)} \frac{\pi(\sigma^*, \theta | Y)}{\pi(\sigma, \theta | Y)} \frac{h(u^* | \sigma^*, \theta^* | Y)}{h(u^* | \sigma, \theta, \sigma^*)}. \tag{B.1}
\]

We choose the proposal density to be the posterior density, that is, 
\( h(u | \sigma, \theta, \sigma^*) = \pi(\sigma^2 | \sigma^*, \sigma) \times \pi(\beta^* | \sigma^*, \sigma^2, \beta) \), More specifically, for \( r = 1, 2, \ldots, d \), we first sample each \( \sigma_r^2 \) from

\[
\pi(\sigma_r^2 | \sigma^*, Y_r) \propto \pi(\sigma_r^2) \int_{\beta_r} \pi(Y_r | \beta_r, \sigma_r^2) \pi(\beta_r) d\beta_r. \tag{B.2}
\]

which is an inverse Gamma density with shape parameter \( a_r + n_r/2 \) and scale parameter \( b_r + Y_r^T Y_r/2 \) with \( V_r = I_{n_r} - X_r(X_r^T X_r + \Lambda^{-1})^{-1}X_r^T \). Next, we sample \( \beta_r \) based on \( \sigma_r^2 \):

\[
\beta_r^* | \sigma^*, \sigma_r^2, Y_r \sim N(\mu_r, \Sigma_r), \tag{B.3}
\]

Under this choice of proposal density function \( h(\cdot) \), we can substitute

\[
\pi(\sigma | \sigma^*) = \pi(Y | \sigma, \theta) \pi(\theta | \sigma, \sigma^*) (a^2 | \sigma) (\beta | \sigma) (\sigma | \sigma^*), \tag{B.4}
\]

with \( m(Y) < \infty \) from (A.5). Therefore, the acceptance ratio in (B.1) becomes

\[
H(\sigma | \sigma^*) \times \pi(\sigma^* | \sigma) \pi(\beta | \sigma, \sigma^*) (a^2 | \sigma) (\beta | \sigma) (\sigma | \sigma^*), \tag{B.4}
\]

Using the fact that

\[
\pi(\sigma | \sigma) = \frac{\pi(Y | \sigma, \beta, \sigma^2) \pi(\beta, \sigma^2 | \sigma)}{\pi(\sigma, \sigma^2 | \sigma)}, \tag{B.4}
\]

the ratio (B.4) reduces to

\[
H(\sigma | \sigma^*) \times \pi(Y | \sigma^*) \times \pi(\sigma^* | \sigma). \tag{B.4}
\]

When both the prior and proposal density of \( \sigma \) are diffuse, for instance, under tiny values of \( \kappa \) and uniform proposal density \( H(\cdot) \), the acceptance ratio in (B.5) is mainly determined by the marginal likelihood ratio \( r = \pi(Y | \sigma^*)/\pi(Y | \sigma) \) with

\[
\pi(Y | \sigma) = \prod_{r=1}^d \int_{\beta_r} \pi(Y_r | \sigma_r^2, \beta_r) \pi(\beta_r | \sigma_r^2) \pi(\sigma_r^2) d\beta_r d\sigma_r^2. \tag{B.6}
\]
which is a product of the multivariate Student’s-T densities under (A.3), or more specifically

$$\pi(Y \mid \sigma) = \prod_{i=1}^{d} \Gamma(a_\sigma + n_i/2) (1 + Y_i^2 V_i Y_i / (2b_\sigma))^{-(a_\sigma + n_i/2)} \prod_{i=1}^{d} [1 + \lambda X_i^T X_i |_a^2]^{-\frac{d}{2}},$$

(B.7)

where $V_i = I_{n_i} - X_i (X_i^T X_i + \Lambda^{-1})^{-1} X_i^T$.

Next, we consider specifying the probability mass function $H(\cdot)$ for proposing $\sigma^*$, which involves constructing plausible move types that give higher chance of exploring the set of clustering configurations that resemble the current $\sigma$ in $\Theta$, potentially with different dimensions, that is, varying numbers of clusters $d$. The efficiency of implementing the reversible jump MCMC algorithm can heavily rely on searching such neighborhoods of $\sigma$ in the full set $\Theta$.

To propose a new clustering configuration $\sigma^*$, we consider one of the three steps: a Growth step that creates one more clusters, a Merge step that deletes an existing cluster, and a Shift step with the same number of clusters but different cluster centers, with prescribed probabilities.

