Statistics and Its Interface Volume 5 (2012) 149–158

Bayesian areal wombling using false discovery rates*

Pei Li, Sudipto Banerjee†,
Alexander M. McBean and Bradley P. Carlin

Spatial data arising in public health services are often reported as case counts or rates aggregated over *areal* regions (e.g. counties, census-tracts or ZIP codes), rather than being referenced with respect to the geographical coordinates of individual residences. For such *areal data*, subsequent inferential interest often resides in the formal identification of "barriers", or "difference boundaries", on the map, where "boundary" refers to a border with sharp changes in outcome on either side. This boundary detection problem is often referred to as "wombling" or, more specifically, "areal wombling" for aggregated areal data, after a foundational article by Womble (1951). Existing statistical frameworks for areal wombling usually follow a two stage procedure: (i) estimate the spatial effects from an appropriate spatial model, and (ii) detect boundaries from those estimates using appropriate discrepancy metrics on those estimates. Lu and Carlin (2005), and several subsequent articles, explored areal wombling within this framework.

This article treats wombling as a hypothesis-testing problem, where we are testing a substantial number of hypotheses – one for each geographical boundary – and seek to provide policy-makers and analysts with a final set of difference boundaries. Here we must reckon with a lurking multiplicity problem arising from the large number of individual hypothesis we are testing. We proffer a computationally feasible framework to estimate hierarchical spatial models that account for dependence between adjacent regions and test for equality of spatial effects, while adjusting for multiplicities using false discovery rates (FDR); see, e.g., Benjamini and Hochberg (1995). A simulation study is conducted to first illustrate and assess the new approach, which is then applied to detect boundaries on a county map of Minnesota that records pneumonia and influenza hospitalization rates from the SEER-Medicare program.

AMS 2000 subject classifications: Primary 62F15, 62H11; secondary 62F03.

Keywords and phrases: Areal data, Bayesian inference, Hierarchical models, False discovery rates, Spatial moving averages.

*Pei Li is Statistician at Medtronic Inc., Sudipto Banerjee is Professor and Bradley P. Carlin is Mayo Professor and Head of Biostatistics, and Alexander M. McBean is Professor of Health Policy and Management, School of Public Health at the University of Minnesota, Minneapolis, MN 55414 (e-mail: *sudiptob@biostat.umn.edu*). This work was supported in part by NIH grants 1-R01-CA95995 and 1-RC-1-GM092400-01.

†Corresponding author.

1. INTRODUCTION

Geographical Information Systems (GIS) software has revolutionized the analysis of spatially referenced health data with its depiction of counts and rates over study areas. In public health services, to protect patient privacy, spatial data are usually available as case counts or rates aggregated over *areal* regions (e.g. counties, census-tracts or ZIP codes), rather than as geographical locations of individual residences. When spatial dependence in the data renders the ordinary least squares regression model unsuitable, alternative models that incorporate spatial dependence should be considered. For areally-referenced, or simply areal, data, the association structures are built upon adjacencies or neighborhood structures for the regions. Here we regard observations from a region to be more similar to those from its neighboring regions than those arising from regions farther away.

Statistical models for areal data have been widely employed for smoothing maps and evincing spatial trends and clusters in econometrics (e.g. Anselin, 1988; Le Sage and Pace, 2009) and public health (e.g., Banerjee et al., 2004; Waller and Gotway, 2004). Subsequent inferential interest often resides in the formal identification of "barriers" or "boundaries" on the spatial surface or map, where 'boundary' refers to a border with sharp changes in outcome on either side. A ripe area of research is the statistical detection of spatial or geographical barriers (also known as *difference boundaries*) that can represent major differences in outcomes between neighboring areal units. Statistical models can help analysts separate significant boundaries from those arising due to random noise in the data. This boundary detection problem is often referred to as "wombling", after a foundational article by Womble (1951). Here we focus exclusively upon finding boundaries on maps for areal data; for other types of spatial data see, e.g., Banerjee (2010).

Algorithmic approaches to areal wombling, also known as polygonal wombling, have been addressed by Jacquez and Greiling (2003a, 2003b). While attractive in their simplicity and ease of use, the algorithmic approaches fail to reckon

with all sources of uncertainty and can produce spurious statistical inference. For instance, public health data often reveal extremeness in counts and rates for thinly populated regions that are attributable to random variation in the observed data, rather than any systemic differences. The algorithmic approaches are unable to adjust for such variability across regions.

A more detailed review of the existing algorithmic approaches and their deficiencies can be found in Lu and Carlin (2005), who were among the first to propose a fully model-based framework for areal wombling using hierarchical conditionally autoregressive models (also see Wheeler and Waller, 2008). Lu and Carlin (2005) explored different metrics for measuring the differences in the estimates (posterior means) of the spatial effects. In the same vein, Lu et al. (2007) and Ma, Carlin and Banerjee (2010) investigated estimating the adjacency matrix within a hierarchical framework using priors on the edges. Li, Banerjee, Hanson and McBean (2010) proffered a class of non-parametric Bayesian hierarchical models for areally aggregated health outcome data that provide stochastic assessments regarding the presence of geographical barriers. These models circumvent the identifiability issues arising from the aforementioned "edge effects" by modeling the spatial effects as almost surely discrete realizations of areally dependent stick-breaking processes (including the Dirichlet process). Subsequent inference is based upon the posterior probability that two spatial effects in neighboring regions are equal.

These methods, however, do not reckon with the multiplicity issues afflicting inference from marginal posterior estimates. This article pursues a simpler formulation that attempts to resolve the multiplicities using false discovery rates (FDR). We formulate the problem of areal wombling as one of testing different boundary hypotheses. A boundary hypothesis posits whether a pair of neighbors have equal spatial random effects or not. We want to test, for each pair of adjacent geographical regions (i.e. neighbors) in a map, a null model that posits equal spatial effects for the two regions against an alternative model that allows unconstrained, but spatially correlated, regional effects. As such, we will have as many hypothesis as there are geographical boundary segments on our map. For example, there are 211 such segments in the county map for the state of Minnesota. Each hypothesis corresponds to a two-component mixture distribution that assigns a point mass to the null hypothesis and distributes the remaining mass to the alternative.

When multiple hypotheses are tested simultaneously, classical inference is usually concerned about controlling the overall Type I error rate. Benjamini and Hochberg (1995) introduced the FDR as an error criterion in multiple testing and described procedures to control it. The FDR is the expected proportion of falsely rejected null hypotheses among all rejected null hypotheses. Bayesian versions of FDR have been proposed and discussed by several authors including Storey (2002; 2003), Genovese and Waserman (2002), Newton et al. (2004) and Broet et al. (2004). Mueller et al. (2008) used a decision theoretic perspective and set up decision problems that lead to the use of FDR-based rules and generalizations. We adapt this framework to our "areal wombling" problem. We depart from the more traditional conditionally autoregressive (CAR) and simultaneous autoregressive (SAR) models used for areal data analysis as they create problems (see Section 2) in implementing the mixture models in our hypothesis framework.

