Bayesian Nonparametric Areal Wombling for Small Scale Map with an Application in Urinary Bladder Cancer Data from Connecticut

Rajarshi Guhaniyogi
Mailing Address: Department of Applied Mathematics & Statistics
University of California, Santa Cruz, CA 95064
SOE2
Email: rguhaniy@ucsc.edu

September 17, 2016

Abstract

With increasingly abundant spatial data in the form of case counts or rates combined over areal regions (e.g ZIP codes, census-tracts or counties), interest turns to formal identification of difference "boundaries", or barriers on the map, rather than the estimated statistical map itself. "Boundary" refers to a border that describes vastly disparate outcomes in the adjacent areal units, perhaps caused due to latent risk factors. This article focuses on developing a model based statistical tool equipped to identify difference boundaries in maps with a small number of areal units, also referred to as *small scale maps*, where detecting "boundaries" becomes relatively more challenging. This article proposes a novel and robust nonparametric boundary detection technique, coined as the Dirichlet Process Wombling (DPW) rule, by employing Dirichlet Process based mixture models for *small scale maps*. Unlike the recently proposed nonparametric boundary detection rules based on false discovery rates [1,2], DPW is free of ad-hoc parameters, computationally simple and readily implementable in freely available software for public health practitioners such as `JAGS` and `OpenBUGS` and yet provides statistically interpretable boundary detection in small scale wombling. We compare DPW with the popular and widely used boundary detection technique based on the conditional autoregressive (CAR) model and show significant advantage in terms of boundary detection through various simulation studies. Finally, we offer an application of our proposed approach to a urinary bladder cancer incidence rates dataset between 1990-2012 in the eight counties in Connecticut.

keywords Areal wombling; CAR model; Dirichlet process; Multiple hypotheses testing; Urinary bladder cancer.

1 Introduction

Public health practitioners often encounter spatial datasets where the concerned problem requires protecting individual privacy. Such datasets, referred to as *areal datasets* in spatial parlance, come in the form of case counts or rates combined over areal regions (e.g ZIP codes, census-tracts or counties). Of late, there has been a growing interest among public health practitioners in the spatial problem of identifying "boundaries" or barriers between two adjacent areal regions which reveal sharp changes in certain quantities of interest on either side. The general problem of boundary identification in spatial statistics, when addressed in areal data, is referred to as "areal wombling". Areal wombling is useful, especially in public health, for detecting rapid changes in disease mortality or incidence between two neighboring regions, and plays a key role in forming public health policies. In many applications concerning areal wombling, the geographical map of interest is a *small scale map* with only a few areal units (see Figure 1). As an example, consider the problem of identifying "health barriers" between the counties of Connecticut with respect to urinary bladder cancer. As a matter of fact, Connecticut is a small scale map with only eight counties. The issues with areal wombling in such small scale maps, though loosely stated in the literature, have not been given due attention. It has been observed that in a small scale map with a few adjacent counties, statistical models tend to overestimate/underestimate the difference between adjacent counties, thereby making it relatively more challenging to separate random variations from systematic variations.

Deterministic approaches for areal wombling [3,4] detect boundaries by implementing non-stochastic algorithms on areal data. Though there is a readily available `BoundarySEER` software to execute such algorithms, it fails to account for inherent variability and spatial association in the data. Matters become worse in *small scale* maps where extremities in counts and rates corresponding to certain thinly populated regions arise due to random variation in the observed data rather than due to systematic differences. In small scale maps, therefore, the differences between adjacent regions tend to be overemphasized by these algorithmic approaches, strongly indicating the need for a probability based modeling framework. Probabilistic modeling framework separates systematic variation from random variation in the data by borrowing information from neighboring units, enabling more accurate boundary detection. Such a model-based framework using a hierarchical Bayesian conditional autoregressive (CAR) model has been proposed by Lu and Carlin [5], where wombling probabilities along with the uncertainties are presented for every boundary in the map. Lu and Carlin [5] has a number of attractive features including its easy implementation in freely available software for public health practitioners. Unfortunately, calculation of posterior probabilities of wombling boundaries in [5] is dependent on a user defined cut-off c (see Sections 2,4,5) that is context specific and yields different posterior probabilities of wombling boundaries for changing cut-off values. The issue is more severe in small scale maps, where due to the smoothing effect of the CAR model, difference metrics for adjacent regions often show a continuous pattern. A small difference in the cut-off c might drastically change the probabilities of every boundary. We demonstrate such effects both in simulation studies and in real data application. Lu *et al.* [6] and Ma *et al.* [7] propose hierarchical models with priors on the

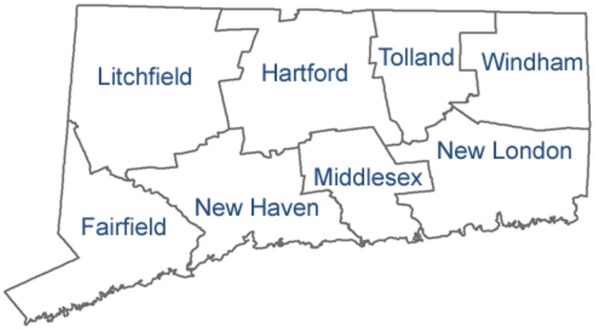


Figure 1: County map of Connecticut.

edges to estimate the adjacency matrix. However, inference from these models is usually highly sensitive to prior specifications on certain parameters.

Recently Li *et al.* [1] reformulate the areal wombling problem as one of Bayesian hypothesis testing within a class of spatial moving average models and adjust multiple tests using false discovery rates. Further, Li *et al.* [2] propose a nonparametric Bayesian hierarchical model that provides stochastic assessment regarding the presence of geographical barriers. This method alleviates "identifiability issues" in the aforementioned techniques and models spatial random effects as almost surely discrete realizations from the areally dependent stick breaking process (a derivative of Dirichlet process). Any subsequent detection of difference boundaries is based on whether spatial random effects corresponding to the two adjacent areal units are equal. This approach, therefore, establishes an important equivalence between the boundary detection problem and testing a set of hypotheses (where the number of hypotheses equals the number of edges) on equality between spatial random effects of adjacent regions. To circumvent the multiplicity issue arising from multiple hypotheses testing, Li *et al.* [2] propose using false discovery rate (FDR) based control. An advantage of [2] is that it permits the probabilistic estimation of an edge as a difference boundary. However, the model is computationally expensive, depends on the ad-hoc "target level" parameter introduced by FDR, and above all, cannot be fitted into any existing widely used software for public health practitioners (e.g. `JAGS`,`WinBUGS`).

The primary contribution of this article is to systematically study areal wombling for small scale maps. We propose a novel nonparametric Bayesian approach that delivers accurate boundary detection in small scale maps, and yet is free of "ad-hoc" target level parameters and is fast and easily implementable in freely available popular software for public health practitioners. We pursue an approach based on what we define as a "similarity zone." A *Similarity zone* corresponds to a connected set of areal units with no internal difference boundary. We propose a nonparametric model based hypothesis testing framework where each hypothesis corresponds to a specific partition of the areal map of interest into a number of "similarity zones". A class of Dirichlet process priors or its stick breaking approximation is then employed to calculate the posterior probabilities of each of these hypotheses simultaneously. Stochastic assessments on wombling boundaries are readily available from such posterior probabilities. The proposed method is conceptually simple and avoids FDR based multiple testing issues prominent in [2]. It is also easy to implement in freely available software for public health practitioners such as `JAGS`, `OpenBUGS` and provides state-of-the-art boundary detection for small scale maps. We clearly demonstrate the usefulness of our method over CAR based wombling in the simulation study and urinary bladder cancer dataset.

Our approach is fundamentally different from the massive literature on *small area estimation* [8,9] as our interest does not lie in finding accurate local estimates for regions with small sample sizes compared to the entire population, in the survey sampling data. Our reference to small scale map bears a different meaning than geography/cartography literature where a small scale map actually refers to a map covering a very large geographic region where the ratio of distance units on the map to distance on the ground is smaller than the same ratio for a zoomed in area. There is related literature in disease mapping where the focus lies on finding statistically significant clustering of areal units with respect to a disease [10]. In addition to providing the most probable clustering of areal units as a by-product, our approach yields the probability of wombling boundary between any two adjacent areal units. The Bayesian hierarchical model in Zhang *et al.* [10] does not provide any such assessment of difference boundaries.

Section 2 offers a brief exposition to boundary detection with the conditional autoregressive models for areally referenced rate incidence and case count data. Section 3 elucidates the proposed Bayesian nonparametric hypothesis testing framework for areal wombling. A detailed simulation study is conducted in Section 4 to first illustrate the new approach, which is then applied in Section 5 to detect boundaries on a county map of Connecticut that records the number of cases of urinary bladder cancer from the SEER-Medicare program. Finally, Section 6 concludes the article with an eye towards future work.

