An extended conditional autoregressive model for Bayesian disease mapping

Lee, Duncan

University of Glasgow, School of Mathematics and Statistics
15 University Gardens
Glasgow, UK
E-mail: Duncan.Lee@glasgow.ac.uk

Mitchell, Richard

University of Glasgow, School of Medicine
1 Lilybank Gardens
Glasgow, UK
E-mail: Richard.Mitchell@glasgow.ac.uk

1. Introduction and motivation

Disease mapping is the area of statistics that quantifies the spatial pattern in disease risk over an extended geographical region, such as a city or country. Ideally, every disease case or mortality event would be available with an exact geographical location, but due to patient confidentiality this is often not the case. Instead, the study region is partitioned into a number of small non-overlapping areal units, and only the total number of cases in each unit is available. These data are denoted by $y = (y_1, \ldots, y_n)$, and disease risk is quantified by comparing them to $E = (E_1, \ldots, E_n)$, the expected number of cases given the size and demographic structure of the population in each area. The simplest measure of disease risk is the standardised incidence ratio (SIR), which for area $k$ is given by

$$\hat{R}_k = \frac{y_k}{E_k}.$$  

Values above one represent areas with elevated risks of disease, while values less than one correspond to relatively healthy areas. However, elevated SIR values can occur by chance in areas where $E_k$ is small, as only a few additional cases are required to greatly inflate the risk. Therefore the set of disease risks for all $n$ areas are typically estimated using a Bayesian hierarchical model (see for example Lawson (2008)) and a general form is given by

$$Y_k|E_k, R_k \sim \text{Poisson}(E_k R_k) \quad \text{for} \quad k = 1, \ldots, n,$$

$$\ln(R_k) = \phi_k.$$  

(1)

The disease risks for all $n$ areas are modelled by a set of random effects $\phi = (\phi_1, \ldots, \phi_n)$, which model any spatial correlation and overdispersion in the data. The most common prior for $\phi$ is a conditional autoregressive (CAR, Besag et al. (1991)) model, which is specified through $n$ univariate full conditional distributions $f(\phi_k|\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_n)$ for $k = 1, \ldots, n$. The spatial correlation structure amongst these random effects is controlled by a binary $n \times n$ adjacency matrix $W$, which has elements $w_{kj}$ that are equal to one or zero depending on whether areas $(k,j)$ are defined to be neighbours. A common specification is that areas $(k,j)$ are neighbours if they share a common border, which corresponds to $w_{kj}=1$ and is denoted by $k \sim j$. In contrast, areas that do not share a common border have $w_{kj}=0$. Based on this adjacency matrix, the CAR prior proposed by Leroux et al. (1999) has full conditional distributions given by
\begin{align}
(2) \quad \phi_k|\phi_{-k}, W, \tau^2, \rho, \mu & \sim 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{\sigma^2}{\rho \sum_{j=1}^{n} w_{kj} + 1 - \rho} \right). 
\end{align}

The conditional expectation is a weighted average of the random effects in neighbouring areas and a global intercept \( \mu \), where the weights are controlled by \( \rho \). This parameter therefore determines the amount of spatial correlation amongst the random effects, with \( \rho = 0 \) corresponding to independence, while \( \rho \) close to one corresponds to strong spatial correlation. Diffuse priors are typically specified for the hyperparameters \( (\mu, \sigma^2, \rho) \), on the real line, positive real line and the interval \([0, 1]\) respectively.

A recent area or research (see for example Boots (2001), Lu and Carlin (2005)) is the statistical identification of boundaries or discontinuities in the risk map, which separate two areas that are geographically adjacent but have very different disease risks. The identification of such boundaries allows researchers to detect the spatial extent of a cluster of high risk areas, which in turn allows health resources to be targeted where they are most needed. Traditional approaches to boundary detection are based on Boundary Likelihood Values (BLV), which for two neighbourhing areas are calculated as

\begin{align}
(3) \quad BLV_{kj} = |\hat{R}_k - \hat{R}_j|,
\end{align}

the difference in risk (estimated from model (1) by posterior medians) between the areas in question. Using these values, the border between neighbouring areas \((k, j)\) is classified as a boundary in the risk surface if its BLV is within the top \(c\)% of values over the study region, for some fixed percentage \(c\). However, this approach has been criticised by numerous authors (see for example Jacquez et al. (2000)) due to it is ad-hoc nature, because the investigator has to choose the constant \(c\). This means that the investigator essentially chooses the number of boundaries that are identified, even though this is unknown and the goal of the analysis.