The remainder of the manuscript proceeds as follows. The next section outlines the spatial moving average (SMA) models we employ and gives some of their core properties. Section 3 discusses our framework for developing decision rules accounting for multiple comparisons. Section 4 presents a synthetic data example as well as a real data application, while Section 5 concludes the paper with some discussion.

2. AREAL WOMBLING USING THE SPATIAL MOVING AVERAGE MODEL

Spatial autoregressive models (see, e.g., Banerjee et al., 2004) have been widely employed to account for spatial dependence in areal data sets. These can be broadly classified into two classes: simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models. These have been applied extensively in econometrics (see, e.g., Anselin, 1988, 1990; Le Sage, 1997; Le Sage and Pace, 2009) and public health (see, e.g., Banerjee et al., 2004; Waller and Gotway, 2004). Both of these model classes assign probability distributions to spatial random effects and, with the help of a geographical proximity matrix (e.g. the underlying regional adjacency matrix), capture spatial associations by assuming that neighboring regions exhibit stronger associations than those that are less proximate. They smooth the outcomes across neighboring regions to produce maps that better reveal where the outcome variable tends to cluster. The SAR models achieve this using joint probability distributions with spatially correlated dispersion structures, while the CAR models build spatial dependencies through spatially correlated neighborhood level random effects.

Focusing upon boundary analysis, we feel that both the SAR and CAR models are less conducive to a computationally simple approach that will account for multiplicities when testing for significant differences across regions. The areal wombling problem seeks to learn about difference boundaries from the data by considering the influence of each edge on these models. The model corresponding to the null hypothesis, therefore, will be constrained by making two spatial effects equal and, whatever areal model one considers, the model needs to be estimated once for each boundary. In this context, the SAR model proves computationally exorbitant because its estimation involves matrix inversions, while the standard CAR runs into technical difficulties that arise from it being an "improper" distribution. A "proper" CAR model that yields integrable joint distributions (e.g.

Banerjee et al., 2004) is an option, but the so called propriety parameter here is often difficult to estimate in practice due to lack of identifiability from the data.

Instead, we employ a class of discrete spatial moving average models (SMA) that incorporate dependencies through a *weighted average* of uncorrelated latent risk factors. As with the SAR and CAR models, we form a geographical proximity matrix W, whose (i,j)-th entry, w_{ij}, connects areal units i and j spatially in some fashion. Customarily w_{ii} is set to 0. Possibilities include binary choices, i.e. $w_{ij}=1$ if i and j share some common boundary or perhaps a vertex (as in a regular grid), and $w_{ij}=0$ otherwise. Alternatively, w_{ij} could reflect "distance" between units, e.g., a decreasing function of inter-centroidal distances between the units. The entries in W can be viewed as weights; more weight will be associated with j's closer (in some sense) to i than those farther away from i. Henceforth, unless otherwise stated, we use a binary adjacency matrix.

Let Y_i be the outcome variable (e.g. count or rate) and $\boldsymbol{x}_i$ be a vector of explanatory variables for areal unit i. The hierarchical smoothed moving average (SMA) model is

$$(1) \qquad Y_i \,|\, \boldsymbol{\beta}, \phi_i \sim Poisson\left(e^{\boldsymbol{x}_i'\boldsymbol{\beta}+\phi_i}\right); \quad i=1,\ldots,n$$

where

$$\phi_i = \alpha\psi_i + (1-\alpha)\sum_{l=1}^{n}\frac{w_{il}}{w_{i+}}\psi_l; \quad \psi_i \overset{i.i.d.}{\sim} N(0,\sigma^2).$$

Here the ψ_i's can be thought of as latent unobserved risk factors associated with regions indexed by i, and ϕ_i is the spatial random effect associated with region i which represents the cumulative effect of these unmeasured risk factors in each area. The ϕ_i's are spatial effects that borrow strength by averaging the latent risk factors over their neighbors.

The SMA is a very natural and flexible type of spatial process that involves integrals or sums of independent and identically distributed random variables. Note that (1) may be looked upon as a discretized version of SMA's often used to describe continuous spatial processes, particularly in geostatistical applications. Such models are constructed by integrating a simple two-dimensional random noise process with a smoothing kernel that is a function of distance. Cressie and Pavlicova (2002) defined a Gaussian SMA by the stochastic integral $Z(\boldsymbol{s}) = \int k(\boldsymbol{s},\mu)V(d\mu); \boldsymbol{s}\in\mathcal{D}$, where $\mathcal{D}$ is the spatial domain where $\boldsymbol{s}$ resides, $V(\cdot)$ is a spatial process of independent increments defined on $\Re^d$, and $k(\cdot,\cdot)$ is a kernel function that mitigates the random noise process in two-dimensional space to yield smoother surfaces. In the discrete version, instead of using kernel functions based upon distance, we employ a weight matrix which is based upon the adjacency structure of the map. See also Haining (1978), who employed similar ideas to study spatial interactions on a rectangular lattice.

The joint distribution of the spatial effects in (1) arises from the linear transformation $\boldsymbol{\phi} = (\alpha\boldsymbol{I}+(1-\alpha)\tilde{W})\boldsymbol{\psi}$, where $\boldsymbol{\phi} = (\phi_1,\phi_2,\ldots,\phi_n)'$ and $\boldsymbol{\psi} = (\psi_1,\psi_2,\ldots,\psi_n)'$. Writing $B = (\alpha\boldsymbol{I}+(1-\alpha)\tilde{W})$, where $\tilde{W}$ is the row-normalized weight matrix with w_{ij}/w_{i+} as its (i,j)-th element, the spatial effects $\boldsymbol{\phi}$ follow a multivariate normal distribution with zero mean and variance-covariance matrix σ^2BB'. This will admit a proper density if and only if B is nonsingular. Let D be a diagonal matrix whose i-th diagonal element is the number of neighbors of region i. Then B, though itself not symmetric, can be written as

$$B = D^{-1/2}\left(\alpha\boldsymbol{I}+(1-\alpha)D^{-1/2}WD^{-1/2}\right)D^{1/2},$$

