Biostatistics (2012), **13**, 3, *pp.* 415–426
doi:10.1093/biostatistics/kxr036
Advance Access publication on October 31, 2011

Boundary detection in disease mapping studies

DUNCAN LEE*
School of Mathematics and Statistics,University of Glasgow, Glasgow, UK
Duncan.Lee@glasgow.ac.uk

RICHARD MITCHELL
Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK

Summary

In disease mapping, the aim is to estimate the spatial pattern in disease risk over an extended geographical region, so that areas with elevated risks can be identified. A Bayesian hierarchical approach is typically used to produce such maps, which represents the risk surface with a set of random effects that exhibit a single global level of spatial smoothness. However, in complex urban settings, the risk surface is likely to exhibit localized rather than global spatial structure, including areas where the risk varies smoothly over space, as well as boundaries separating populations that are geographically adjacent but have very different risk profiles. Therefore, this paper proposes an approach for capturing localized spatial structure, including the identification of such risk boundaries. The effectiveness of the approach is tested by simulation, before being applied to lung cancer incidence data in Greater Glasgow, UK, between 2001 and 2005.

Keywords: Boundary detection; Conditional autoregressive models; Disease mapping; Spatial correlation.

1. Introduction

In disease mapping, the aim is to quantify the spatial pattern in disease risk over an extended geographical region, which is partitioned into small nonoverlapping areal units, such as electoral wards. Bayesian hierarchical models are typically used to produce such maps, where the spatial pattern in disease risk is represented by a set of random effects. These random effects are often assigned by a conditional autoregressive (CAR) prior (see, e.g. Elliott *and others*, 2000; Wakefield, 2007; Lee, 2011), which forces them to exhibit a single global level of spatial smoothness. However, in complex urban settings, the risk of disease is likely to exhibit more localized spatial structure, including areas where it varies smoothly over space, as well as locations where there are sharp changes in its value. The latter correspond to "boundaries" (or "cliffs" or "discontinuities") in the risk surface and may occur where rich and poor communities live side by side. In this context, one might expect to observe similar (correlated) disease risks within each community but not at the border where they meet. The identification of risk boundaries has a number of benefits, including the ability to detect the spatial extent of a cluster of high-risk areas. Such cluster detection is important economically because it allows health resources to be targeted at communities with the highest disease risks. Furthermore, the locations of any boundaries may coincide with borders between different neighborhoods, that is, groups of people with the same social circumstances, culture, and behavior. Such

*To whom correspondence should be addressed.

neighborhood boundaries are of interest because "their locations reflect underlying biological, physical, and/or social processes" (Jacquez *and others*, 2000).

However, relatively few disease mapping studies have attempted to identify boundaries in the risk surface, with the majority of studies focusing on quantifying the effects of covariates using ecological regression. One of the earliest solutions to this problem was proposed by Womble (1951), with more recent work being based on local statistics (Boots, 2001), mixture models (Knorr-Held and Rasser, 2000), and CAR models (Lu and Carlin, 2005; Lu *and others*, 2007). The approach we propose here is also based on CAR models but has the duel advantages over existing work of being parsimonious and not requiring the user to specify a tuning constant. The rationale for our approach is that risk boundaries are likely to occur between populations living side by side that are very different because homogeneous populations should exhibit similar disease risks. The remainder of this paper is organized as follows. Section 2 provides a review of disease mapping and boundary detection methods, while Section 3 presents our proposed methodology. Section 4 assesses the efficacy of our approach using simulation, while Section 5 identifies boundaries in the risk surface for lung cancer cases in Greater Glasgow. Finally, Section 6 contains a concluding discussion and outlines future developments.

2. BACKGROUND

2.1 *Disease mapping and ecological regression with covariates*

The data used to quantify disease risk are denoted by $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{E} = (E_1, \ldots, E_n)$, the former being the numbers of disease cases observed in each of n nonoverlapping areal units within a specified time frame. The latter are the numbers of disease cases expected to occur in each area, which depend on the size and demographic structure of the populations living there. The simplest measure of disease risk is the standardized incidence ratio, which for area k is given by $\hat{R}_k = y_k/E_k$. However, to estimate the effects of covariate factors on disease risk, a Bayesian hierarchical model is typically used. A general specification is given by

$$\begin{aligned} Y_k|E_k, R_k &\sim \text{Poisson}(E_k R_k) \quad \text{for } k = 1, \ldots, n, \\ \ln(R_k) &= \mathbf{x}_k^{\mathrm{T}}\beta + \phi_k, \end{aligned} \tag{2.1}$$

and has been described in detail by Elliott *and others* (2000), Wakefield (2007), and Lawson (2008). Disease risk in area k is denoted by R_k and is represented by covariates $\mathbf{x}_k^{\mathrm{T}} = (x_{k1}, \ldots, x_{kp})$ and random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$, the latter allowing for any overdispersion and spatial correlation in the disease data after the covariate effects have been accounted for. The most common prior for $\boldsymbol{\phi}$ is a CAR (Besag *and others*, 1991) model, which induces spatial correlation into the random effects via a binary $n \times n$ neighborhood matrix W. Element w_{kj} of this matrix is equal to one if areas (k, j) share a common border and is zero otherwise. A number of priors have been proposed within the general class of CAR models, and the one we adopt here was originally proposed by Leroux *and others* (1999). It is specified via the following n univariate full conditional distributions:

$$\phi_k|\boldsymbol{\phi}_{-k}, W, \tau^2, \rho, \mu \sim \mathbf{N}\left(\frac{\rho\sum_{j=1}^{n} w_{kj}\phi_j + (1-\rho)\mu}{\rho\sum_{j=1}^{n} w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{kj} + 1 - \rho}\right). \tag{2.2}$$