2 Areal Wombling with the Bayesian CAR Model: A Popular Approach

Broadly, an analysis of areal data proceeds through Bayesian hierarchical models that incorporate geographical effects. Assume y_{it} is the observed response in the i th geographical region at time t, $i = 1, ..., k$ and $t = 1, ..., T$. Two most common response types appearing in the public health sector are rate data and count data. When y_{it} represents the rate of incidence detected in the areal unit i at time t, the popular model for y_{it} following [11] becomes,

$$y_{it} = \mu_{it} + \epsilon_{it},\ \epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2), \tag{1}$$

$\mu_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \phi_i$, where $\boldsymbol{x}_{it}$'s are the region level covariates associated with the outcome, $\boldsymbol{\beta}$ represents the corresponding coefficients and σ^2 is the noise variability. When y_{it} is the number of events detected in the areal unit i, $i = 1, ..., k$ at time t, and E_i is the expected number of event occurrences in the areal unit i, a Poisson approximated binomial model (see

Waller & Gotway [12]) is employed as follows,

$$y_{it} \overset{ind}{\sim} Poisson(E_i e^{\mu_{it}}), \;\; i = 1, ..., k. \tag{2}$$

While y_{it} is assumed to be a random variable for (2), E_i's are fixed and known. ϕ_i represents the spatial random effect associated with region i. A popular practice in areal wombling is to model $\phi = (\phi_1, ..., \phi_k)'$ with the conditional autoregressive (CAR) model [5,13], denoted by CAR($\boldsymbol{W}, \tau$), where τ is a spatial dispersion parameter and $\boldsymbol{W} = ((w_{ij}))$ is the neighborhood matrix with $w_{ij} = 1$ if i is a neighbor of j (henceforth written as $i \sim j$) and $= 0$ otherwise.

CAR model has been immensely popularized in the last decade for its simple and efficient Bayesian implementation using popular statistical software such as `WinBUGS`. Usefulness of this model as a tool for areal wombling was pioneered by Lu and Carlin [5]. They suggested basing wombled boundaries on the posterior distributions of boundary likelihood values (BLVs),

$$\kappa_{ij} = |\eta_i - \eta_j|, \; i \sim j,$$

where

$$\eta_i = \begin{cases} \frac{E_i \exp\{\mu_i\}}{E_i} \text{ for (2)} \\ \phi_i \text{ for (1)} \end{cases}$$

Lu and Carlin [5] proposed the concept of *crisp* and *fuzzy* wombling boundaries from these boundary likelihood values. *Crisp* technique detects boundaries between units i and j if $E[\kappa_{ij}]$ exceeds some user defined cut-off. It has a serious disadvantage in that it does not provide any probabilistic assessment of difference boundaries. Fuzzy wombling technique assigns a number between 0 and 1 (not a probability) to judge which boundary segments are more likely to be considered as difference boundaries. However, fuzzy wombling technique, not being generated from a probability model, is unable to assign a degree of confidence to each segment. The most widely accepted strategy for areal wombling proposed in Lu and Carlin [5] appears as follows. Let c be a user defined cut-off; the probability that the boundary segment between units i and j is a difference boundary is assessed by $P(\kappa_{ij} > c|\boldsymbol{y})$. Within the Bayesian framework, Markov chain Monte Carlo(MCMC) samples are used to calculate the empirical estimate of $P(\kappa_{ij} > c|\boldsymbol{y})$, given by $\hat{P}(\kappa_{ij} > c|\boldsymbol{y}) = \frac{\#\kappa_{ij}^{(l)} > c}{L}$ [5,13], where $\kappa_{ij}^{(l)}$'s are MCMC samples and L is the total number of MCMC samples after burn-in. Since this is a binomial proportion, where its components are independent, basic binomial theory implies an approximate standard error for the estimate as $SE(\hat{P}_{ij}) = \sqrt{\frac{\hat{P}(\kappa_{ij}>c|\boldsymbol{y})(1-\hat{P}(\kappa_{ij}>c|\boldsymbol{y}))}{L}}$. In applications concerning areal wombling, one simply presents maps of $\hat{P}(\kappa_{ij} > c|\boldsymbol{y})$ and $SE(\hat{P}_{ij})$ for various choices of the cut-off c. Several variations of Lu and Carlin [5] are available in the literature, see e.g. Wheeler and Waller [14] for further references.

Given that the choice of c is not stochastically determined, it has to be chosen by ballpark estimation or some adhoc context specific rule. Clearly, depending on different choices of c, different probabilities of wombling boundaries appear. However, it is not clear which one of them should be chosen. The next section describes an alternative modeling strategy for spatial random effects ϕ that enables efficient and *automated* boundary detection without involving tuning parameters. Building upon (1) and (2) for rate and case count data respectively, the proposed approach models ϕ as a realization from a Dirichlet process. In due course, we will detail out how $P(\phi_i = \phi_j | \boldsymbol{y})$ is readily available from such specifications, leading to unambiguous boundary detection.

3 Nonparametric Dirichlet process based areal wombling

3.1 Areal wombling as a statistical hypothesis testing problem

Let $\boldsymbol{\Phi} = \{\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k) : \phi_i \in \mathcal{R}, i = 1, 2, \ldots, k\}$ be the k-dimensional parameter space. Equality and inequality relationships among the ϕ_i's induce statistical hypotheses which include $H_0 : \Phi_0 = \{\phi_i : \phi_1 = \phi_2 = \cdots = \phi_k\}, H_1 : \Phi_1 = \{\phi_i : \phi_1 \neq \phi_2, \phi_2 = \phi_3 = \cdots = \phi_k\}$,and so on up to $H_N : \Phi_N = \{\phi_i : \phi_1 \neq \phi_2 \neq \cdots \neq \phi_k\}$ as subsets. The total number of hypotheses N as a function of the number of areal units k is given by the famous Bell number B_k defined by the recursive relationship, $B_k = \sum_{i=0}^{k-1} \binom{k-1}{i} B_i$, $B_0 = B_1 = 1$. Let $\mathcal{H}_{hyp} = \{H_r : \Phi_r; r = 0, 1, \ldots, N\}$ be the set of all such hypotheses and $\boldsymbol{\Phi} = \bigcup_{i=1}^{N} \Phi_i$. Prior probabilities on these hypotheses can possibly be induced through discrete prior distributions on ϕ_i's. The posterior probabilities of different hypotheses indicate how likely it is for different ϕ's to be similar to or different from each other aposteriori.

Let the k geographical units be assigned numbers $\{1, 2, ..., k\}$. For the purpose of studying areal wombling, only a subset of $\mathcal{H}_{hyp}$ is of interest to us. We proceed to characterize them by introducing the concept *similarity zones* in an areal map.

Definition: A similarity zone is a connected set of geographical units with no internal difference boundary. Mathematically, $(i_1, ..., i_m)$, $i_1, ..., i_m \in \{1, ..., k\}, m \leq k$, is a similarity zone if the areal units $(i_1, ..., i_m)$ form a contiguous region in the map and $\phi_{i_1} = \cdots = \phi_{i_m}$.

Example: In the Connecticut map presented in Figure 1, denote “Litchfield” = 1, “Hartford” = 2, “Tolland” = 3, “Windham” = 4, “New London” = 5, “Middlesex” = 6, “New Haven” = 7 and “Fairfield” = 8. Then, $(1, 2, 8)$ corresponds to a similarity zone if $\phi_1 = \phi_2 = \phi_8$. However, $(1, 4, 5)$ is not a similarity zone even if $\phi_1 = \phi_4 = \phi_5$, as the areal units $(1, 4, 5)$ do not form a contiguous region.

Let

$$\mathcal{H}_{sim} = \{ \text{ All possible partitioning of k units into similarity zones} \} . \tag{3}$$

Clearly, a hypothesis that corresponds to partitioning geographical units into a number of *similarity zones* must belong to the class of all hypotheses $\mathcal{H}_{hyp}$ so that $\mathcal{H}_{sim} \subseteq \mathcal{H}_{hyp}$. Hereon, our focus resides upon assigning prior distributions on $\phi_1, ..., \phi_k$ that enable efficient estimation of $P(H_i|\boldsymbol{y})$ for every $H_i \in \mathcal{H}_{sim}$. Ideally, our goal is to assign prior distributions on $\phi_1, ..., \phi_k$ that ensure apriori (hence aposteriori) probability one to $\mathcal{H}_{sim}$. One could possibly extend the recent literature on Constrained Dirichlet processes (DP) [15] to facilitate a solution to this problem. However, it is less desirable for multiple reasons. Firstly, constrained

DP might involve complicated model formulation that increases computational complexity. Secondly, complex formulation using constrained DP might not enjoy easy implementation in a popular software (e.g. WinBUGS). In the next few sections an alternate strategy is proposed based on ordinary Dirichlet process based prior distribution on ϕ_i's.