which implies that $D^{1/2}BD^{-1/2}$ is symmetric. Assuming the diagonal elements of D are strictly positive (i.e., the map is connected), B is nonsingular if and only if $D^{1/2}BD^{-1/2}$ is. Let $\lambda(D^{-1/2}WD^{-1/2}) = \left\{\lambda_{(1)},\lambda_{(2)},\ldots,\lambda_{(n)}\right\}$ be the set of eigenvalues of the adjacency matrix $D^{-1/2}WD^{-1/2}$, arranged in ascending order. The eigenvalues of $\left(\alpha\boldsymbol{I}+(1-\alpha)D^{-1/2}WD^{-1/2}\right)$ are then given by the set $\{\alpha+(1-\alpha)\lambda_{(1)},\alpha+(1-\alpha)\lambda_{(2)},\ldots,\alpha+(1-\alpha)\lambda_{(n)}\}$. Since $\mathrm{tr}(D^{-1/2}WD^{-1/2}) = \sum_{i=1}^{n}\lambda_{(i)} = 0$, we have $\lambda_{(1)} < 0 < \lambda_{(n)}$. The nonsingularity of B is now assured when each $\alpha+(1-\alpha)\lambda_{(i)}$ is nonzero. Therefore, we obtain

$$\alpha \neq \frac{-\lambda_{(i)}}{1-\lambda_{(i)}} \quad \text{for } i=1,2,\ldots,n$$

as a sufficient condition for the nonsingularity of B. It can also be shown that $\lambda_{(n)}=1$, but this is not of much use to us here.

A simpler sufficient condition for the nonsingularity of B is to let $1/2<\alpha<1$, which ensures diagonal dominance, and hence nonsingularity, of B irrespective of the rank of $\tilde{W}$. In other words, any prior for α with support on $(1/2,1)$ will yield proper distributions for the spatial effects on any connected map. If the map has islands, then we simply use this prior for each connected component (island) of the map. To be precise, the spatial effects for regions within an island are assigned their own SMA prior (i.e. there is a different α and W for each island), while spatial effects across different islands are assumed independent. We note that in our current application (using a Minnesota map), we do not have islands so we have not pursued this in detail.

It is worth pointing out the related work in Ickstadt and Wolpert (1998) and Best et al. (2000). They proposed Poisson-gamma spatial moving average models for use in identity-link Poisson regression models. Identical and independent gamma priors are assumed for the underlying risk factors, as this enables the MCMC sampler to exploit conjugacy with the Poisson likelihood. The Poisson-gamma SMA model may be implemented in `WinBUGS` using the readily available `pois.conv` distribution (Lunn et al., 2000). However, in our multiple hypothesis testing framework, computations involving the adjacency matrix W makes it more computationally intensive to execute `pois.conv`, while the model in (1) is fairly effective.

3. FDR BASED DECISION RULES

Our current objective is to test whether the geographical boundary given by the ordered pair (i,j) is a difference boundary. Let $\mathcal{E} = \{(i,j) : w_{ij} \neq 0; i,j = 1,2,\ldots,n\}$ be the set of all geographical boundaries on a map with adajacency matrix W. For each $(i,j) \in \mathcal{E}$, our null hypothesis posits that (i,j) is not a difference boundary, i.e. $\phi_i = \phi_j$, while the alternative is $\phi_i \neq \phi_j$. In other words, we want to look at each ordered pair (i,j) such that $w_{ij} = 1$ and test whether the spatial effects ϕ_i and ϕ_j are "equal". At the outset, note that the prior densities for the spatial effects in (1) are continuous, so $P(\phi_i = \phi_j) = 0$ both a priori and a posteriori. For the prior on the spatial effects we will, therefore, adopt a two-component mixture density that places some positive mass on the null hypothesis. To be more precise, we impose an equality constraint upon the spatial effects for regions i and j, which yields, for every (i,j) that represents a geographical neighbor, the following two-component mixture prior for the spatial random effects:

$$f_{(i,j)}(\boldsymbol{\phi}) = H_{0,(i,j)} f_{0,(i,j)}(\boldsymbol{\phi}) + (1 - H_{0,(i,j)}) f_1(\boldsymbol{\phi}), (i,j) \in \mathcal{E}. \tag{2}$$

Here, $H_{0,(i,j)}$ is an indicator variable that equals one if (i,j) is *not* a difference boundary (i.e. $\phi_i = \phi_j$) and equals zero if (i,j) is a difference boundary; $f_1(\boldsymbol{\phi})$ is the prior density for $\boldsymbol{\phi}$ as specified in (1), which is precisely $N(\mathbf{0}, \sigma^2 BB')$, and $f_{0,(i,j)}(\boldsymbol{\phi})$ is the density obtained from $f_1(\boldsymbol{\phi})$ subject to the constraint $\phi_i = \phi_j$. Note that this is a linear constraint on $\boldsymbol{\phi}$, which yields a constrained (singular) normal density (e.g. Rao, 1973, Sec 8a.4) for $f_{0,(i,j)}(\boldsymbol{\phi})$ in (2). This density exists on an $n-1$ dimensional subspace and will yield a valid joint posterior density for the spatial effects as long as $H_{0,(i,j)}$ is not a degenerate random variable. For each (i,j) such that $w_{i,j} = 1$, we estimate the hierarchical model

$$\begin{aligned} Y_k \,|\, \boldsymbol{\beta}, \phi_i &\sim Poisson\left(e^{\mathbf{x}_k'\boldsymbol{\beta} + \phi_k}\right); \quad k = 1,\ldots,n \\ \boldsymbol{\phi} = \{\phi_k\}_{k=1}^n &\sim f_{(i,j)}(\boldsymbol{\phi}) \\ &= H_{0,(i,j)} f_{0,(i,j)}(\boldsymbol{\phi}) + (1 - H_{0,(i,j)}) f_1(\boldsymbol{\phi}); \\ H_{0,(i,j)} &\sim \text{Ber}(\pi); \quad \pi \sim \text{Beta}(a,b) \end{aligned} \tag{3}$$

using a Gibbs sampler with Metropolis steps. This yields posterior samples for $\boldsymbol{\beta}$, $\{\phi_k\}$, $H_{0,(i,j)}$ and π.