Therefore in this paper we develop an alternative statistical approach to detecting boundaries in disease risk maps, which separate populations that exhibit high and low risks of disease. Our approach detects boundaries by measuring the dissimilarity between populations living in neighbouring 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 behaviour. Our approach has the advantage of being fully automatic, so that the number of boundaries in the risk surface is determined by the data and not \textit{a-priori} by the investigator. The remainder of this paper is organised as follows. Section 2 presents our proposed model, and compares it against other approaches that have been proposed for this problem. Section 3 applies our model to hospital admission data for alcohol related conditions in Greater Glasgow, Scotland, while Section 4 contains a concluding discussion.

2. Methods

2.1. General approach

Our methodological approach is not based on boundary likelihood values (equation (3)), because after their calculation, an ad-hoc rule has to be applied to determine which ones are large enough to correspond to boundaries in the risk surface. Instead, we model the elements of the neighbourhood
matrix that correspond to adjacent areas as binary random quantities, rather than being fixed and equal to one. The motivation for this follows from the correlation structure induced between the random effects in areas \((k,j)\), and hence between the risk estimates \((R_k, R_j)\), by the value of \(w_{kj}\). If \(w_{kj}\) is estimated as zero the random effects \((\phi_k, \phi_j)\) and hence the risk estimates \((R_k, R_j)\) are conditionally independent, which corresponds to a boundary in the risk surface. In contrast, if \(w_{kj}\) equals one the random effects and risks are correlated, which corresponds to no boundary.

This general approach to boundary detection has previously been proposed by Lu et al. (2007), Ma and Carlin (2007), and Ma et al. (2010), who model the set of \(w_{kj}\) by logistic regression, a CAR prior, or an Ising model. However, this requires the large set of \(w_{kj}\) for all pairs of neighbouring areas to be estimated, which Li et al. (2011) argue are not well identified from the data. As an example, in the Greater Glasgow study presented in Section 3 there are 701 random neighbourhood relations \(w_{kj}\), that would need to be estimated from data in only 271 areas. Therefore, here we model the set of 701 \(w_{kj}\) as a function of a small number of parameters \(\alpha = (\alpha_1, \ldots, \alpha_q)\), rather than treating each \(w_{kj}\) as a separate unknown quantity. This results in a parsimonious yet flexible model for detecting boundaries in the risk surface, which avoids the weak parameter identifiability and slow MCMC convergence experienced by Li et al. (2011).

### 2.2. Modelling details

The likelihood component of our model is identical to (1), and represents the natural log of the set of disease risks by a set of random effects \(\phi\). These effects are modelled by a CAR prior similar to (2), which is given by

\[
\phi_k | \phi_{-k}, \mu, \alpha, \tau^2 \sim N \left( \frac{0.99 \sum_{j=1}^{n} w_{kj}(\alpha) \phi_j + 0.01 \mu}{0.99 \sum_{j=1}^{n} w_{kj}(\alpha) + 0.01}, \frac{\sigma^2}{0.99 \sum_{j=1}^{n} w_{kj}(\alpha) + 0.01} \right).
\]

The first difference from equation (2) is that the elements of the neighbourhood matrix are now random quantities, and depend on unknown parameters \(\alpha\). The second is that the overall spatial correlation parameter \(\rho\) is fixed at 0.99, rather than being estimated. A value close to one is chosen because it enforces strong spatial correlation, which allows the presence or absence of boundaries in the risk surface to be determined by the set of \(\{w_{kj}(\alpha)\}\). We note that if instead \(\rho\) was fixed at zero (corresponding to independence of the random effects), the set \(\{w_{kj}(\alpha)\}\) would disappear from equation (2), rendering boundary detection impossible. We also note that we do not fix \(\rho\) equal to one, because it is both theoretically and practically unappealing if an area is surrounded by boundaries, i.e. if \(\sum_{j=1}^{n} w_{kj}(\alpha) = 0\) for some area \(k\). From a theoretical perspective, the joint distribution of \(\phi\) in this situation is improper, while the full conditional distribution given by (2) would have an infinite mean and variance.