3.2 Prior probabilities on boundaries using DP

In the context of (1) and (2), a Dirichlet process mixture prior is assigned on $\phi_1, ..., \phi_k$ with the baseline distribution G_0 and the concentration parameter M, formally written as

$$\phi_1, ..., \phi_k \overset{iid}{\sim} G, \;\; G \sim DP(M, G_0).$$

Dirichlet process prior induces nonzero prior probabilities on every hypothesis in $\mathcal{H}_{hyp}$. To elaborate further, let $\phi_1, \phi_2, ..., \phi_k$ be a sample of size k from a DPP. The sample belongs to the class $C(m_1, m_2, ..., m_s)$, written as $(\phi_1, \phi_2, \ldots, \phi_k) \in C(m_1, m_2, \ldots, m_s)$, if there are m_1 distinct values of ϕ that occur once, m_2 that occur exactly twice, . . . , m_s that occur s times. It follows that the number of means is $k = \sum_{i=1}^{s} i m_i$, and the total number of distinct values that occur is $p = \sum_{i=1}^{s} m_i$. The following equation from [16,17] gives the prior probability of a hypotheses in $\mathcal{H}_{hyp}$ in terms of its C-class.

$$P\{(\phi_1, \ldots, \phi_k) \in C(m_1, \ldots, m_s)\} = \frac{k!}{\prod_{i=1}^{s} i^{m_i} m_i!} \frac{M^{\sum_{i=1}^{s} m_i}}{M^{(s)}} \tag{4}$$

where $M^{(s)} = \prod_{i=1}^{s}(M + i - 1)$. (4) leads to a closed form expression of conditional prior probabilities for a hypothesis $H_i \in \mathcal{H}_{sim}$

$$P(H_i \,|\, \mathcal{H}_{sim}) = \frac{P(H_i)}{\sum_{H_\alpha \in \mathcal{H}_{sim}} P(H_\alpha)},$$

ensuring $\sum\limits_{H_i \in \mathcal{H}_{sim}} P(H_i | \mathcal{H}_{sim}) = 1$. Let $\mathcal{H}_{ij} \subseteq \mathcal{H}_{sim}$ be the set of all hypotheses with $\phi_i = \phi_j$. It follows that the prior probability of $\phi_i = \phi_j$ conditioned on $\mathcal{H}_{sim}$ is given by

$$P(\phi_i = \phi_j | i \sim j, \mathcal{H}_{sim}) = \sum_{H_r \in \mathcal{H}_{ij}} \hat{P}(H_r | \mathcal{H}_{sim}).$$

Since ϕ_i's are all exchangeable apriori, $P(\phi_i = \phi_j)$ is same for all $i \sim j$ apriori.

The concentration parameter M acts as a tuning parameter. It is well known that the choice of higher values of M favors hypotheses where more ϕ_i's are equal, while smaller values of M favor inequality between ϕ_i's. To facilitate a data driven choice of M, we treat M as a parameter for this analysis and assign a prior distribution $M \sim Gamma(2, 1)$. All the variance components are assigned $IG(1, 2)$ prior while each β_i is assigned a vague prior. We check sensitivity of the prior choice by changing hyperparameters and find robust inference. Posterior inference for the proposed model is based upon Markov chain Monte Carlo simulations on the DP models [18, 19].

3.3 Posterior probabilities of boundaries using DP

In each iteration, MCMC sample on $(\phi_1, ..., \phi_k)'$ corresponds to a hypothesis in $\mathcal{H}_{hyp}$. After L (here $L = 9000$) MCMC samples are drawn following burn-in, the posterior probabilities $P(H_i|\boldsymbol{y}),\ H_i \in \mathcal{H}_{hyp}$ are estimated using

$$\hat{P}(H_i|\boldsymbol{y}) = \frac{1}{L}\,(\text{No. of times } H_i \text{ occurs among L MCMC samples})\,.$$

Thereupon we calculate conditional posterior probabilities for a hypothesis $H_i \in \mathcal{H}_{sim}$

$$\hat{P}(H_i\,|\,\mathcal{H}_{sim}, \boldsymbol{y}) = \frac{\hat{P}(H_i\,|\,\boldsymbol{y})}{\sum_{H_\alpha \in \mathcal{H}_{sim}} \hat{P}(H_\alpha\,|\,\boldsymbol{y})}, \tag{5}$$

ensuring $\sum\limits_{H_i \in \mathcal{H}_{sim}} \hat{P}(H_i|\mathcal{H}_{sim}, \boldsymbol{y}) = 1$. It follows from (5), using the earlier definition of $\mathcal{H}_{ij}$,

$$\hat{P}(\phi_i = \phi_j|i \sim j, \mathcal{H}_{sim}, \boldsymbol{y}) = \sum_{H_r \in \mathcal{H}_{ij}} \hat{P}(H_r|\mathcal{H}_{sim}, \boldsymbol{y}). \tag{6}$$

Henceforth, for notational convenience, the above probability is written as $\hat{P}(\phi_i = \phi_j|i \sim j, \boldsymbol{y})$ after omitting $\mathcal{H}_{sim}$ in the conditioning set. $\hat{P}(\phi_i \neq \phi_j|i \sim j, \boldsymbol{y}) = 1 - \hat{P}(\phi_i = \phi_j|i \sim j, \boldsymbol{y})$ is taken as the posterior probability of existence of a wombling boundary between units i and j. To consider boundary detection analogous to crisp wombling, one can possibly choose a cut-off c_2 and detect (i, j)th boundary as a wombling boundary if $\hat{P}(\phi_i \neq \phi_j|i \sim j, \boldsymbol{y}) > c_2$. Alternatively, one can simply present an areal map with these probabilities, as is the popular practice with Lu and Carlin [5]. One important advantage of Dirichlet process wombling (DPW) lies in the fact that it completely eliminates the need for a user defined choice of c and automates areal wombling. The next two sections demonstrate the inferential advantages of DPW over LC.

4 Illustration with Synthetic Data

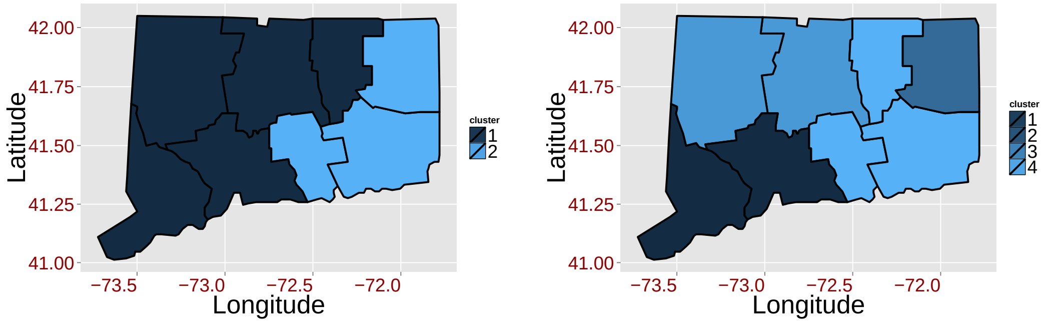
This section presents a detailed simulation analysis to assess performance of the proposed DPW as a boundary detector. For this section, the Connecticut county map is assumed to be the template. To facilitate implementation of DPW in `WinBUGS` and `JAGS`, we adopt truncated approximation of the stick breaking representation of DP. The stick breaking representation of the DP [17] says that a draw from the Dirichlet process can be written as $G(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$ a.s., where δ_{θ_i} is the Dirac measure (point mass) located at θ_i, each θ_i is a random draw from the base distribution G_0, and $p_1 = V_1$,..., $p_i = V_i \prod\limits_{l=1}^{i-1}(1 - V_l)$, where each $V_i \overset{iid}{\sim} Beta(1, M)$. θ_i's are known as atoms and p_i's are the stick breaking weights whose infinite sum is 1 a.s. In practice, the infinite sum is often replaced by a finite number (say S) of stick breaking weights $p_1, ..., p_S$ with $p_S = 1 - \sum_{i=1}^{S-1} p_i$. This is known as the truncated approximation of the Dirichlet process and is easily implementable in a popular software for public health practitioners. Since our focus is on the small area map, $S = k$ is chosen, which works well in every simulation study. Regarding the truncation bias, we must add that we have also implemented the full Dirichlet process mixture model in `R` and found identical boundary detection.

As described in Table 1, two different sets of simulation studies are conducted for this article, one for the rate incidence data and the other for the case count data, using the template of the Connecticut county map. As shown in Figure 1, Connecticut has 8 counties and 13 pairs of neighborhood counties. Following the simulation set ups in Li et al. [1,2], we divide the Connecticut map into a number of clusters (or similarity zones) and simulate data by using the same spatial random effect ϕ_i for each cluster (see Figure 2). In our simulations, $\boldsymbol{x}_i = (1, x_i)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ both in (1) and (2), where x_i's are drawn iid from N(0,1). β_0 is always kept at 1, while the true β_1 is fixed at 0.5 for both (1) and (2) respectively for all the simulations.

In the study of areal wombling, it is instructive to see the changes in performance when the difference between true cluster means varies. For the rate incidence data, we consider three cases by varying the differences between cluster means (a) within 2σ (case 1), (b) between 2σ and 3σ, and (c) above 3σ (case 3). For the case count data, cases 1 to 3 show gradual increase in the difference between cluster means. Table 1 shows different simulation cases. It is more difficult to detect difference boundaries with a smaller difference between cluster means than those with a higher difference.

To compare the performance of DPW, [5] (hereafter referred to as "LC" method) is implemented as a competitor. Both these methods are implemented in `OpenBUGS` and `JAGS` and they take less than 30 seconds to run 10000 MCMC iterations in a standard laptop. Detailed `OpenBUGS` code for implementing the proposed approach is presented in the Appendix. Convergence is diagnosed after 1000 iterations of burn-in using Gelman-Rubin diagnostics and autocorrelation plots. A subsequent 9000 MCMC samples are used for posterior inference. As suggested in Lu and Carlin [5], LC is implemented with three different choices of c. To elaborate further, $E(\kappa_{ij}|\boldsymbol{y})$ is estimated from 9000 MCMC samples for every $i \sim j$. c is chosen as the δth percentile of these estimated values over all $i \sim j$, where three different choices of c arise from setting $\delta = 5, 20, 30$.