From a practical implementation standpoint, we will avoid working with the singular density $f_{0,(i,j)}(\boldsymbol{\phi})$ in (3). Recall from (1) that the ϕ_i's are linear transformations of the ψ_i's, the latter being independently and identically distributed normal random variables. This means that posterior samples of the ψ_l's will immediately deliver samples of the ϕ_i's. It can be shown, after some algebra, that $\phi_i = \phi_j$ is equivalent to

$$\begin{aligned} &\alpha(\psi_i - \psi_j) + (1-\alpha)\sum_{l=1}^n \left(\frac{w_{il}}{w_{i+}} - \frac{w_{jl}}{w_{j+}}\right)\psi_l = 0 \\ &\iff \psi_i = \frac{\alpha - (1-\alpha)\frac{w_{ij}}{w_{i+}}}{\alpha - (1-\alpha)\frac{w_{ji}}{w_{j+}}}\psi_j \\ &\quad - \frac{1-\alpha}{\alpha - (1-\alpha)\frac{w_{ji}}{w_{j+}}} \sum_{l \neq i,j} \left(\frac{w_{il}}{w_{i+}} - \frac{w_{jl}}{w_{j+}}\right)\psi_l \\ &\iff \psi_j = \frac{\alpha - (1-\alpha)\frac{w_{ji}}{w_{j+}}}{\alpha - (1-\alpha)\frac{w_{ij}}{w_{i+}}}\psi_i \\ &\quad - \frac{1-\alpha}{\alpha - (1-\alpha)\frac{w_{ij}}{w_{i+}}} \sum_{l \neq j,i} \left(\frac{w_{jl}}{w_{j+}} - \frac{w_{il}}{w_{i+}}\right)\psi_l. \end{aligned} \tag{4}$$

This constraint, though somewhat daunting in appearance, is straightforward to program in the `BUGS` language (Lunn et al., 2000) allowing us to easily sample the ψ_l's in (1) using a Gibbs sampler. Equation (4) imposes a linear constraint on the ψ_l's, which means that there are only $n-1$ free parameters among the ψ_l's. Replacing one of them, say ψ_j, in Model (1) with the constraint in (4), enables us to express (1) as a function of the remaining $n-1$ free parameters. This is easily specified in the `BUGS` language – we simply assign independent and identically distributed normal distributions (as in (1)) to each of the ψ_l's *except* ψ_j, which is set to the expression in (4).

Simpler, albeit somewhat restrictive, formulations of the null are also possible. The following, for instance, sets $\phi_i = \phi_j$ to be equal to their average,

$$\phi_k = \begin{cases} \frac{\alpha}{2}(\psi_i + \psi_j) + \frac{1-\alpha}{2}\sum_{l=1}^n (\frac{w_{il}}{w_{i+}} + \frac{w_{jl}}{w_{j+}})\psi_l, & \text{if } k = i,j \\ \alpha\psi_k + (1-\alpha)\sum_{l=1}^n \frac{w_{kl}}{w_{k+}}\psi_l, & \text{if } k \neq i,j, \end{cases} \tag{5}$$

where $\psi_i \overset{i.i.d.}{\sim} N(0,\sigma^2)$. This would correspond to a model where the entries in the i-th and j-th rows of the adjacency matrix have been replaced by their simple averages.

From a Bayesian decision-making perspective, the spatial analyst will want to identify a boundary (i,j) as a difference boundary if the posterior probability that $H_{0,(i,j)}$ equals zero exceeds a certain threshold t. In more formal parlance, we define a *critical region* or rejection region for the null hypothesis to be the set

$$A_{(i,j)}(Y;t) = \{Y : v_{(i,j)} = P(H_{0,(i,j)} = 0 \,|\, Y) > t\},$$

where $Y = \{Y_1, Y_2, \ldots, Y_n\}$. The choice of t will be based upon controlling the false discovery rate (FDR) below a level $\delta = 0.05$. Specifically, the FDR is defined to be

$$FDR = \frac{\sum_{(i,j)\in\mathcal{E}} H_{0,(i,j)} 1(v_{(i,j)} > t)}{\sum_{(i,j)\in\mathcal{E}} 1(v_{(i,j)} > t)}; \quad \mathcal{E} = \{(i,j) : w_{ij} = 1\}. \tag{6}$$

Estimation of the above quantity simplifies considerably in a Bayesian setting. The $v_{(i,j)}$'s are functions purely of the

data, which, for computing posterior expectations, is constant. Therefore the only unknown in the above definition are the $H_{0,(i,j)}$'s in the numerator. The posterior estimate of the FDR is now given by

$$(7)\quad \widehat{FDR} = \mathrm{E}[FDR\,|\,Y] = \frac{\sum_{(i,j)\in\mathcal{E}}(1-v_{(i,j)})1(v_{(i,j)}>t)}{\sum_{(i,j)\in\mathcal{E}}1(v_{(i,j)}>t)},$$

where an estimate of the posterior probability $v_{(i,j)} = P(H_{0,(i,j)} = 0\,|\,Y)$ is computed as a Monte Carlo mean of the posterior samples for $H_{0,(i,j)}$, i.e., the number of times when $H_{0,(i,j)} = 0$ divided by the length of the simulation run. Rejection rules can be then constructed to bound the FDR at target level δ: reject if $v_{i,j} > t$, where

$$t = \sup\left\{u : \frac{\sum_{(i,j)\in\mathcal{E}} I(v_{(i,j)}>u)(1-v_{(i,j)})}{\sum_{(i,j)\in\mathcal{E}} I(v_{(i,j)}>u)} \le \delta\right\}.$$

The above bound depends upon the estimated FDR and its accuracy and can be sensitive to the choice of the priors. We discuss this further in the simulation example.

4. ILLUSTRATIONS

We implemented our approach in Section 3 using the constraint in (4) and the slightly more specific constraint in (5). These models were run within the `R` statistical framework using the `BRugs` package (http://www.stats.ox.ac.uk/pub/RWin/) that can execute embedded `WinBUGS` scripts from within `R`. Both these approaches yielded essentially indistinguishable inference with regard to boundary detection, but the latter is easier to program and is computationally more efficient, delivering post burn-in posterior samples with approximately 20% savings in CPU time. Therefore, in our subsequent examples we present only the results from (5). We illustrate our proposed approach in a simulation study, and then apply it to real data analysis in Section 4.2. On a workstation using an Intel dual core 4 GHz processor our entire simulation exercise, where we analyzed 50 simulated datasets on a Minnesota map, took less than five hours of CPU time. Our analysis of the Minnesota Pneumonia and Influenza data consumed less than fifteen minutes of CPU time.

4.1 Synthetic example

The simulation study serves two main purposes: 1) to evaluate the proposed model performance in detecting true difference boundaries as compared to existing methods; and 2) to identify which levels of FDR can be accurately estimated by the SMA model.

Our synthetic example is based on a Minnesota county map. There are 87 counties and 211 geographical boundaries between counties on the map; thus, there are 211 different boundary hypotheses in our analysis. We divided the Minnesota map into six regions, and let $\mu_i \in \{0, 0.5, 1, 1.5, 2, 2.5\}$. Letting Y_i be the simulated number of cases in county i, we generate $\{Y_i\} \sim Poisson(5 \times \exp(\mu_i))$ for $i = 1, 2, \ldots, 87$. This resulted in six well defined clusters with 47 true difference boundaries mapped in Figure 1. Note that two of the clusters are shaded white. The one in the interior comprises a single county (Sherburne) and has mean 0, while the other has a mean of 0.5. This configuration creates a county with all its boundaries being true difference boundaries.