The conditional expectation has an attractive mixture representation, comprising spatial smoothing to the neighboring random effects (with weight ρ) and nonspatial smoothing to a global mean μ (with

weight $(1 - \rho)$). The single parameter ρ determines the global level of spatial correlation between the random effects, with $\rho = 0$ corresponding to independence everywhere, while ρ close to 1 defines strong spatial correlation throughout the region. The conditional variance is equally attractive, being constant if the random effects are independent ($\rho = 0$), and inversely proportional to the number of neighboring areas (measured by $\sum_{j=1}^{n} w_{kj}$ for area k) if the random effects are highly correlated ($\rho = 1$). If $\rho \in [0, 1)$, (2.2) corresponds to a proper multivariate Gaussian distribution for $\boldsymbol{\phi}$, which is given by $\boldsymbol{\phi} \sim \mathrm{N}(\mu \mathbf{1},\ \tau^2[\rho W^* + (1 - \rho)I]^{-1})$. Here, I denotes an $n \times n$ identity matrix, $\mathbf{1}$ is an n-vector of ones, while W^* has diagonal elements $w^*_{kk} = \sum_{i=1}^{n} w_{ki}$ and nondiagonal elements $w^*_{kj} = -w_{kj}$.

2.2 *Boundary detection*

Two main approaches have been proposed for identifying boundaries in disease maps using CAR models. The first was proposed by Lu and Carlin (2005), who used the model outlined by (2.1) and (2.2) except that ρ is fixed at one. They identify boundaries by calculating $|\hat{R}_k - \hat{R}_j|$ for all pairs of neighboring areas, where $\hat{R}_k$ is an estimate of R_k such as a posterior median. These quantities are called boundary likelihood values (BLV), and the border between neighboring areas (k, j) can be classified as a boundary if its BLV is either bigger than some constant c_1 or is in the top $c_2\%$ of all BLVs. However, Jacquez *and others* (2000) have criticized this approach because (c_1 or c_2) need to be specified by the investigator, who in doing so essentially chooses the number of boundaries that are identified, even though this is unknown and the goal of the analysis.

An alternative approach was proposed by Lu *and others* (2007), who again fix ρ at one but also remove the covariate component $\mathbf{x}_k^{\mathrm{T}}\beta$. The second of these changes means that the spatial structure among the random effects and the disease risks are identical, as one is the logarithm of the other. They model the set of $\{w_{kj}\}$ as binary random quantities if areas (k, j) share a common border, rather than assuming they are fixed at one. If $w_{kj} = 0$, then (ϕ_k, ϕ_j) are conditionally independent, corresponding to a boundary in the risk surface, whereas $w_{kj} = 1$ corresponds to correlated random effects and no risk boundary. Lu *and others* (2007) model the set of $\{w_{kj}\}$ using a logistic regression model, while Ma and Carlin (2007) and Ma *and others* (2010) propose similar models using a CAR prior and an Ising model, respectively. However, modeling each w_{kj} as a separate random quantity results in a highly overparameterized covariance model for $\boldsymbol{\phi}$, which Li *and others* (2011) argue is not well identified from the data and requires informative prior information.

3. Methods

Following the general approach of Lu *and others* (2007), we propose capturing localized spatial correlation and identifying risk boundaries by modeling the set of $\{w_{kj}\}$ for geographically adjacent areas as binary random quantities. However, unlike their model, we propose determining the set of $\{w_{kj}\}$ using a small number of regression parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$, rather than treating each w_{kj} as a separate unknown quantity. This leads to a parsimonious covariance structure for $\boldsymbol{\phi}$, whist still allowing the number and locations of any boundaries to be identified by the data, rather than a priori by the investigator. Inference for this model is based on Markov Chain Monte Carlo (MCMC) simulation, and a description of the algorithm is given in the supplementary material available at *Biostatistics* online accompanying this paper. In addition, a function to run the model in the statistical package `R` (R Development Core Team, 2009) as well as the data set analyzed in Section 5 are also provided in the supplementary material available at *Biostatistics* online.