Figures 3,4,5 and 6 show posterior probabilities of wombling boundaries for DPW and LC along with the estimated standard errors for LC in case 1. For our exposition, we only show figures for case 1 where boundary detection is hardest among the three cases, both in rate incidence and case count data. Since true difference boundaries are known in the simulation studies, one can compare accuracy of DPW rule as a boundary detector. It is evident from the figures that whenever there is a true difference boundary between two neighboring units i and j, estimated $P(\phi_i \neq \phi_j | i \sim j, \boldsymbol{y})$ is very close to 1 for DPW. On the other hand, all four figures reveal high probabilities for LC even on a few non-boundary elements. Additionally, Figures 4 and 6 show disparity in terms of posterior wombling boundaries for LC with three different cut-offs. For example, Figure 6 shows probability around 0.4 of having a boundary between Fairfield and New Haven for LC with $\delta = 5$, which sharply decreases to 0.09 for LC



(a) Configuration 1

(b) Configuration 2

Figure 2: A map of the simulated data in color-scales showing different clusters for a) configuration 1 having 2 clusters and b) configuration 2 having 4 clusters. Each cluster has its own mean and a cluster with the darker shade has a smaller mean. For cases a) and b), there are 5 and 9 boundary segments respectively that separate regions with different means(shades).

with $\delta = 30$.

To systematically study relative performances of all competitors, operating characteristics are thoroughly investigated. Recall that for a cut-off $c_2 \in [0, 1]$, DPW and LC detect boundary between units i and j if $P(\phi_i \neq \phi_j|\boldsymbol{y}) > c_2$ and $P(\kappa_{ij} > c|\boldsymbol{y}) > c_2$ respectively. Since the true boundaries are known for the simulation studies, it is easy to monitor True Positive (TP) and False Positive (FP) rates for DPW and LC (for$c$ corresponding to $\delta = 5, 20, 30$) with varying c_2. Let $TPR_{DPW}(c_2)$, $TPR_{LC,5}(c_2)$, $TPR_{LC,20}(c_2)$, $TPR_{LC,30}(c_2)$ be the true positive rates of DPW and LC with $\delta = 5, 20, 30$ respectively for $c_2 =$ `seq(0,1,0.02)`. Similarly, let $FPR_{DPW}(c_2)$, $FPR_{LC,5}(c_2)$, $FPR_{LC,20}(c_2)$, $FPR_{LC,30}(c_2)$ be the false positive rates of DPW and LC with three cutoffs. We take the average of true positive and false positive rates over $c_2 =$ `seq(0,1,0.02)` for all the competitors. Figures 7 and 8 plot the average true and false positive rates for all competitors for the rate incidence and case count data respectively. Few important observations are made from these figures. First, DPW shows uniformly smaller average false positive rates across all the simulation scenarios, while maintaining competitive true positive rates. Second, the false positive rates are much higher across all the competing methods for configuration 2 than for configuration 1. Perhaps the adverse effect of local smoothing is more in configuration 2. Third, the true positive rates are consistently higher in case 3 than in case 1 and 2 since case 3 allows relatively easy detection of boundaries. Understandably, false positive rates show an exact opposite behavior.

As discussed earlier, there is a significant discrepancy in terms of operating characteristics of LC with different c. While LC with c corresponding to $\delta = 30$ consistently outperforms LC with other δ in terms of the false positive rate, it shows lower true positive rates than the latter in many simulation studies. No clear choice of δ emerges from various simulation studies. The entire issue stems from the adhoc choice of c for the LC method. DPW solves this issue by eliminating the need to choose δ. More importantly, its easy implementation in popular software can be leveraged upon for potentially wide applicability among public health researchers.

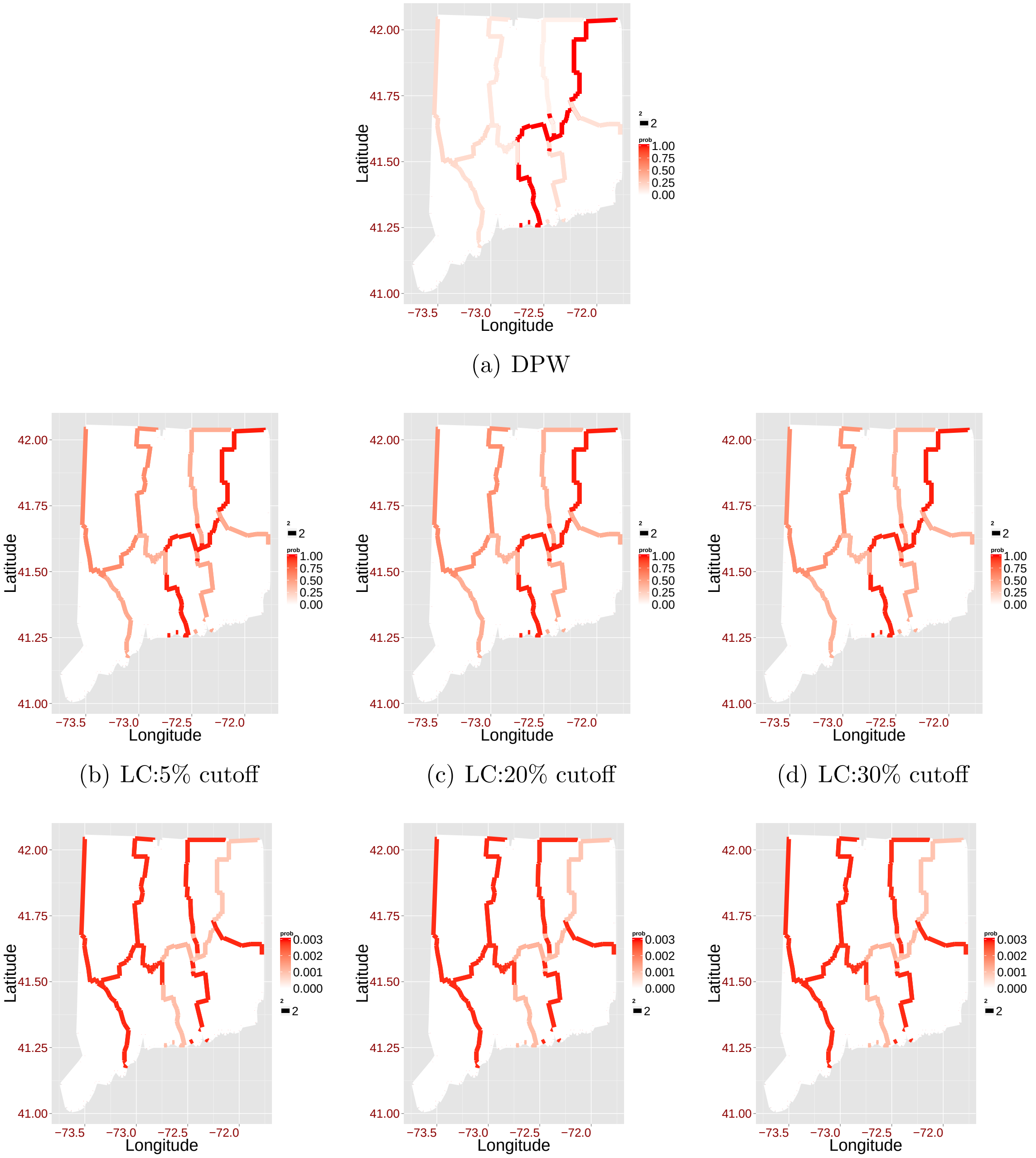
5 Analysis of Connecticut Urinary Bladder Cancer Dataset

This section demonstrates usefulness of our new areal wombling approach in the Connecticut urinary bladder cancer dataset. Bladder cancer is one of several types of cancers arising from the epithelial lining (i.e., the urothelium) of the urinary bladder. Urinary bladder cancer represents about 4.5% of the total cancer cases in 2015 and corresponds to the fifth largest number of cancer patients (after breast cancer, lung and bronchus cancer, prostrate cancer and colon cancer) among the entire U.S. population. Identifying "difference boundaries" for bladder cancer can fuel initiatives based on several public health programs to help identify the root cause of the "health barrier".

We analyzed bladder cancer data combining records of patients in all the age groups for eight counties in Connecticut. Our study consists of all the patients who have bladder cancer incidence between 1990 and 2012, measured every two years. Details of the data can be found in the Surveillance, Epidemiology and End Results (SEER) website `http://seer.cancer.gov/statfacts/html/urinb.html`, released April 2015, based on the November 2014 submission. Population data for all these years and counties, released in January 2015, can also be found from the SEER website. It is popular wisdom that bladder cancer becomes more common with age. Therefore, the analysis proceeds after adjusting for the average age of patients in a county (i.e. treating the average age of the patients for every county as a covariate in the regression) to investigate health barriers in bladder cancer between every two adjacent counties.