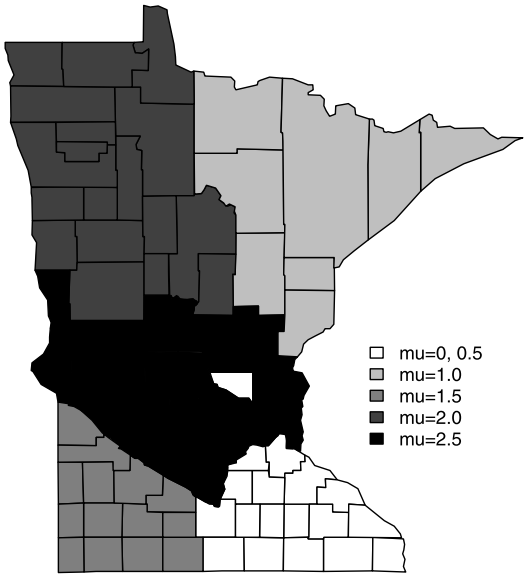

Figure 1. A map of the simulated data with the grey-scales showing the six different clusters, each having its own mean. There are 47 boundary segments that separate regions with different means (shades).

A uniform $(0, 1)$ prior is assigned to α. The conditions for positive-definiteness of the variance-covariance matrix for $\boldsymbol{\phi}$ were discussed in Section 2. The prior distribution for the precision in (1) was taken to be a weakly informative Gamma distribution, $p(\frac{1}{\sigma^2}) \propto \Gamma(0.01, 0.01)$, and we also take a flat prior for β. For the hyperparameters, we set $a = 1$ and $b = 9$ after some preliminary explorations. Ideally, the hyperparameters a and b should reflect the proportion of true difference boundaries on the map. In practice, unfortunately, such prior information is rarely available and naive choices of a and b may detract from the performance of our approach. However, in our experience, even informal approaches such as obtaining eye-ball estimates of difference boundaries from choropleth maps of the raw data can suggest hyperparameter values that seem to deliver robust inference with regard to boundary detection. Alternative approaches include using Boundary Likelihood Values (BLV's) (Jacquez and Greiling, 2003a, 2003b) or the model-based (LC) method of Lu and Carlin (2005) to arrive at initial estimates for robust boundary detection. A sensitivity analysis was carried out with varying number of true difference boundaries and we

found that setting hyperparameters using any of the above approaches delivered essentially indistinguishable posterior inference.

For each of 50 simulated datasets, we estimate the 211 hierarchical models described in (1) and (3), one for each geographical boundary, using Markov chain Monte Carlo methods (see, e.g., Carlin and Louis, 2009; Gelman et al. 2004). We assumed only an intercept term (i.e. $\boldsymbol{x}_i \equiv 1$ in (1)) in the mean, with β as the corresponding global mean parameter. Upon convergence, each model yields posterior samples of β, the ϕ_i's and α. For a typical simulated dataset, the posterior means across the 211 models for β hovered between -0.23 and -0.21, while the posterior standard deviation ranged between 0.13 and 0.17; the posterior mean for α was between 0.37 and 0.43 and the posterior standard deviation was between 0.021 and 0.045.

For every pair of geographical neighbors (i, j), we compute the posterior probability $P(H_{0,(i,j)} = 0 \,|\, \{Y_k; k = 1, 2, \ldots, n\})$; higher posterior probabilities provide evidence in favor of (i, j) being a difference boundary. For illustrative purposes, we choose the top $T = 35, 40, 45, 50$ and 55 edges with the highest posterior probabilities. In practice, health professionals might seek to identify a "top bracket" of difference boundaries. Our choices of T considers 17% to 26% of the most probable difference boundaries (based upon their posterior probabilities). Since there are 47 true difference boundaries, these choices encompass settings where we could, theoretically have obtained 100% accuracy (when $T = 35, 40, 45$) and also where we are assured of a few false positives (when $T = 50, 55$).

Since we know the true difference boundaries, we can obtain the sensitivity and specificity for the SMA model; the sensitivity corresponds to the probability of correctly detecting a true difference boundary, while specificity corresponds to the probability of correctly rejecting a difference boundary. We compare the performance of our method with two existing methods: the deterministic Boundary Likelihood Value (BLV) algorithm of Jacquez and Greiling (2003a, 2003b) and the model-based approach of Lu and Carlin (2005). The average detection rates for these different methods applied to the 50 simulated datasets are listed in Table 1. Our proposed method seems to be slightly outperforming the two existing methods in both sensitivity and specificity under all five scenarios.

In order to investigate if the $\widehat{FDR}$ based decision rule can be used for controlling false discovery rates, we compare the estimated FDR ($\widehat{FDR}$) and the true FDR after choosing a threshold t. The estimated FDR can be worked out by equation (7), while the true FDR is given by (6). By examining the closeness of the estimated FDR to the true FDR, we are able to assess the accuracy of the FDR estimation by the proposed approach. Figure 2 plots the $\widehat{FDR}$ against the number of edges selected from a cutoff value t. Also plotted are the realized FDR computed (dashed line) based on the discoveries and the underlying truth from which we simulated the data. Both estimated and realized FDR curves are averages over the 50 simulated datasets for each number of edges selected. The two curves follow each other closely on the plot, demonstrating that the FDR is well estimated by the proposed approach. The FDR in the lower level region (<0.6) is slightly overestimated, while in the higher level region (>0.6) is slightly underestimated. This suggests that we are being conservative here because only the lower region is of interest when the FDR is controlled by a certain target, say 10%.

Table 1. Sensitivity and specificity in the simulation study (50 datasets) for the SMA model, the Boundary Likelihood Value approach of Jacquez and Greiling (2003a) and the approach of Lu and Carlin (2005) model. The simulation study was based on a Minnesota county map

T	Method	Sensitivity	Specificity
35	SMA	0.740	0.998
	BLV	0.711	0.990
	LC	0.702	0.989
40	SMA	0.818	0.991
	BLV	0.778	0.979
	LC	0.767	0.976
45	SMA	0.872	0.975
	BLV	0.831	0.964
	LC	0.813	0.959
50	SMA	0.901	0.955
	BLV	0.869	0.944
	LC	0.859	0.941
55	SMA	0.925	0.930
	BLV	0.891	0.920
	LC	0.881	0.917

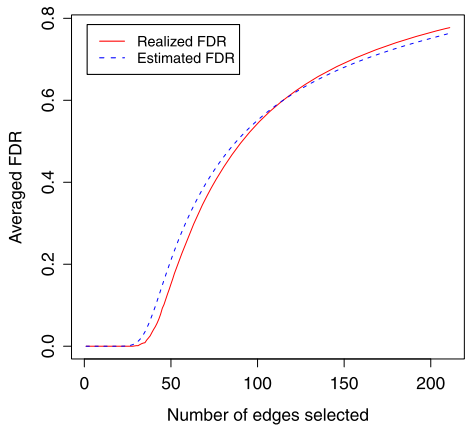

Figure 2. The realized FDR and the estimated FDR curves in the simulation study. The x-axis is number of edges selected as difference boundaries.