3.1 *Level 1: observation model*

The first level model is similar to that described in Section 2 and is given by

$$\begin{aligned} Y_k|E_k, R_k &\sim \text{Poisson}(E_k R_k) \quad \text{for } k = 1, \ldots, n, \\ \ln(R_k) &= \phi_k, \\ \phi_k|\boldsymbol{\phi}_{-k}, \mu, \boldsymbol{\alpha}, \tau^2 &\sim \mathbf{N}\left(\frac{0.99 \sum_{j=1}^{n} w_{kj}(\boldsymbol{\alpha})\phi_j + 0.01\mu}{0.99 \sum_{j=1}^{n} w_{kj}(\boldsymbol{\alpha}) + 0.01}, \frac{\tau^2}{0.99 \sum_{j=1}^{n} w_{kj}(\boldsymbol{\alpha}) + 0.01} \right). \end{aligned} \tag{3.1}$$

No covariates are included in this model, so that the spatial structure in the random effects surface and the risk surface is identical. In addition, ρ is fixed at 0.99, so that the spatial correlation structure can be determined locally by $\{w_{kj}(\boldsymbol{\alpha})\}$ rather than globally by ρ (if ρ was fixed at 0, $\{w_{kj}(\boldsymbol{\alpha})\}$ would disappear from the model). A value of one is not chosen because then (2.2) would have an infinite mean and variance if area k is surrounded by boundaries, that is, if $\sum_{j=1}^{n} w_{kj}(\boldsymbol{\alpha}) = 0$. This model corresponds to a proper multivariate Gaussian distribution for $\boldsymbol{\phi}$, which is given by $\boldsymbol{\phi} \sim \text{N}(\mu\mathbf{1},\ \tau^2 Q(\boldsymbol{\alpha})^{-1})$, where $Q(\boldsymbol{\alpha}) = [0.99W^*(\boldsymbol{\alpha}) + 0.01I]$.

3.2 *Level 2: neighborhood model*

We believe that boundaries in the risk surface are likely to occur between populations that are very different because homogeneous populations should have similar risk profiles. Therefore, we model the presence or absence of a boundary in the risk surface between geographically adjacent areas (k, j) using q nonnegative dissimilarity metrics $\mathbf{z}_{kj} = (z_{kj1}, \ldots, z_{kjq})$, where $z_{kji} = |z_{ki} - z_{ji}|/\sigma_i$, for $i = 1, \ldots, q$. Each z_{kji} measures the absolute difference in the value of a covariate between the 2 areas in question, and the rescaling by σ_i, the standard deviation (SD) of $|z_{ki} - z_{ji}|$ over all pairs of contiguous areas, improves the mixing and convergence of the MCMC algorithm. It is these dissimilarity metrics that drive the detection of boundaries in the risk surface, and examples could include differences in 2 populations social characteristics (e.g. average income) or risk-inducing behavior (e.g. smoking prevalence). Using these metrics, our proposed neighborhood model is

$$w_{kj}(\boldsymbol{\alpha}) = \begin{cases} 1 & \text{if } \exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right) \geqslant 0.5 \text{ and } j \sim k \\ 0 & \text{otherwise} \end{cases}, \tag{3.2}$$

where pairs of areas that do not share a common border have $w_{kj}(\boldsymbol{\alpha})$ fixed at zero. For areas that are contiguous, the model detects a boundary in the risk surface if $\exp(-\sum_{i=1}^{q} z_{kji}\alpha_i)$ is less than 0.5, while a value greater than or equal to 0.5 corresponds to no boundary. The regression parameters, $\boldsymbol{\alpha}$, are constrained to be nonnegative, so that the greater the dissimilarity between 2 areas the more likely there is to be a boundary between them. In addition, there is no intercept term in (3.2), so that 2 areas with homogeneous populations (i.e. all $z_{kji} = 0$) cannot have a boundary between them.

If there is only one dissimilarity metric in the model, then no boundaries will be detected if the sole regression parameter $\alpha \leqslant -\ln(0.5)/z^{\max}$, while all borders in the study region will be identified as boundaries (unless $z_{kj} = 0$) if $\alpha > -\ln(0.5)/z^{\min}$. Here, $(z^{\min}, z^{\max})$ denote the minimum positive and maximum values of the dissimilarity metric. More generally, if there are q dissimilarity metrics,

then boundaries are identified if $\prod_{i=1}^{q} \exp(-z_{kji}\alpha_i) < 0.5$. In this case, if $\alpha_i \leqslant -\ln(0.5)/z_i^{\max}$, then z_{kji} is not solely responsible for detecting any boundaries because the component $\exp(-z_{kji}\alpha_i)$ would be greater than 0.5 for all pairs of contiguous areas. Thus, if the 95% credible interval for α_i lies below $\alpha_i^{\min} = -\ln(0.5)/z_i^{\max}$, then the dissimilarity metric can be said to have no effect on detecting boundaries. In contrast, if the uncertainty interval lies completely above $\alpha_i^{\min}$, then the metric can be said to have a substantial effect on identifying risk boundaries. We note that our approach does not guarantee that the boundaries we detect will be closed (form an unbroken line), which allows us to detect boundaries that enclose an entire subregion, as well as those that just separate highly different areas.

3.3 *Level 3: Hyperpriors*

Finally, we require hyperpriors for $(\mu, \tau^2, \boldsymbol{\alpha})$, which respectively control the mean, variance, and correlation structure of $\boldsymbol{\phi}$. Weakly informative priors are specified for (μ, τ), in the form of Gaussian (μ, mean 0 and variance 10) and uniform (τ, limits of 0 and 10) distributions. We propose a noninformative uniform prior for each regression parameter, $\alpha_i \sim \text{Uniform}(0, M_i)$, which corresponds to our prior ignorance about the number of boundaries in the risk surface. An alternative would be a reciprocal prior, $f(\alpha_i) \propto \frac{1}{\alpha_i} I[0 \leqslant \alpha_i \leqslant M_i]$, which represents our prior belief that the risk surface is spatially smooth. In both cases, a natural upper limit would be $M_i = -\ln(0.5)/z_i^{\min}$, the value at which the dissimilarity measure solely identifies all borders as boundaries in the risk surface. However, we believe that large parts of the risk surface will be spatially smooth (otherwise, we would use an independence model for $\boldsymbol{\phi}$), so we fix M_i, so that at most 50% of borders in the study region can be classified as boundaries. A sensitivity analysis to this choice is presented in the supplementary material available at *Biostatistics* online.