Let y_{it} and P_{it} be the total number of bladder cancer patients observed and the total population respectively in the year $t = 1990, 1992, ..., 2012$ for the counties $i = 1, ..., 8$. Then the expected number of cancer cases under the assumption of no spatial variation among counties is given by $E_i = \frac{\sum_{t=1}^{T} P_{it}}{T} \left(\frac{\sum_{i=1}^{k} \sum_{t=1}^{T} y_{it}}{\sum_{i=1}^{k} \sum_{t=1}^{T} P_{it}} \right)$, where $T = 12$ refers to the number of time points. Figure 9 shows the choropleth map of the raw data averaged over time, without adjusting for the age covariate. Based on the raw data from Figure 9, there seems to be a discrepancy in cancer incidence among a vast proportion of adjacent counties. It is to be seen as to how many of these differences are statistically significant after adjusting for the age covariate.

The urinary bladder cancer data does not offer an obvious choice of the cut-off c for LC. In absence of a context specific choice of c, we follow the procedure adopted in the simulation studies (and in Lu and Carlin [5]), i.e. choose three different c's as the δ-th



(a) DPW

(b) LC:5% cutoff

(c) LC:20% cutoff

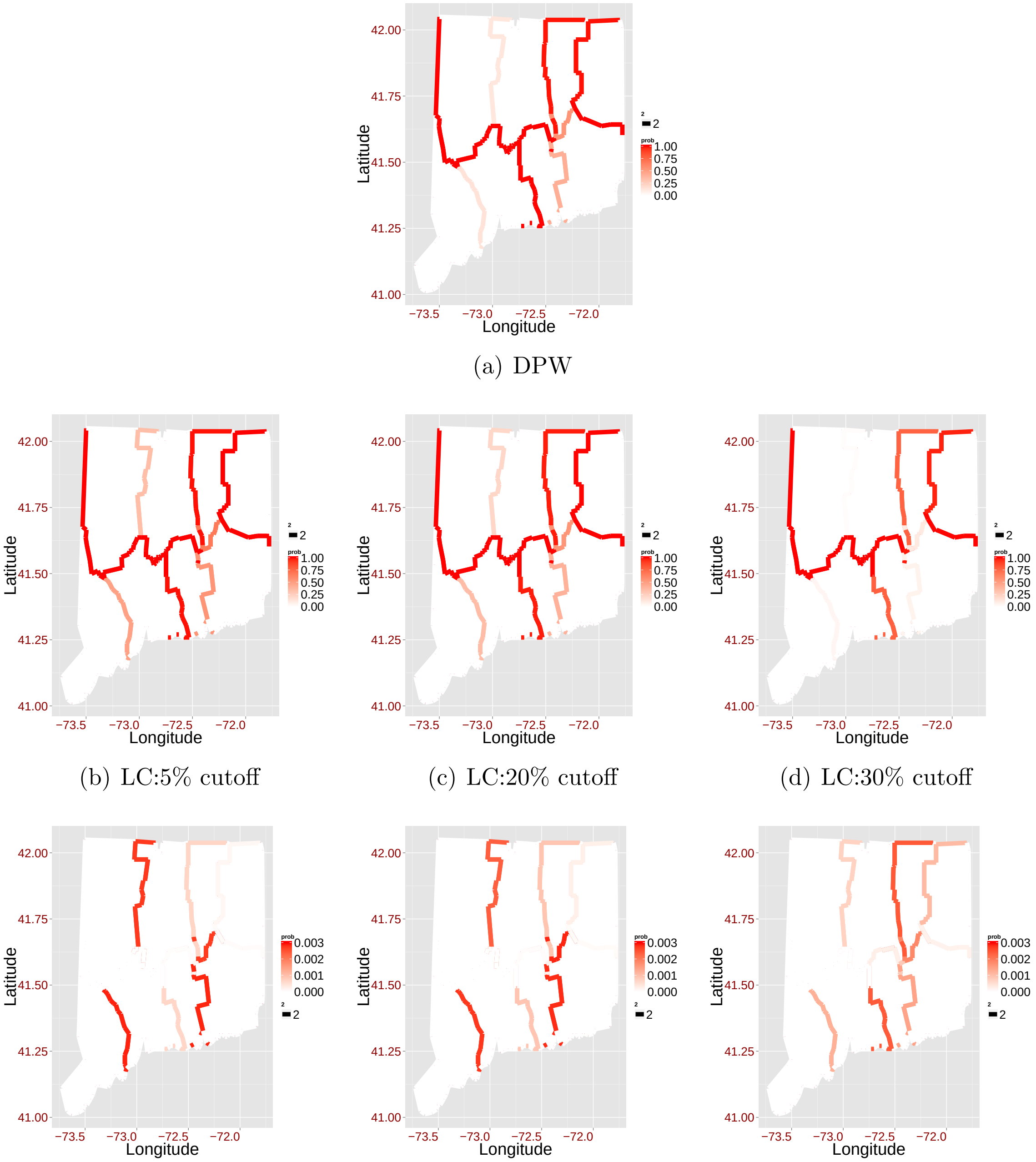
(d) LC:30% cutoff

(e) LC uncertainty:5% cutoff

(f) LC uncertainty:20% cutoff

(g) LC uncertainty:30% cutoff

Figure 3: Figures show the posterior probability areal wombling for different competitors under case 1 and configuration 1 with rate incidence data. First row shows boundary probabilities for DPW. Second row shows boundary probabilities for LC with c chosen as the (3(b)) 5%, (3(c)) 20%, (3(d)) 30% quantiles of estimated $E(\kappa_{ij})$ values respectively from left to right. Third row shows the corresponding standard errors of LC from left to right.



(a) DPW

(b) LC:5% cutoff

(c) LC:20% cutoff

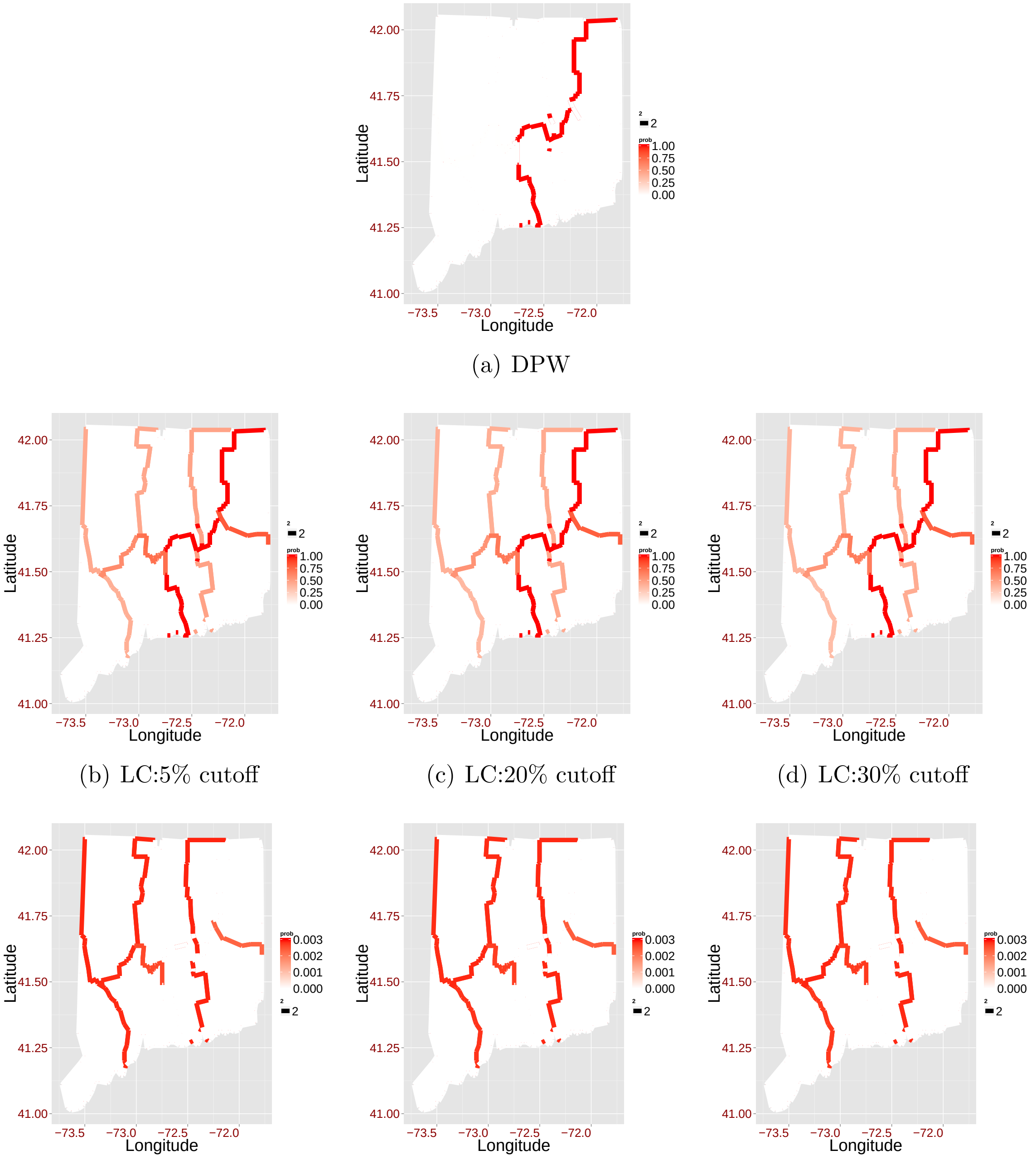
(d) LC:30% cutoff

(e) LC uncertainty:5% cutoff

(f) LC uncertainty:20% cutoff

(g) LC uncertainty:30% cutoff

Figure 4: Figures show the posterior probability areal wombling for different competitors under case 1 and configuration 2 with rate incidence data. First row shows boundary probabilities for DPW. Second row shows boundary probabilities for LC with c chosen as the (4(b)) 5%, (4(c)) 20%, (4(d)) 30% quantiles of estimated $E(\kappa_{ij})$ values respectively from left to right. Third row shows the corresponding standard errors of LC from left to right.