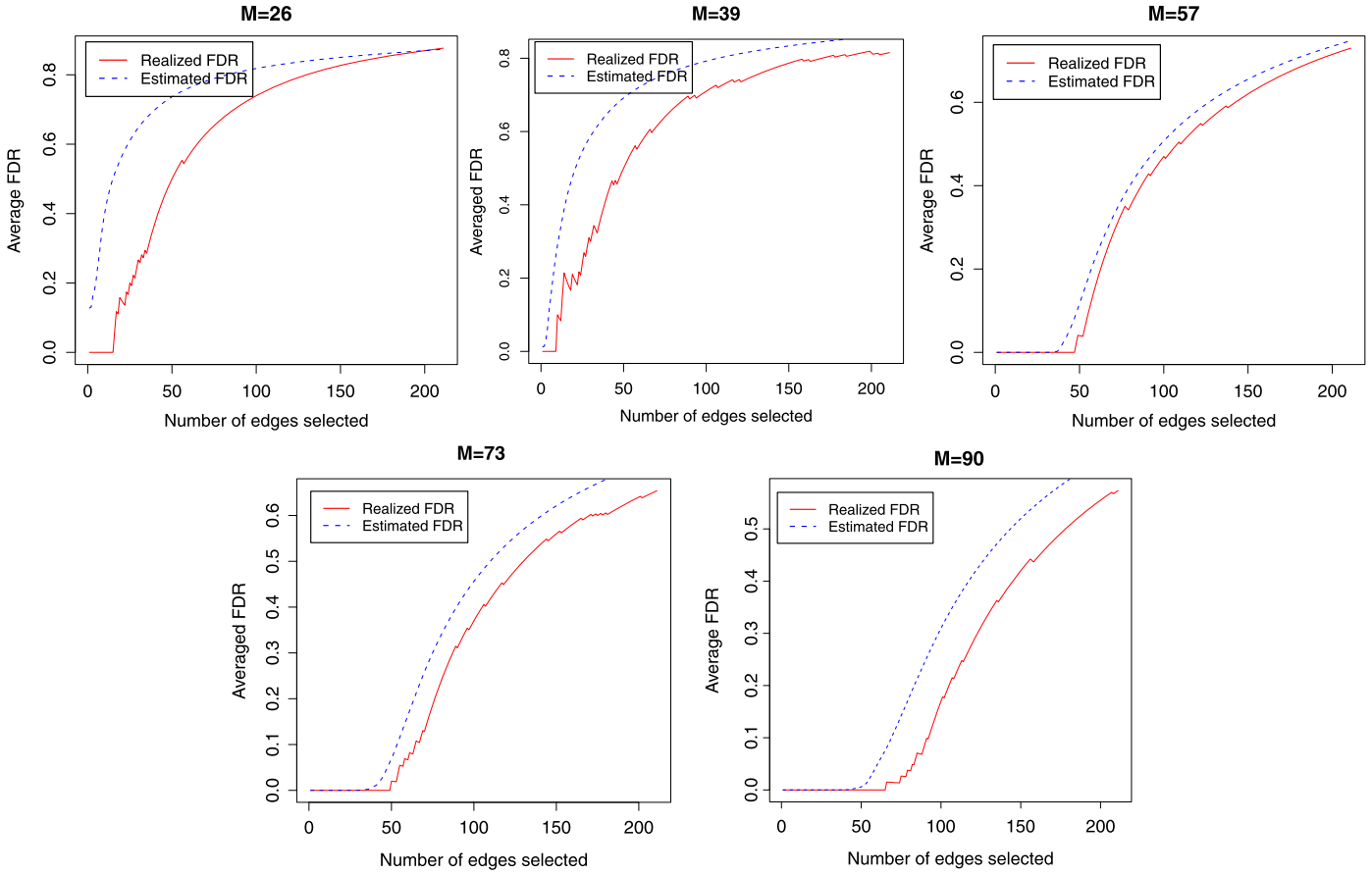

Figure 3. The realized FDR and the estimated FDR curves in the simulation study for $M = 26, 39, 57, 73, 90$. The x-axis is number of edges selected as difference boundaries.

Finally, we check the sensitivity of the prior selection to the number of true difference boundaries M. Figure 3 compares the estimated FDR with the realized FDR when $M = 26, 39, 57, 73$ and 90. Rather than the patterns of true difference boundaries, here we are more interested in comparisons between the estimated and the true FDR for varying numbers of true difference boundaries. The plots show that the estimated FDR lies slightly above the realized FDR most of the time. An exception to this conservatism is when both of them are exactly zero and when M is 57, 73 or 90. Overestimating the FDR would make the practitioner declare more errors than he actually made, which is arguably better than underestimating the FDR to achieve a pre-specified level. Although this may sacrifice some power, the model allows for more conservative control of the FDR as is desirable.

4.2 The Minnesota Pneumonia and Influenza data analysis

We illustrate our model comparison approach in the context of a Minnesota Pneumonia and Influenza ($P\&I$) diagnosis dataset. Influenza and pneumonia are major causes of illness and death. In 2005, these conditions ranked as the eighth leading cause of death in the United States and the sixth leading cause in people over 65 years of age. An active surveillance program for an influenza-like illness can help impede the spread of the infection by appropriate intervention. Boundary analysis can help identify "health barriers" separating counties that experience different impacts of the influenza virus. Identifying difference boundaries can improve coordination between neighbors and execute plans for hospital needs and antiviral or vaccine interventions.

Our dataset includes subjects older than 65 years who were enrolled in both Medicare part A and part B in December 2001. Residents of Minnesota who were 65 years of age and older and who were enrolled in the Medicare fee-for-service program as of December 31, 2001, formed our study population. This population had been identified as part of a multi-year study regarding the impact of vaccinations on elderly Minnesota residents. The Medicare Denominator file for 2001 was used to define the cohort. In addition to meeting the age and state of residence criteria, to be eligible for inclusion in the study, the person had to be enrolled in both Medicare Part A and Medicare Part B, not be enrolled in