4. Simulation study

This section presents a simulation study that assesses whether the model proposed in Section 3 can detect "true" boundaries in the risk surface, as well as the extent to which it falsely identifies boundaries that do not exist.

4.1 *Data generation*

The study is based on the $n = 271$ intermediate geographies (IG) that comprise the Greater Glasgow and Clyde health board, which is the region considered in the cancer mapping study presented in Section 5. Disease counts are generated from the Poisson model (3.1), where the expected numbers of admissions, $\mathbf{E}$, relate to the Glasgow cancer data. A new risk surface $\mathbf{R} = \exp(\boldsymbol{\phi})$ is generated for each set of simulated disease data because this ensures that the results are not affected by a particular realization of $\boldsymbol{\phi}$. Each simulated risk surface has fixed boundaries, which are shown by the bold black lines in Figure 1. There are 74 boundaries in total, which corresponds to approximately 10% of the set of borders in the study region. This set of boundaries partition the study region into 6 groups, the main area shaded in white and the remaining 5 smaller areas shaded in gray. To produce risk surfaces with boundaries, the random effects ($\boldsymbol{\phi}$) are generated from a multivariate Gaussian distribution with a piecewise constant mean, which in the white region is equal to zero, while in the gray regions it is equal to k_1. The correlation function is from the Matern class with smoothness parameter equal to $\kappa = 2.5$, while the spatial range is fixed, so that the median correlation between areas is 0.5. The dissimilarity metrics z_{kj} are generated from $|\text{N}(1, 0.5^2)|$ if areas (k, j) are not separated by a boundary in Figure 1 and from $|\text{N}(1 + k_2, 0.5^2)|$ if they are separated by a boundary. Therefore, larger values of k_2 correspond to dissimilarity metrics that better identify the true boundaries in the risk surface; that is, have larger values for the boundaries in Figure 1 than for the nonboundaries.

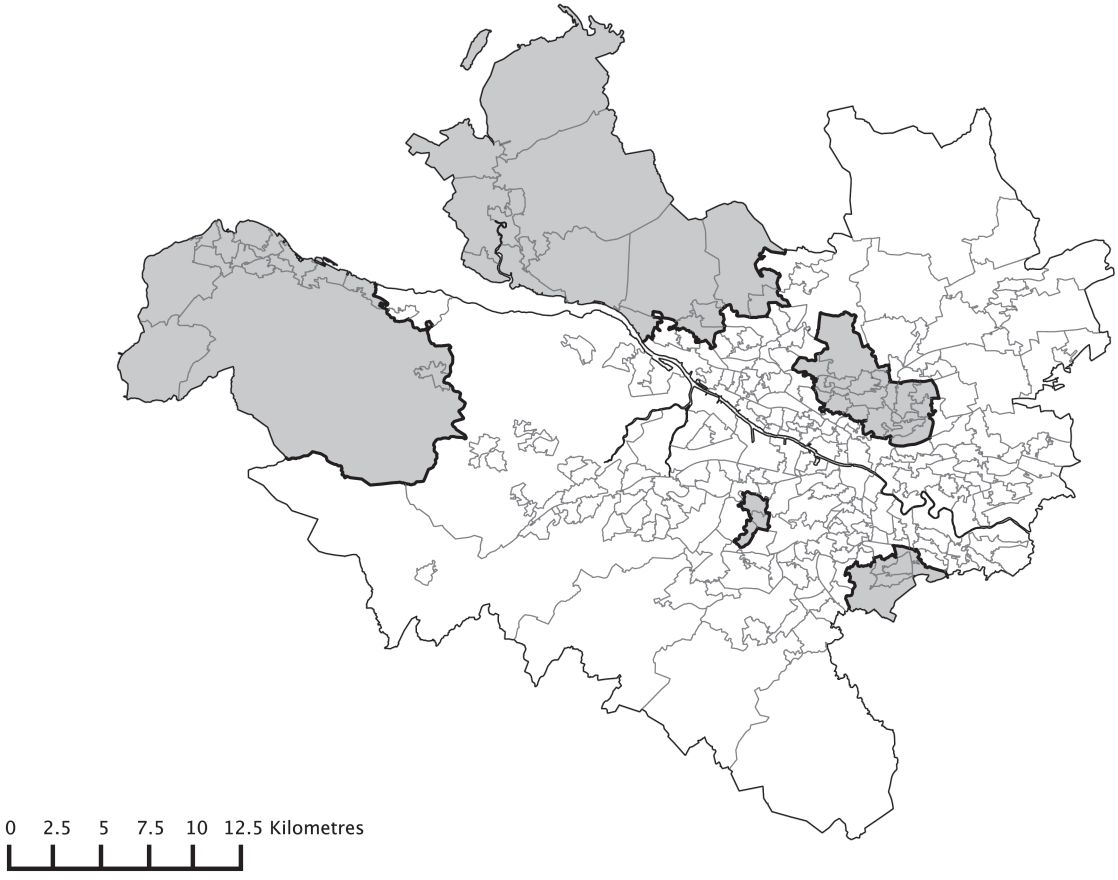

Fig. 1. Locations of the true boundaries in the simulated risk surfaces.