(a) DPW

(b) LC:5% cutoff

(c) LC:20% cutoff

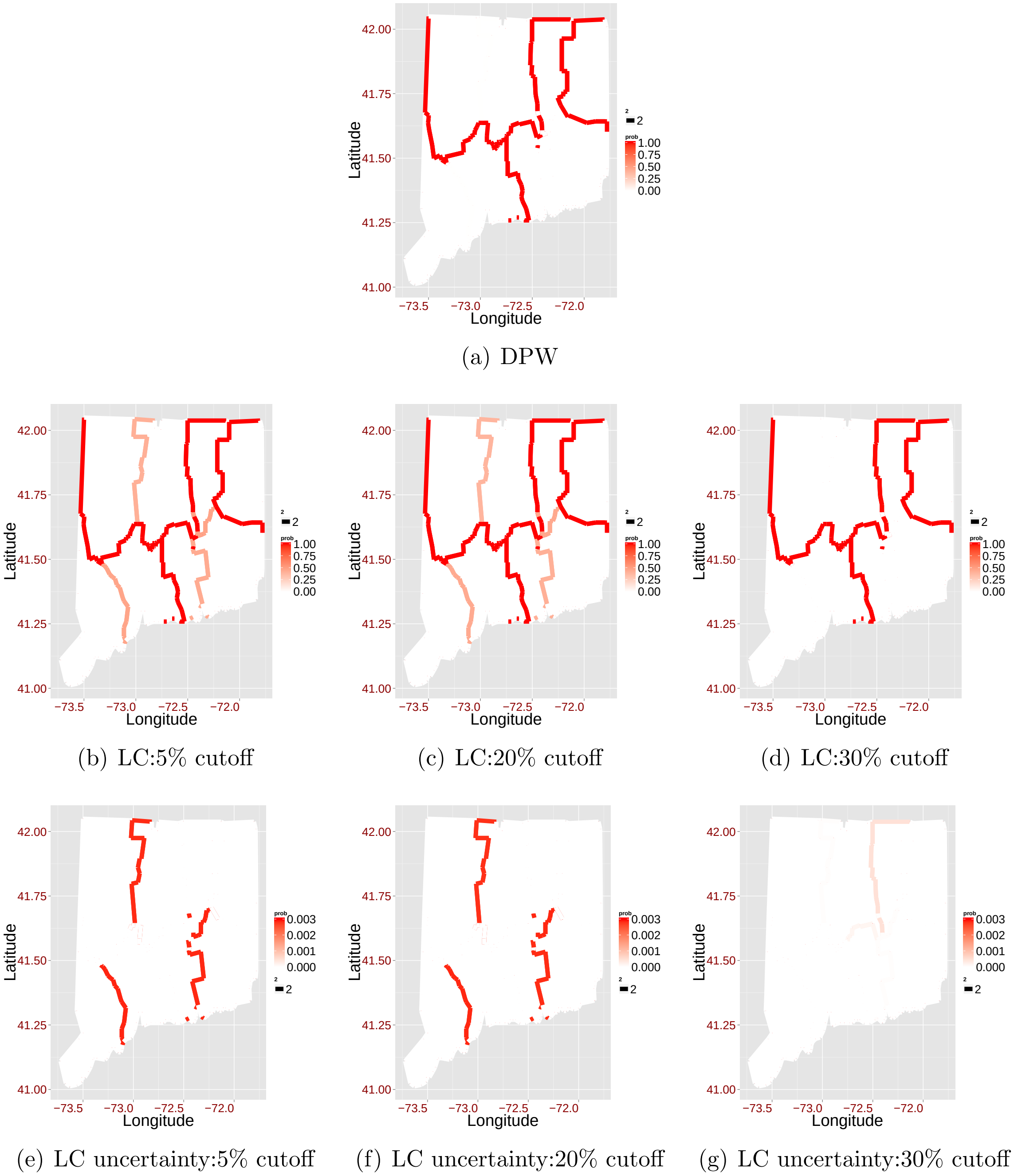
(d) LC:30% cutoff

(e) LC uncertainty:5% cutoff

(f) LC uncertainty:20% cutoff

(g) LC uncertainty:30% cutoff

Figure 5: Figures show the posterior probability areal wombling for different competitors under case 1 and configuration 1 with case count data. First row shows boundary probabilities for DPW. Second row shows boundary probabilities for LC with c chosen as the (3(b)) 5%, (3(c)) 20%, (3(d)) 30% quantiles of estimated $E(\kappa_{ij})$ values respectively from left to right. Third row shows the corresponding standard errors of LC from left to right.



(a) DPW

(b) LC:5% cutoff

(c) LC:20% cutoff

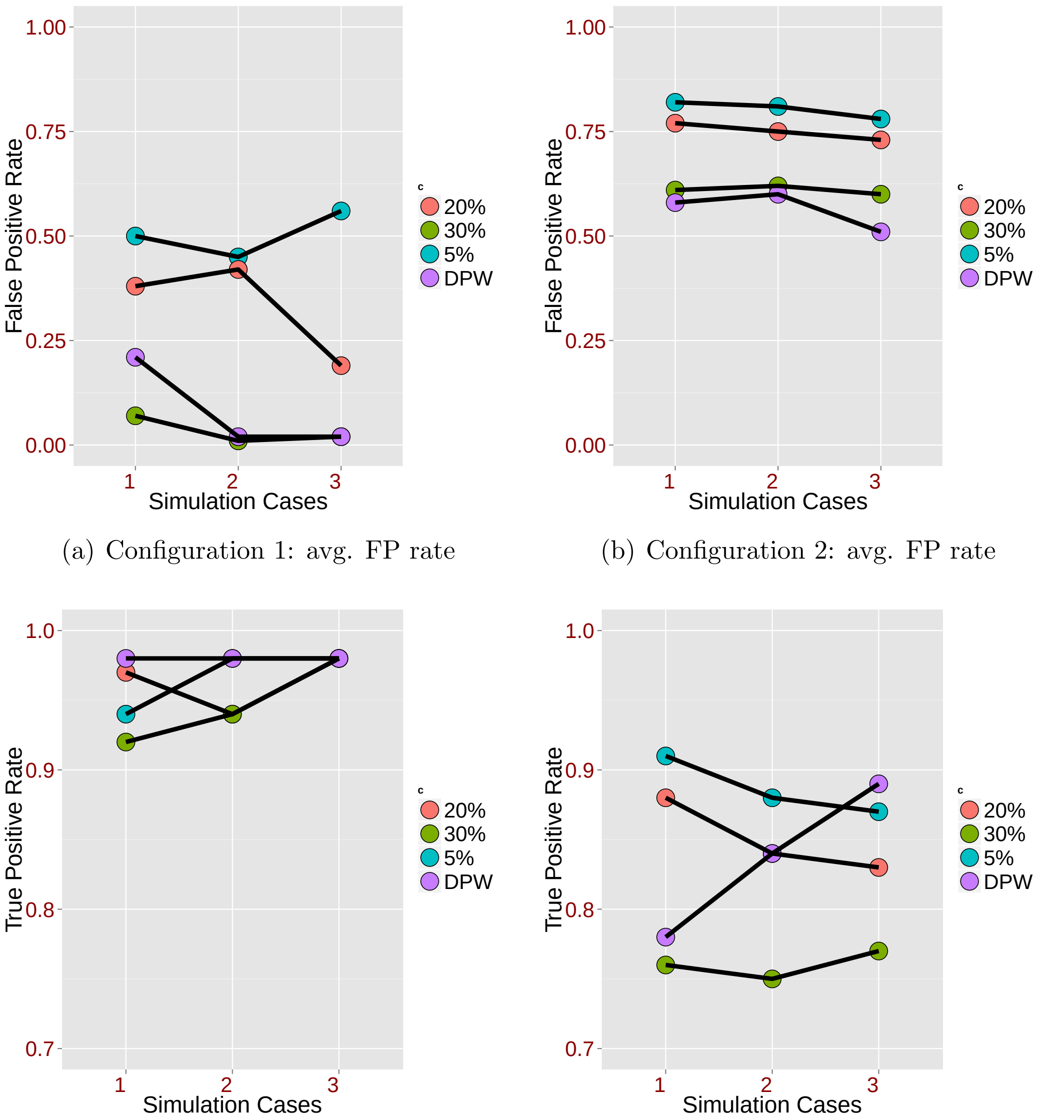
(d) LC:30% cutoff

(e) LC uncertainty:5% cutoff

(f) LC uncertainty:20% cutoff

(g) LC uncertainty:30% cutoff

Figure 6: Figures show the posterior probability areal wombling for different competitors under case 1 and configuration 2 with case count data. First row shows boundary probabilities for DPW. Second row shows boundary probabilities for LC with c chosen as the (3(b)) 5%, (3(c)) 20%, (3(d)) 30% quantiles of estimated $E(\kappa_{ij})$ values respectively from left to right. Third row shows the corresponding standard errors of LC from left to right.