2. Update ($\beta, \sigma^2$): In this step, we consider sequentially updating $\beta$ and $\sigma^2$ given the remaining parameters including $\sigma$, to guarantee newly sampled values for ($\beta, \sigma^2$) even when the proposed $\sigma^*$ with ($\beta^*, \sigma^2*$) is not accepted in the preceding step, to potentially improve the mixing. Specifically, for each cluster $r = 1, \ldots, d$, we sample $\beta_r$ from (B.3), and then sample $\sigma^2_r$ from its full conditional density given $\beta_r$, also, which is an inverse Gamma density with shape parameter $a_\sigma + n_r/2$ and scale parameter $b_\sigma + \sum_{i=1}^{d} (\beta_i^2/\sigma^2_i)^2/2$. Finally, we sample $\kappa$ from $\pi(\kappa | Y, \sigma) \propto \pi(\beta_1 | \kappa) \pi(\kappa)$, which is proportional to the prior density (1) under the uniform prior choice for $\kappa$. One can use the inverse cumulative density function (CDF) method by numerically approximating the CDF of (1).

APPENDIX C: POSTERIOR ANALYSIS OF THE SPATIAL CLUSTERING MODEL

To get posterior samples, we run three MCMC chains with different initial numbers of clusters and hence different starting values for the remaining parameters. For each chain, we run the model for 100,000 iterations, which take around 30 min to complete on a 2.66 GHz Octocore Intel Xeon E7-8837 processor. The convergence is concluded after 50,000 iterations by monitoring the full and marginal likelihoods, and parameters at the site level. We then draw a posterior sample at every 100th iteration for each chain to reduce the autocorrelations of samples and repetitions of clustering configurations. We have a total of 1500 MCMC samples for inference. The posterior mean and 95% credible intervals for the scaling parameters $\lambda_r$ for $\beta^*_r$’s and penalty parameter $\kappa$ for number of clusters $d$ are shown in Table 2. The posterior estimates of $\lambda_r$ for the covariates percentage of male residents and percentage of white residents are relatively small comparing to that for the intercept. This indicates that the covariates effects are not far from zero with the adaptive scales. Comparing to the posterior probabilities of $d$ in Table 1, a larger value for the data-driven estimates of $\kappa$ indicates larger penalty, and generally yields smaller $d$.

We extract a central clustering configuration $\hat{\sigma}_C$ from the samples $\sigma^{(b)}$ for $b=1, 2, \ldots, B$ using the model-averaging procedure proposed in Dass et al. (in press). More specifically, we define the distance $D$ measure between a pair of ZCTAs $i$ and $j$ as

$$D(i, j) = 1 - \sum_{b=1}^{B} \int^{b}(i, j)/B,$$

where $\int^{b}(i, j)$ is an indicator function of the event that area $i$ and $j$ are in the same cluster under the $b$th clustering configuration $\sigma^{(b)}$. We then use a traditional hierarchical clustering procedure to obtain the central clustering configuration $\hat{\sigma}_C$ based on the distance measure $D$ and the posterior mode $\hat{d}$ of the number of clusters. Comparing to the posterior modal sample $\hat{\sigma}_M$ of $\sigma$, which is commonly used for summarizing clusters (see, e.g., Kim, Mallick, and Holmes 2005; Gramacy and Lee 2008), $\hat{\sigma}_C$ is not restricted to be a Voronoi tessellation, rather it involves clusters with more general shapes by integrating the clustering information over all posterior samples of $\sigma$. On the other hand, the posterior inference conditioning on $\hat{\sigma}_M$ can involve only partial information from the full posterior samples, while discarding considerable samples of $\sigma^{(b)}$’s that are similar to $\hat{\sigma}_M$ as its neighbor states in the model space $\Theta$. This can potentially lead to biased results in particular when the posterior probability mass on $\hat{\sigma}_M$ is low (for instance, around 0.1 for year 2000 in our analysis). Moreover, $\hat{\sigma}_M$ is restricted to a Voronoi tessellation with less flexibility and capability of summarizing the high-risk disease maps than $\hat{\sigma}_C$. One can further evaluate the dispersion of the posterior samples of $\sigma$ by comparing $\sigma^{(b)}$’s with $\hat{\sigma}_C$.
Table 2. Posterior mean (2.5%, 97.5% quantile) of the scaling parameter $\lambda_k$ for $k = 1$ (intercept), $k = 2$ (percentage of male population), $k = 3$ (percentage of white population), and $\kappa$