4.2 *Results*

The results are displayed in Table 1 and relate to 200 simulated data sets for each combination of (k_1, k_2). In all cases, the model produces a good representation of the risk surface, with bias and root-mean-square error being less than 0.3% and 7.2% of the true values. Panel A shows the impact of boundary size on model performance, in the idealized situation of having perfect dissimilarity metrics ($k_2 = 3$ corresponds to the mean values of the dissimilarity metrics for boundaries and nonboundaries being separated by 6 SDs). The table shows that the model can detect larger risk boundaries (larger values of k_1) more often than smaller ones as expected (BA column in the table), although it still achieves over a 90% detection rate when the true boundaries only represent a mean difference in risk of 0.2. The much lower detection rates for smaller values of k_1 are also not surprising, as they correspond to risk surfaces that do not exhibit large discontinuities at the boundary locations. In addition, the false positive rates (the percentage of nonboundaries incorrectly identified as boundaries) are very low regardless of the boundary size (generally less than 1%), which suggests that detected boundaries are likely to be real.

Panel B in Table 1 displays the effects of having imperfect dissimilarity metrics, in the situation where the risk boundaries are relatively large ($k_1 = 0.4$). The table shows that when k_2 is at least 2, the model

Table 1. *The effect of boundary size (as measured by k_1) and the quality of the dissimilarity metrics (as measured by k_2) on the effectiveness of the model. The table displays the percentage of true boundaries correctly detected by the model (BA), and the percentage of nonboundaries correctly not detected as boundaries (NBA). It also displays the bias and root-mean-square error (RMSE) of the estimated risk surface, which are presented as a percentage of their true value*

Panel	k_1	k_2	BA (%)	NBA (%)	Bias	RMSE
A	0.4	3	99.97	98.70	−0.123	5.689
	0.3	3	99.57	99.16	−0.195	5.700
	0.2	3	93.76	99.43	−0.187	5.689
	0.1	3	48.31	99.89	−0.092	5.706
	0.05	3	25.84	100	−0.139	5.663
B	0.4	3	99.97	98.70	−0.123	5.689
	0.4	2	98.89	95.07	−0.134	5.747
	0.4	1.5	96.27	88.37	−0.140	5.866
	0.4	1	87.19	80.85	−0.206	6.226
	0.4	0.5	55.74	80.93	−0.242	6.927
	0.4	0	1.85	98.82	−0.248	7.178
C	0.2	1.5	85.60	95.88	−0.108	5.792
	0.2	0	1.54	98.87	−0.249	6.282
	0.05	1.5	21.97	99.72	−0.194	5.797
	0.05	0	1.05	99.12	−0.161	5.747

performs well, exhibiting over 95% accuracy in detecting both boundaries and nonboundaries. However, as k_2 decreases, the information content in the dissimilarity metrics also decrease, as there is a greater overlap between the z_{kj} at boundary and nonboundary locations. As a result, model performance reduces both in terms of the boundary and the nonboundary agreement. In the extreme case that the dissimilarity metric contains no information about the locations of the boundaries ($k_2 = 0$), the model only identifies 1.86% of the true boundaries, although in this situation, the false positive rate returns to being low at around 1%.

The top line of panel C shows that when $k_1 = 0.2$ and $k_2 = 1.5$, model performance remains good, with 85.6% and 95.9% accuracy in detecting boundaries and nonboundaries, respectively. However, if either the risk boundaries are small ($k_1 = 0.05$) or the dissimilarity metrics contain no information ($k_2 = 0$), then the model is largely unable to detect risk boundaries, with detection rates ranging between 1.1% and 22.0%. In this setting, however, the false positive rates remain very low, as less than 1.2% of nonboundaries are false identified as risk boundaries. Finally, the sensitivity of these results were assessed by changing a number of aspects of the study, and to save space, the details are presented in the supplementary material available at *Biostatistics* online.

5. Case study—cancer risk in Greater Glasgow

We illustrate our methods by presenting a study mapping lung cancer risk in Greater Glasgow, UK, between 2001 and 2005. We apply both the ecological regression model (Section 2.1) and the boundary detection model (Section 3) to our data, because they are complementary, and address fundamentally different research questions.

5.1 *Data description*

The study region is the Greater Glasgow and Clyde health board, which is partitioned into $n = 271$ administrative units called IG. Glasgow is an ideal region for this study because it has rich and poor communities that are geographically adjacent. All the data used are freely available and can be downloaded from the Scottish Neighbourhood Statistics database (http://www.sns.gov.uk). The disease data are the numbers of people diagnosed with lung cancer between 2001 and 2005 in each IG, which corresponds to ICD-10 codes C33 and C34. The expected numbers of cases are calculated by external standardization, using age- and sex-adjusted rates for the whole of Scotland. The first covariate we consider is a modeled estimate of the percentage of the population in each IG that smoke (for details, see Whyte *and others*, 2007), which is included because the causal relationship between smoking and lung cancer risk is well known (Doll *and others*, 2005). Cancer risk also varies by ethnic group (National Cancer Intelligence Network, 2009) and socioeconomic status (Woods *and others*, 2006), which are represented by the percentage of school children from ethnic minorities and the natural log of the median house price, respectively. The first of these is the only proxy measure of ethnicity available, while the second is the measure of deprivation available least correlated with the smoking covariate (Pearson's $r = -0.69$).