(a) Configuration 1: avg. FP rate

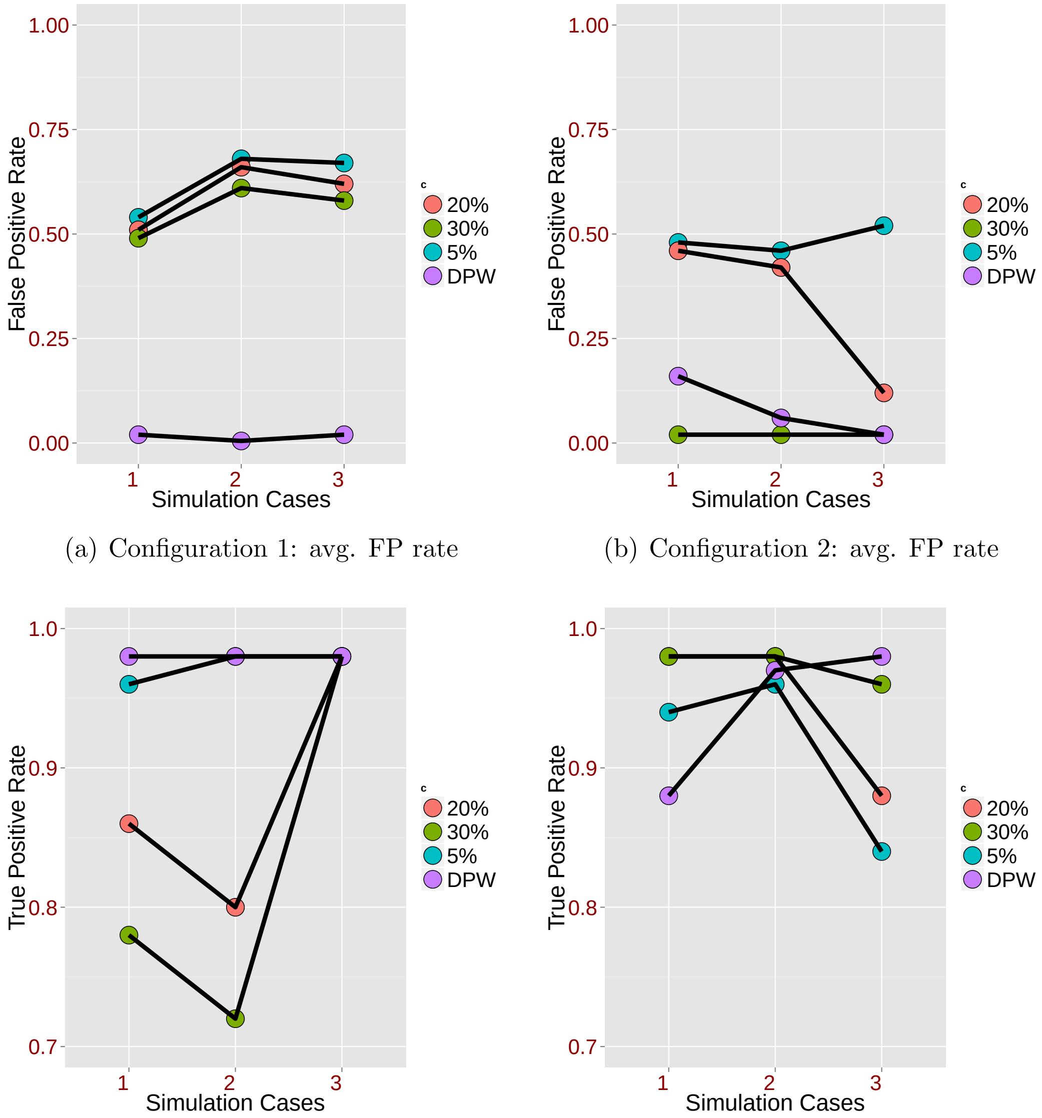
(b) Configuration 2: avg. FP rate

(c) Configuration 1: avg. TP rate

(d) Configuration 2: avg. TP rate

Figure 7: Figures 7(a) and 7(b) show average *False Positive* rates for DPW and LC with three values for the cutoff c. Figures 7(c) and 7(d) show average *True Positive* rates for DPW and LC with three values for the cutoff c. The average false positive and true positive rate is calculated by taking average over true positive and false positive rates respectively for 50 cutoff values of c_2 in a grid between 0 and 1. The first and second column display results for *rate incidence* data generated under configurations 1 and 2 respectively.

quantile of estimated $E(\kappa_{ij}|y)$'s, with $\delta = 5, 20, 30$. $\hat{P}(\phi_i \neq \phi_j | i \sim j, \boldsymbol{y})$ for DPW and $\hat{P}(\kappa_{ij} > c)$ for LC with three different c along with their estimated standard errors are shown in Figure 10. DPW shows strong evidence that out of 13 geographical boundaries, 10 are



(a) Configuration 1: avg. FP rate

(b) Configuration 2: avg. FP rate

(c) Configuration 1: avg. TP rate

(d) Configuration 2: avg. TP rate

Figure 8: Figures 8(a) and 8(b) show average *False Positive* rates for DPW and LC with three values for the cutoff c. Figures 8(c) and 8(d) show average *True Positive* rates for DPW and LC with three values for the cutoff c. The average false positive and true positive rate is calculated by taking average over true positive and false positive rates respectively for 50 cutoff values of c_2 in a grid between 0 and 1. The first and second column display results for *case count* data generated under configurations 1 and 2 respectively.

highly likely to be wombling boundaries. Only boundaries between (New Haven, Middlesex), (Middlesex, New London) and (New Haven, Litchfield) have very low posterior probabilities of being included as wombling boundaries. On the contrary, the posterior probabilities of

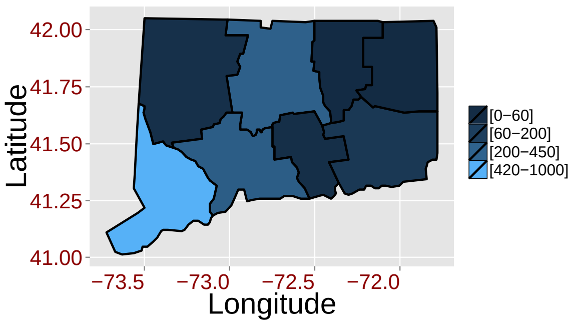


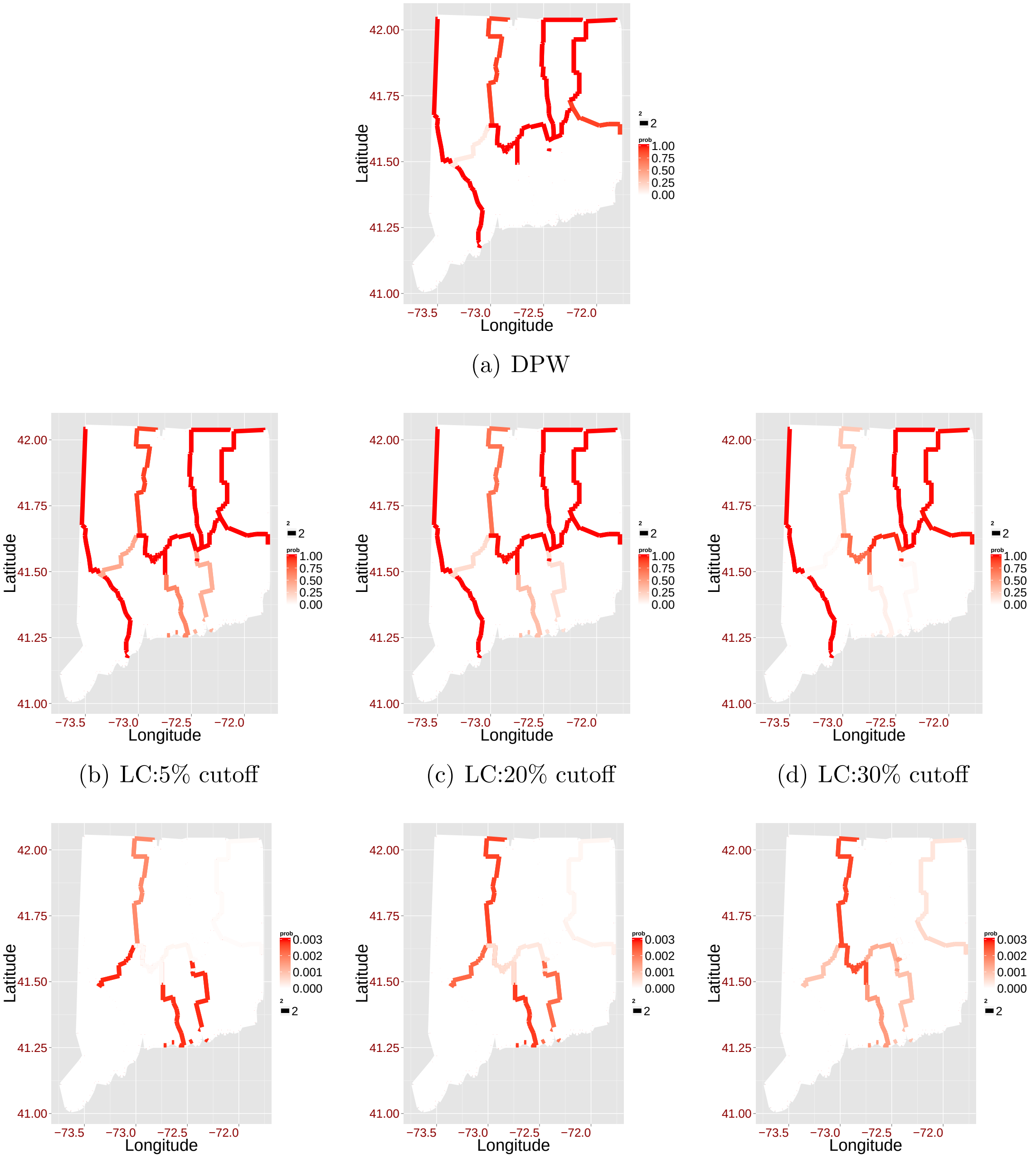
Figure 9: Choropleth map of the urinary bladder cancer data averaged over 12 years. Darker colors present counties with lesser cancer patients.