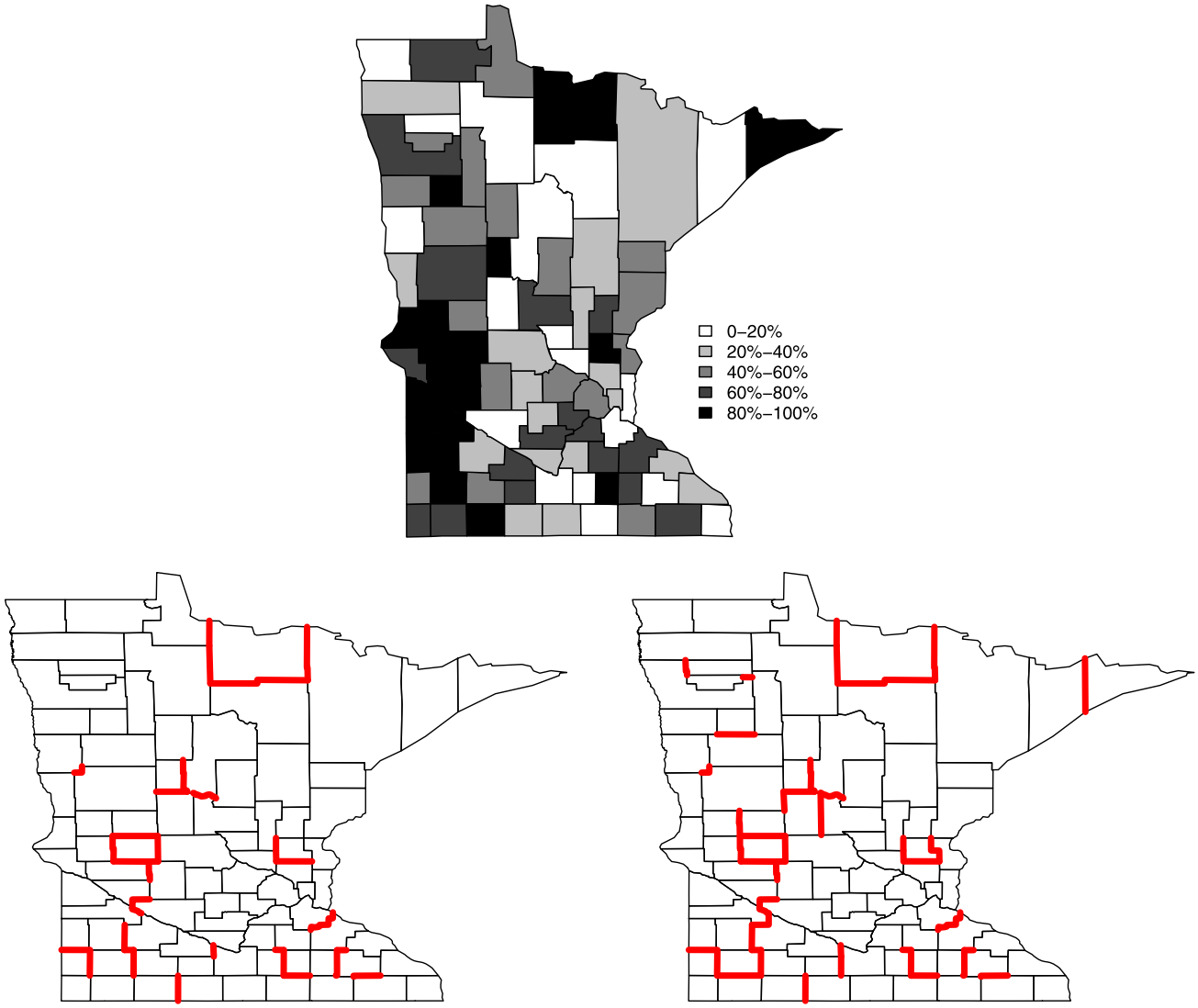

Figure 4. The plot on the top is the choropheth map of pneumonia and influenza hospitalizations in the MN (P&I) dataset. The plot on the lower left highlights the difference boundaries declared when FDR is controlled at 5%*, and the plot on the lower right highlights the ones when FDR is controlled at* 10%.

a Medicare Advantage health plan, and not have end-stage renal disease. The denominator file also indicated the county of residence for each person.

Hospitalizations for Pneumonia and Influenza were identified by the Medicare Provider Analysis and Review (MedPAR) short stay inpatient file for the above Minnesota residents. This annual file contains extensive patient records per hospitalization based on the date of discharge. Rates of $P\&I$ hospitalization are traditional measures of the impact of influenza virus in the elderly population. We identify the boundaries that separate the more affected areas from the less affected areas. Here we studied the number of hospitalizations from $P\&I$ in both influenza and a "shoulder" period among persons at risk in each county. The definition we, and others, have used to define the influenza season is period of weeks that include the 2.5-th to 97.5-th percentile of all influenza isolates for a given influenza year (July 1, 2000 through June 2001, for example). The "shoulder" period includes the weeks on either end of the influenza season beginning with the week of the first influenza isolate and extending through the week of the last isolate.

Let Y_i be the observed number of hospitalizations in county i, O_i be the population of county i, $E_i = \frac{\sum_{k=1}^{n} Y_i}{\sum_{k=1}^{n} O_i} O_i$ be the expected number of cases (under the assumption of no spatial variation in rates), where n is the total number of counties. The map appearing in the top row of Figure 4 shows the raw data. The high-hospitalization counties are scattered over the map, with a clump in the southwest

Table 2. Names of adjacent counties that have significant boundary effects from the SMA model. The county pairs (i, j) are arranged in descending order based upon the estimated posterior probability for $H_{0(i,j)}$ being equal to zero. Numbers 1–33 are detected to be difference boundaries when the FDR is controlled at 5%, and numbers 1–42 are detected when the FDR is controlled at 10%

1	Beltrami, Koochiching	22	Isanti, Mille Lacs
2	Cass,Wadena	23	Lyon, Redwood
3	Douglas, Pope	24	Todd, Wadena
4	Goodhue, Olmsted	25	Pope, Swift
5	Itasca, Koochiching	26	Fillmore, Olmsted
6	Kandiyohi, Pope	27	Jackson, Martin
7	Koochiching, St. Louis	28	Dodge, Olmsted
8	Pope, Stearns	29	Murray, Pipestone
9	Steele, Waseca	30	Kandiyohi, Swift
10	Anoka, Isanti	31	Rice, Waseca
11	Dakota, Goodhue	32	Chippewa, Renville
12	Freeborn, Steele	33	Blue Earth, Brown
13	Pope, Stevens	34	Otter Tail, Todd
14	Cass, Morrison	35	Murray, Nobles
15	Cottonwood, Murray	36	Pennington, Polk
16	Isanti, Sherburne	37	Becker, Mahnomen
17	Lincoln, Pipestone	38	Cook, Lake
18	Clay, Otter Tail	39	Blue Earth, Watonwan
19	Murray, Redwood	40	Chisago, Isanti
20	Renville, Yellow Medicine	41	Redwood, Yellow Medicine
21	Koochiching, Lake of the Woods	42	Morrison, Todd

and some isolated regions surrounded by sparsely inhabited counties that also have lower counts.