5.2 *Modeling*

Inference for all models is based on 50 000 MCMC samples generated from 5 Markov chains, which were each burnt-in until convergence (40 000 iterations). Both the ecological regression and boundary detection models were applied to the data, using the 3 covariates described in the previous section. The residuals from each model were assessed for the presence of spatial correlation using a Moran's I permutation test, and in all cases, the p values suggested no evidence of residual spatial structure.

For the ecological regression model, all 3 covariates had substantial impacts on the response and were hence all retained in the model. In contrast, in the boundary detection model, only smoking prevalence had any effect in identifying risk boundaries, which can be seen from the bottom half of Table 2. The table displays posterior medians and 95% credible intervals for each α_i, as well as the threshold values $\alpha_i^{\min}$, below which each dissimilarity metric does not solely detect any boundaries in the risk surface. For smoking prevalence, the estimate and credible interval lie above the threshold value of $\alpha_i^{\min} = 0.131$, while for the remaining dissimilarity metrics, the credible intervals lie below the "no effect" thresholds. The deviance information criterion (DIC) values are 1711.2 and 1753.5 respectively for the ecological regression and boundary detection models, which suggests that the former is the better fit to the data. This result is not surprising, as the ecological regression model is using the covariates in the mean structure to provide a better description of the data, while the boundary detection model is using them to identify boundaries via the covariance structure.

5.3 *Results*

The effects of the covariates on lung cancer risk from the ecological regression model are shown in the top half of Table 2, which displays posterior medians and 95% credible intervals. All results are presented on the relative risk scale, for a one SD increase in each covariates value. There is strong evidence that increasing the percentage of the population in each IG that smoke is related to increased lung cancer risk, with a relative risk of 1.316 for a 9.6% increase in the smoking prevalence. Populations living in areas that are less deprived (as measured by the log of average house price) are also less at risk of cancer, with a relative risk of 0.929 for a one SD increase in house price. Finally, the evidence for an effect of ethnicity on lung cancer risk is marginal, as the 95% credible interval just contains the null risk of one.

The top half of Figure 2 displays the spatial pattern in lung cancer risk from the ecological regression model, which shows that the highest risks are in the heavily deprived east end of Glasgow as well as along