| Year | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\kappa$ |
|------|--------------|--------------|--------------|----------|
| 2000 | 785.9        | 0.100        | 0.008        | 0.127    |
| 2001 | 210.2        | 0.011        | 0.009        | 0.097    |
| 2002 | 228.0        | 0.009        | 0.008        | 0.102    |
| 2003 | 431.9        | 0.036        | 0.036        | 0.261    |
| 2004 | 230.9        | 0.039        | 0.033        | 0.130    |
| 2005 | 242.2        | 0.040        | 0.051        | 0.096    |
| 2006 | 157.4        | 0.036        | 0.030        | 0.189    |
| 2007 | 299.4        | 0.039        | 0.036        | 0.129    |
| 2008 | 234.2        | 0.035        | 0.035        | 0.096    |
| 2009 | 508.5        | 0.038        | 0.038        | 0.174    |
| 2010 | 284.2        | 0.040        | 0.040        | 0.112    |

using the agreement measures in Dass et al. (in press), and investigate the representativeness of $\hat{\nu}_C$ for $\nu^{(i)}$'s, which is high in this study. We further obtain the 95% credible intervals from cluster-wise estimated covariates effects $\hat{\beta}_i$ by averaging $\hat{\beta}_i(i)^{(i)}$ over all member sites $i \in C$, under $\hat{\nu}_C$ for posterior sample $b = 1, 2, \ldots, B$. The intervals for percentage of male residents and percentage of white residents span zero for all clusters, indicating the covariates effects are nonsignificant as manifested by posterior estimates of scales $\lambda_k$'s for $k = 2, 3$ which are relatively close to 0 comparing $\lambda_1$, posing less data evidence or high penalty for nonzero $\beta_i$'s.

[Received June 2014. Revised September 2014]

REFERENCES

Amin, R., Rohnert, A., Holmes, L., Rajasekaran, A., and Assanasen, C. (2010), "Epidemiologic Mapping of Florida Childhood Cancer Clusters," *Pediatric Blood Cancer*, 54, 511–518. [120,124,125]

Cressie, N., and Chan, N. H. (1989), "Spatial Modeling of Regional Variables," *Journal of the American Statistical Association*, 84, 393–401. [120]

Curtin, L. R., and Klein, R. J. (1995), "Direct Standardization (Age-Adjusted Death Rates)," *National Center for Health Statistics Health People Statistical Notes*, 6. [120]

Dass, S. C., Lim, C., Maiti, T., and Zhang, Z. (in press), "Clustering Curves Based on Change Point Analysis: A Nonparametric Bayesian Approach," *Statistica Sinica*. DOI: 10.5705/ss.2012.308. [122,127]

Feng, W., Lim, C., Maiti, T., and Zhang, Z. (in press), “Spatial Regression and Estimation of Disease Risks: A Clustering Based Approach,” Technical Report, Department of Statistics and Probability, Michigan State University. [121,122]

Gramacy, R. B., and Lee, H. K. H. (2008), “Bayesian Treed Gaussian Process Models With an Application to Computer Modeling,” *Journal of the American Statistical Association*, 103, 1119–1130. [121,127]

Green, P. J. (1995), “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” *Biometrika*, 82, 711–732. [122,126]

Kim, H., Mallick, B. K., and Holmes, C. C. (2005), “Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes,” *Journal of the American Statistical Association*, 100, 653–668. [121,122,127]

Knorr-Held, L., and Rafter, G. (2000), “Bayesian Detection of Clusters and Discontinuities in Disease Maps,” *Biometrics*, 56, 13–21. [121,122,127]

Konon, B. A., Sang, H., and Mallick, B. K. (2014), “Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets Using Covariance Approximations,” *Journal of Computational and Graphical Statistics*, 23, 802–829. [121]

Kalllavor, M., Feuer, E. J., Miller, B. A., and Freedman, L. S. (1997), “Breast Cancer Clusters in the Northeast United States: A Geographical Analysis,” *American Journal of Epidemiology*, 146, 161–170. [120]

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database, “Mortality—All COD, Aggregated With State, Total U.S. (1969–2007) <Katrina/Rita Population Adjustmement>,” National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released June 2010. Underlying mortality data provided by NCHS (www.cdc.gov/nchs). [120]