wombling boundaries vary widely for LC with three different cutoff values for c. While LC with $\delta = 30$ yields almost identical boundary detection as DPW, LC with $\delta = 5, 20$ assign posterior probability of around 0.6 to the boundary between (Middlesex, New London) and (Middlesex, New Haven). It is noteworthy that these two posterior probabilities also come with high uncertainties.

From simulations and data analysis, no choice of c appears to perform uniformly better than the others. The contribution of DPW lies in eliminating the need of c, thereby automating areal wombling. Most significantly, DPW offers a win-win situation where the automation does not lead to any loss in performance in terms of inference, computation time and easy implementation.

6 Discussion and Future Work

Due to the smoothing effect in a small number of adjacent counties, wombling in small scale maps is relatively more challenging. This article presents DPW, a novel nonparametric model based rule for detecting wombling boundaries for small scale maps. The proposed rule provides accurate stochastic assessment on wombling boundaries by directly calculating $P(\phi_i \neq \phi_j | i \sim j, \boldsymbol{y})$. Unlike the state-of-the-art nonparametric areal wombling approaches [1, 2], DPW rule is simple, easily implemented in `OpenBUGS` or `JAGS` and is not affected by the unavoidable issues of multiple hypothesis testing such as sensitivity to the choice of FDR based cut-offs. The popular areal wombling technique based on Lu and Carlin [5]



(a) DPW

(b) LC:5% cutoff

(c) LC:20% cutoff

(d) LC:30% cutoff

(e) LC uncertainty:5% cutoff

(f) LC uncertainty:20% cutoff

(g) LC uncertainty:30% cutoff

Figure 10: Figures show the posterior probability areal wombling for urinary cancer data. First row shows boundary probabilities for DPW. Second row shows probabilities for LC with c chosen as the (10(b)) 5%, (10(c)) 20%, (10(d)) 30% quantiles of estimated $E(\kappa_{ij})$ values respectively from left to right. Third row shows the corresponding standard errors of LC from left to right.

is dependent on a cut-off c, whose choice is context specific and is not apparent in many applications. We view DPW as a simple, powerful and practical tool for areal wombling in small scale maps that can be implemented with the same computational complexity as the CAR model. Various simulation studies and urinary cancer data from Connecticut show excellent performance for the proposed approach.

Future methodological investigations will focus upon two directions. As a more sophisticated extension to the DPW, we would investigate possible usage of constrained Dirichlet process based approaches that assign probability one to $\mathcal{H}_{sim}$ apriori. Dircihlet process based prior distributions with constraints have been a recent topic of research, see for example [15] for more detailed references. We aim to make use of this literature and propose a computationally flexible template for constrained DP measures geared towards understanding the areal wombling problem.

As discussed in section 5, it is evident that stick breaking representations of DP render themselves to multivariate extensions. In a future work, we seek to study DPW for multivariate areal wombling with matrix stick breaking processes in great detail. It is of special interest to investigate what advantages, if any, matrix stick breaking processes fetch over multivariate CAR models [20]. Does it also lead to unambiguous boundary detection in the multivariate correlated areal data? Is it also easily implementable in popular public health software? We would like to answer these questions in a future manuscript.

REFERENCE

1. Li, P., Banerjee, S., McBean, A.M., Carlin, B.P. Bayesian areal wombling using false discovery rates. *Statistics and Its Interface* 2012; **5**: 149-158.

2. Li, P., Banerjee, S., Hanson, T.A. and McBean, A.M. Nonparametric hierarchical modeling for detecting boundaries in areally referenced spatial datasets. *Statistica Sinica* 2015; **25**: 385-402.

3. Jacquez, G.M., Greiling, D.A. Local clustering in breast, lung and colorectal cancer in long island, New York. *International Journal of Health Geographics* 2003a; **2**.

4. Jacquez, G.M., Greiling, D.A. Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics in long island, New York. *International Journal of Health Geographics* 2003b; **2**.

5. Lu, H., Carlin, B.P. Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis* 2005; **37**: 265-285.

6. Lu, H., Reilly, C., Banerjee, S., Carlin, B.P. Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics* 2007; **14**: 433-452.

7. Ma, H., Carlin, B.P., Banerjee, S. Hierarchical and joint site-edge methods for Medicare hospice service region boundary analysis. *Biometrics* 2010; **66**: 355-364.

8. Rao, J.N.K., Molina., I. *Small area estimation.* John Wiley & Sons, 2015.

9. Lahiri, P. On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* 2003; **18**(2): 199-210.

10. Zhang, Z., Lim, C.Y., Maiti, T. Analyzing 20002010 childhood age-adjusted cancer rates in Florida: A spatial clustering approach. *Statistics and Public Policy* 2014; **1**(1): 120-128.

11. Banerjee, S., Wall, M., Carlin, B.P. Frailty modelling for spatially correlated survival data with application to infant mortality in Minnesota. *Biostatistics* 2003; **4**: 123-142.

12. Waller, L., Gotway, C.A. *Applied spatial statistics for public health data.* John Wiley & Sons, Hoboken, New Jersey, 2004.

13. Banerjee, S., Carlin, B. and Gelfand, A.E. *Hierarchical Modeling and Analysis for Spatial Data.* CRC Press/Taylor & Francis group, Boca Raton FL, 2004.

14. Wheeler D, Waller L. Mountains, valleys, and rivers: the transmission of raccoon rabies over a heterogeneous landscape. *Journal of Agricultural, Biological, and Environmental Statistics* 2008; **13**: 388406.

15. Vlachos, A., Ghahramani, Z., Briscoe, T. Active learning for constrained Dirichlet process mixture models. *Proceedings of the 2010 workshop on geometrical models of natural language semantics*; Association for Computational Linguistics, 2010.

16. Antoniak, C.E. Mixtures of Dirichlet processes with application to nonparametric problems. *The Annals of Statistics* 1974; **2**: 1152-1174.

17. Gopalan, R., Berry, D.A. Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association* 1998; **93**: 1130-1139.

18. Escobar, M.D., West, M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 1995; **90**: 577-588.

19. Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**: 639650.

20. Jin, X., Banerjee, S., Carlin, B.P. Order-free coregionalized lattice models with ap-

plication to multiple disease mapping. *Journal of the Royal Statistical Society Series B* 2007; **69**: 817-838.

7 Appendix

<u>WinBUGS code for the case count data</u>:

```
model
{

   for( i in 1 :  k ) {

   phi[i] < - theta[Z[i]]
Z[i] ~ dcat(p[1:k])

   for( j in 1 :  t ) {
Y[i , j] ~ dpois(mu[i , j])
log(mu[i , j]) < - beta0 + beta1 * x[i,j] + phi[i] + log(E[i])
}
}

    p[1] < - V[1]
for (j in 2:(k-1)) {p[j] < - V[j] * (1-V[j-1])*p[j-1]/V[j-1]}
for (k1 in 1:(k-1)) {V[k1] ~ dbeta(1,M1)}

   ps < - sum(p[1:(k-1)])
p[k] < - 1-ps
V[k] ~ dbeta(1,M1)

   M1 ~ dgamma(2,1)

   raj < - tau.c
for(k3 in 1:k){theta[k3] ~ dnorm(m0,raj )}

   beta0 ~ dnorm(0.0, 1.0E-5)
beta1 ~ dnorm(0.0, 1.0E-5)
tau.c ~ dgamma(1.0,2)

   }
```

Table 1: Configuration 1 and 2 implies Connecticut map with 2 and 4 clusters respectively. Cases 1, 2 & 3 in rate incidence data correspond to settings where the difference between cluster means are within 2σ, between 2σ and 3σ, more than 3σ distances away respectively. For case count data, cases 1, 2, 3 correspond to gradual increase in the difference between cluster means.

	Case	Configuration
case count	1	1
	1	2
	2	1
	2	2
	3	1
	3	2
rate incidence	1	1
	1	2
	2	1
	2	2
	3	1
	3	2