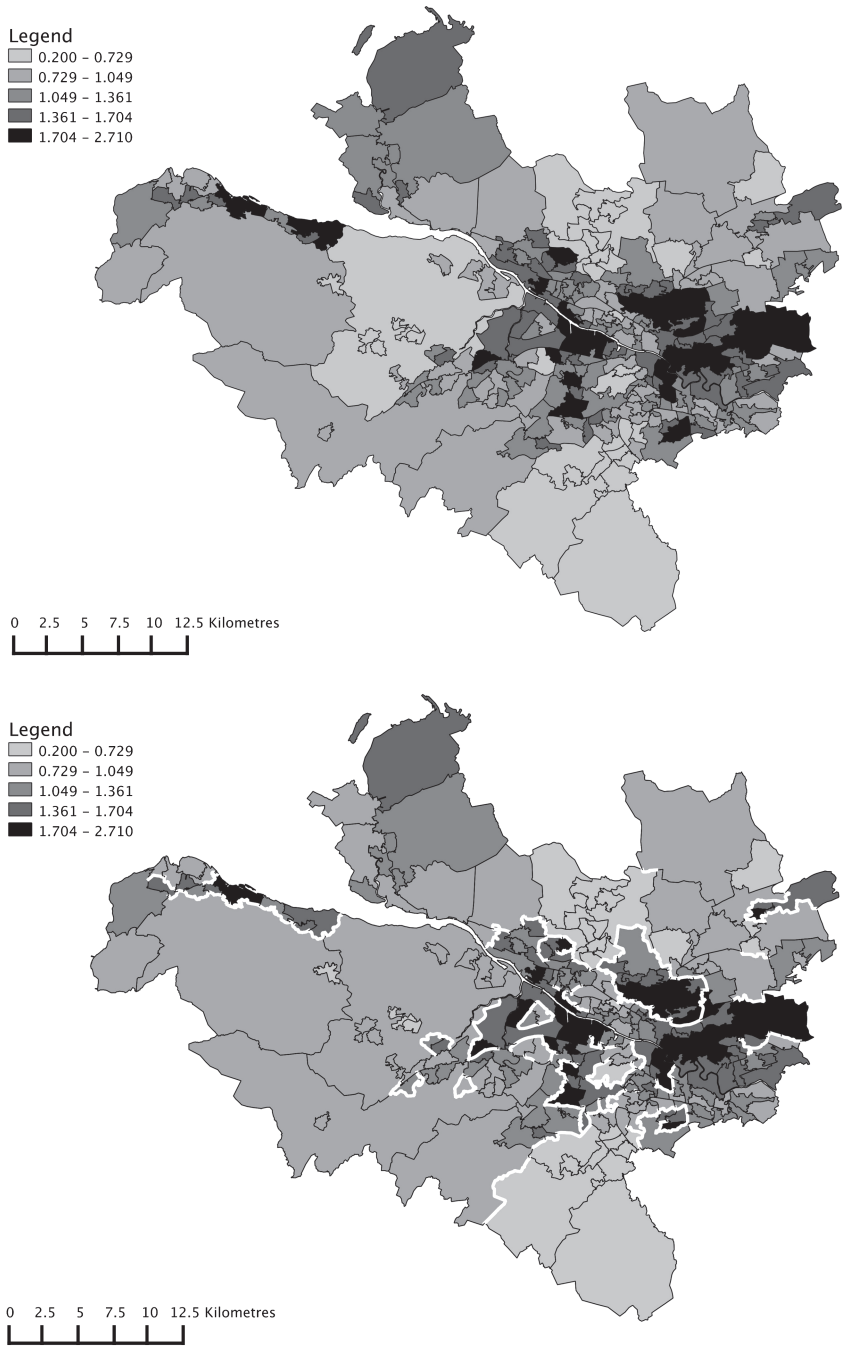

Fig. 2. Estimated risk surface for lung cancer in Greater Glasgow between 2001 and 2005. The top panel displays the results from the ecological regression model, while the bottom panel displays the results from the boundary detection model. In the latter, the risk boundaries are denoted by solid white lines.

Table 2. *Covariate effects from the ecological regression model (top) and the boundary detection model (bottom). Both sets of results are presented as posterior medians and* 95% *credible intervals. The results in the top panel are relative risks for a one SD (shown below) increase in each covariates value. In the bottom panel, the threshold values* $\alpha_i^{\min} = -\ln(0.5)/z_i^{\max}$ *are also presented*

Ecological regression model			
Covariate	Relative risk (β_i)	95% credible interval	SD
Smoking	1.316	(1.249–1.397)	9.637
Ethnicity	0.960	(0.914–1.003)	12.25
House price (log)	0.929	(0.886–0.978)	0.509
Boundary detection model			
Covariate	Estimate (α_i)	95% credible interval	$\alpha_i^{\min}$
Smoking	0.232	(0.171–0.254)	0.131
Ethnicity	0.012	(0.001–0.046)	0.126
House price (log)	0.015	(0.001–0.103)	0.119

the banks of the river Clyde. The risk surface also contains substantial global correlation, as the posterior median for the correlation parameter ρ is 0.53, with a 95% credible interval between 0.13 and 0.90. The bottom half of Figure 2 displays the risk surface estimated from the boundary detection model, where smoking was the only dissimilarity metric used. The 2 risk surfaces exhibit very similar spatial patterns, with a correlation coefficient of 0.97. The absolute difference between the risks estimated from both models is also small relative to the spatial variation in risk, with a median absolute difference of 0.058.

The boundary detection model has the additional benefit of identifying risk boundaries, which correspond to those w_{kj} that have a posterior median of zero. The risk boundaries are presented as solid white lines, and the smoking covariate has identified 162 boundaries in total, which corresponds to 23.1% of all possible borders in the study region. We note that areas on opposite banks of the river Clyde (the thin white line running south east) are not assumed to be neighbors, which explains the absence of boundaries in this area. The majority of these estimated risk boundaries appear to correspond to sizeable changes in the risk surface, suggesting that the smoking covariate appears to be an appropriate dissimilarity metric for detecting such boundaries. However, a few of the boundaries identified show no evidence of separating areas with differing health risks, such as the closed boundary in the south of the city. These "false positives" correspond to 2 areas having different smoking prevalences but similar risk profiles and would be a starting point for a more detailed investigation into why the risk profiles are similar given the vastly different smoking rates. Finally, we assess the sensitivity of the boundary detection model by applying it to the data with different combinations of covariates, including all 3, smoking on its own and smoking with each of the remaining covariates separately. In all cases, both the DIC and the number of boundaries detected remained largely unchanged, reenforcing the suggestion that smoking is the only important covariate for detecting risk boundaries.

6. DISCUSSION

This paper has proposed a novel Bayesian hierarchical model for representing localized spatial structure and identifying boundaries in disease risk maps, by extending the class of CAR prior distributions. This approach detects boundaries by measuring the dissimilarity between populations living in neighboring areas because we believe that abrupt changes in the risk surface are most likely to occur between populations that are geographically adjacent but have very different social characteristics or risk-inducing

behavior. Our model has the dual advantages of being fully automatic and parsimonious, unlike the existing approaches proposed by Lu and Carlin (2005) and Lu *and others* (2007). We view our model as complementary to the ecological regression model that is commonly used in the disease mapping literature, as the 2 models address fundamentally different research questions. The ecological regression model aims to describe the spatial variation in disease risk using covariates, while the model proposed here attempts to use covariate information (termed dissimilarity metrics) to identify the locations of any boundaries in the risk surface.