We fit the model in (3) and used (7) to identify boundaries for the $P\&I$ hospitalization map. Fixing FDR at 5%, the proposed method identifies 33 difference boundaries, while with FDR at 10% it proposes 42 difference boundaries. Table 2 lists these 42 adjacent counties having the boundary effect, ranked by the posterior probabilities of $H_{0,(i,j)=0}$. The boundaries corresponding to 5% and 10% FDR are highlighted in the two maps in the lower row of Figure 4.

Though the method makes no effort to draw connected series of boundary segments, the higher hospitalization region in the southwest is largely isolated from the remainder of the map. Also note that Koochiching county (North-central county with high hospitalization) is completely isolated from its three neighbors, even when FDR is controlled at the 5% level.

5. DISCUSSION AND FUTURE WORK

This article demonstrated how hierarchical spatial moving average mixture models can be used to detect difference boundaries on areal maps by controlling the false discovery rate. The method's appeal lies in that it is easily estimable (can be implemented in `WinBUGS`) and, based upon our simulation experiments, it tends to enjoy higher sensitivity and specificity compared to some existing methods. The proposed method can also produce an estimate of the error measure, such that the practitioner is aware of the errors incurred by any decision.

A potential issue is the sensitivity of the inference to the hyperparameters of the mixture probability π. Simulation experiments, such as in Section 4.1, are often used to ascertain hyperparameter values that can yield robust inference. We use descriptive wombling with BLV's (e.g. Jacquez and Greiling, 2003a, 2003b) and the method of Lu and Carlin (2005) to obtain some idea about the proportion of difference boundaries on the map, and set the hyperparameters accordingly. In our current setting, we did not find significant performance differences between the BLV and LC methods. Nevertheless, a more elaborate exploration to ascertain performance gains achieved by setting hyperparameters with the BLV and LC methods over more ad-hoc approaches can be worthwhile.

An apparent disadvantage of our current approach is that we need to estimate as many models as there are geographical boundaries. This is in contrast to the model-based wombling approaches (e.g. Lu et al., 2007; Ma et al. 2010; Li et al. 2010) that jointly estimate difference boundaries, perhaps thereby circumventing the need to adjust for the false discovery rates. From an implementation standpoint, however, the current approach is much simpler and, as already mentioned, can be entirely executed in `BRugs`.

Finally, we point out that here we rejected null hypotheses by fixing a pre-specified FDR level. Variations of the decision rules under other forms of loss functions are certainly possible and discussed by Mueller et al. (2008). Interesting

options include loss functions that are linear combinations of the false discovery rate and false negative rate, or bivariate loss functions that explicitly acknowledge both. These are future directions of research that extend naturally from our current work.

Received 1 February 2011

REFERENCES

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Boston.

Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science*, **30**, 185–207.

Assuncao, R. and Krainski, E. (2009). Neighborhood dependence in Bayesian spatial models. *Biometrical Journal*, **51**, 851–869. MR2751717

Banerjee, S. (2010). Spatial gradients and wombling. In *Handbook of Spatial Statistics*. Ed(s) P. Diggle, M. Fuentes, A. E. Gelfand and P. Guttorp, Taylor and Francis, Boca Raton, FL. MR2761512

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC Press, Boca Raton, FL.

Banerjee, S. and Gelfand, A. E. (2006). Bayesian wombling: curvilinear gradients assessment under spatial process models. *Journal of the American Statistical Association*, **101**, 1487–1501. MR2279474

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300. MR1325392

Best, N. G., Ickstadt, K., Wolpert, R. L. and Briggs, D. J. (2000). Combining models of health and exposure data: The SAVIAH study. In *Spatial Epidemiology: Methods and Application*. Oxford University Press, 393–414.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis*. 3rd ed. Chapman and Hall/CRC Press, Boca Raton, FL. MR2442364

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, 2nd ed. Wiley, New York. MR1239641

Cressie, N. and Pavlicova, M. (2002). Calibrated spatial moving average simulations. *Statistical Modelling*, **2**, 267–279. MR1951585

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman and Hall/CRC Press, Boca Raton, FL. MR2027492

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society B*, **64**, 499–518. MR1924303

Haining, R. P. (1978). The moving average model for spatial interaction. *Transactions of the Institute of British Geographers*, New Series **3**, 202–225.

Ickstadt, K. and Wolpert, R. L. (1998). Multiresolution assessment of forest inhomogeneity. In *Case Studies in Bayesian Statistics*, **3**. Lecture Notes in Statistics, 121. Ed(s) C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi and N. D. Singpurwalla. Springer-Verlag, New York, 371–386. MR1630084

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

Jacquez, G. M. and Greiling, D. A. (2003b). 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*, **2**, 4.

LeSage, J. P. (1997) Analysis of spatial contiguity influences on state price level formation. *International Journal of Forecasting*, **13**, 245–253.

LeSage, J. P. and Pace, K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC, Boca Raton, FL. MR2485048

Li, P., Banerjee, S., Hanson, T. A. and McBean, A. M. (2010). Nonparametric hierarchical modeling for detecting boundaries in areally referenced spatial datasets. Technical Report rr2010-014, Divison of Biostatistics, School of Public Health, University of Minnesota, Twin Cities.

Lu, H. and Carlin, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, **37**, 265–285.

Lu, H., Reilly, C., Banerjee, S. and Carlin, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, **14**, 433–452. MR2405556

Lunn, D., Thomas, A., Spiegelhalter and Best, N. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

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

Mueller, P., Parmigiani, G. and Rice, K. (2006). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8*. Ed(s) J. M. Bernardo, S. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West. Oxford University Press.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York. 2 MR0346957

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, **64**, 479–498. MR1924302

Storey, J. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, **31**, 2013–2035. MR2036398

Waller and Gotway. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley and Sons, New York. MR2075123

Wheeler, D. and Waller, L. (2008). Mountains, valleys, and rivers: the transmission of raccoon rabies over a heterogeneous landscape. *Journal of Agricultural, Biological and Environmental Statistics*, **13**, 388–406. MR2590936

Wolpert, R. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–267. MR1649114

Womble, W. H. (1951). Differential systematics. *Science*, **114**, 315–322.

Pei Li
710 Medtronic Parkway
Minneapolis, MN 55432-5604
USA
E-mail address: pei.li@medtronic.com

Sudipto Banerjee
420 Delaware Street S.E.
A460 Mayo Building MMC 303
Minneapolis, MN 55455
USA
E-mail address: mcbea002@umn.edu

Alexander M. McBean
420 Delaware Street S.E.
A369-1 Mayo Building
Minneapolis, MN 55455
USA
E-mail address: mcbea002@umn.edu

Bradley P. Carlin
420 Delaware Street S.E.
A369-1 Mayo Building
Minneapolis, MN 55455
USA
E-mail address: carli002@umn.edu