The simulation study presented in Section 4 shows that our model generally performs well, both in terms of detecting true boundaries in the risk surface as well as not detecting large numbers of false positives. The false positive rate is particularly low (generally less than 2%), which suggests that the model is conservative in the number of boundaries that it identifies. The main drawback of our approach is illustrated by panel B in Table 1, which shows that it is crucially dependent on the existence of good quality dissimilarity metrics that have larger values for true boundaries than for nonboundaries. However, even in the worst-case scenario of having poor quality dissimilarity metrics, the model produces less than 20% false positives, which compares favorably with the results presented by Lu *and others* (2007, table 2).

The Glasgow lung cancer example gives insight into the model's performance for real data, which invariably contains more complex structure than is present in idealized simulated data. Figure 2 shows that smoking prevalence (one of the major causes of lung cancer) identifies the majority of the boundaries evident from the risk map, suggesting that it is an appropriate dissimilarity metric to use for this example. However, the use of this covariate also results in a small number of boundaries being identified that do not correspond to sizeable discontinuities in the risk surface. This imperfection could be caused by the omission of other important dissimilarity metrics, or by the fact that the smoking covariate is a modeled estimate rather than being real prevalence data. Therefore, one could view the model proposed here as an exploratory tool, which aids the investigator in understanding the localized spatial nature of disease risk. For example, if a risk boundary is not detected despite their being an apparent discontinuity in the risk surface, further investigation of the 2 areas in question could be carried out to determine what other factors could be causing this discontinuity.

Finally, this paper opens up numerous avenues for future work. The most obvious is to develop a boundary detection method that has the advantages of the method proposed here but does not rely on covariate data to detect risk boundaries. One possibility in this vein is to develop an iterative approach that compares the fit of a number of models with different but fixed neighborhood matrices W. In this way, one could compare a model where all adjacent areas have $w_{kj} = 1$, against an alternative specification where $w_{kj} = 0$ for areas that are suspected of being separated by a risk boundary. Other natural extensions to this approach include the detection of boundaries in multiple diseases simultaneously, as well as adding a temporal dimension to the model.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the valuable comments and suggestions made by an anonymous reviewer, which have improved the focus and content of this paper. The data and shapefiles used in this study were provided by the Scottish Government. *Conflict of Interest:* None declared.

FUNDING

The Economic and Social Research Council (RES-000-22-4256).

REFERENCES

BESAG, J., YORK, J. AND MOLLIE, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics* **43**, 1–59.

BOOTS, B. (2001). Using local statistics for boundary characterization. *GeoJournal* **53**, 339–345.

DOLL, R., PETO, R., BOREHAM, J. AND SUTHERLAND, I. (2005). Mortality from cancer in relation to smoking: 50 years observations on British doctors. *British Journal of Cancer* **92**, 426–429.

ELLIOTT, P., WAKEFIELD, J., BEST, N. AND BRIGGS, D. (2000). *Spatial Epidemiology: Methods and Applications*, 1st edition. Oxford: Oxford University Press.

JACQUEZ, G., MARUCA, S. AND FORTIN, M. (2000). From fields to objects: a review of geographic boundary analysis. *Journal of Geographical Systems* **2**, 221–241.

KNORR-HELD, L. AND RASSER, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**, 13–21.

LAWSON, A. (2008). *Bayesian Disease Mapping: Hierarchical Modelling in Spatial Epidemiology*, 1st edition. Boca Raton, FL: Chapman and Hall.

LEE, D. (2011). A comparison of conditional autoregressive model used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology* **2**, 79–89.

LEROUX, B., LEI, X. AND BRESLOW, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In: Halloran, M. and Berry, D (editors), *Chapter Statistical Models in Epidemiology, the Environment and Clinical Trials*. New York: Springer, pp. 135–178.

LI, P., BANERJEE, S. AND MCBEAN, A. (2011). Mining boundary effects in areally referenced spatial data using the Bayesian information criterion. *Geoinformatica* **15**, 435–454.

LU, H. AND CARLIN, B. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis* **37**, 265–285.

LU, H., REILLY, C., BANERJEE, S. AND CARLIN, B. (2007). Bayesian areal wombling via adjacency modelling. *Environmental and Ecological Statistics* **14**, 433–452.

MA, H. AND CARLIN, B. (2007). Bayesian multivariate areal wombling for multiple disease boundary analysis. *Bayesian Analysis* **2**, 281–302.

MA, H., CARLIN, B. AND BANERJEE, S. (2010). Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. *Biometrics* **66**, 355–364.

NATIONAL CANCER INTELLIGENCE NETWORK (2009). Cancer incidence and survival by major ethnic group, England, 2002–2006. *Technical Report*. Cancer Research UK Cancer Survival Group and London School of Hygiene and Tropical Medicine.

R DEVELOPMENT CORE TEAM (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

WAKEFIELD, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics* **8**, 158–183.

WHYTE, B., GORDON, D., HAW, S., FISCHBACHER, C. AND HARRISON, R. (2007). *An Atlas of Tobacco Smoking in Scotland: A Report Presenting Estimated Smoking Prevalence and Smoking Attributable Deaths within Scotland*. Edinburgh, UK: NHS Health Scotland.

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

WOODS, L., RACHET, B. AND COLEMAN, M. (2006). Origins of socio-economic inequalities in cancer survival: a review. *Annals of Oncology* **17**, 5–19.

[*Received May 16, 2011; revised August 11, 2011; accepted for publication September 21, 2011*]