The proposed model for \(\{w_{kj}(\alpha)\}\) is based on the belief that boundaries (discontinuities) in the risk surface are likely to occur between populations that are very different, because homogeneous populations should have similar risk profiles. Therefore we define a vector of \(q\) non-negative dissimilarity metrics

\[
z_{kji} = \frac{|z_{ki} - z_{ji}|}{\theta_i} \quad \text{for } i = 1, \ldots, q,
\]
which represent the absolute difference in the value of a covariate $z_i$ between the two areas in question. Here, $\theta_i$ represents the standard deviation of $|z_{ki} - z_{ji}|$ over all pairs of contiguous areas, and we re-scale the dissimilarity metrics to improve the mixing and convergence of the MCMC algorithm. It is these dissimilarity measures that drive the detection of boundaries in the risk surface, and examples could include differences in the population’s social characteristics (e.g. average income) or risk inducing behaviour for the disease in question (e.g. smoking prevalence in the case of lung cancer). Using these data, $w_{kj}(\alpha)$ are modelled as

\[(6) \quad w_{kj}(\alpha) = \begin{cases} 
1 & \text{if } \exp(-\sum_{i=1}^{q} z_{kji} \alpha_i) \geq 0.5 \text{ and } j \sim k, \\
0 & \text{otherwise}
\end{cases}
\]

where pairs of areas that do not share a common border have $w_{kj}(\alpha)$ 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. Therefore we constrain the regression parameters to be non-negative, so that the greater the dissimilarity between two areas the more likely there is to be a boundary between them. In contrast, if two areas have identical covariate values (and hence homogeneous populations) there cannot be a boundary between them, regardless of the value of $\alpha$. This is why we do not include an intercept term in (6), as doing so would allow boundaries to be detected between areas with homogeneous populations. The regression parameters determine the number of risk boundaries across the study region, with larger values of $\alpha$ corresponding to more boundaries being detected. If only one dissimilarity metric $z$ is included in (6), then a plausible range of values for the single regression parameter $\alpha$ can be determined. At one extreme, no boundaries will be detected if $\alpha \leq -\ln(0.5)/z_{\text{max}}$, while at the other, all borders in the study region will be considered as boundaries (unless $z_{kj}=0$) if $\alpha > -\ln(0.5)/z_{\text{min}}$. Here $(z_{\text{min}}, z_{\text{max}})$ denote the minimum positive and maximum values of the dissimilarity metric. More generally, if there are $q$ dissimilarity metrics then boundaries are identified if

$$\exp(-z_{kji} \alpha_i) \times \ldots \times \exp(-z_{kjq} \alpha_q) < 0.5,$$

where the value of each component, $\exp(-z_{kji} \alpha_i)$, must lie between zero and one. Therefore, if $\alpha_i \leq -\ln(0.5)/z_{\text{max}}^i$, the dissimilarity measure $z_{kji}$ is not solely responsible for detecting any boundaries, because $\exp(-z_{j} \alpha_i)$ would be greater than 0.5 for all pairs of contiguous areas. Therefore in terms of interpretation, the dissimilarity metric can be said to have no effect on detecting boundaries if the entire 95% credible interval for $\alpha_i$ is less than $\alpha_{\text{min}} = -\ln(0.5)/z_{\text{max}}^i$. In contrast, if the interval lies completely above $\alpha_{\text{min}}$, then the metric can be said to have a substantial effect on identifying risk boundaries. We note that the usual statistical representation of ‘no effect’ (credible interval that includes zero) is not possible in this context, because the regression parameters are constrained to be non-negative. We also note that the approach we outline 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.

Each parameter $\alpha_i$ is non-negative and has an absolute upper limit of $M_i = -\ln(0.5)/z_{\text{max}}^{kji}$. However, in a boundary detection analysis one is looking to identify boundaries between collections of areas, which have similar risks within each collection but differ across the boundary. This would not be the case if $\alpha_i = M_i$, as it corresponds to each area having a markedly different risk from each of its neighbours. Therefore in this paper we fix $M_i$ so that at most 50% of borders can be classified as boundaries. Two possible priors for $\alpha_i$ are uniform ($\alpha_i \sim \text{Uniform}(0, M_i)$) or reciprocal
\( f(\alpha_i) \propto \frac{1}{\alpha_i} I[0 \leq \alpha_i \leq M_i] \), the first of which represents prior ignorance, while the second represents our prior belief that the risk surface is spatially smooth.

We also note that the model represented by (4) and (6) provides a much more realistic description of the spatial correlation structure amongst the random effects \( \phi \), than is assumed by model (2). In the latter, the level of spatial correlation in the data is controlled globally by a single parameter \( \rho \), which forces the same level of spatial correlation across the entire region. In contrast, the approach proposed here allows the random effects in geographically adjacent areas to be conditionally independent or correlated, depending on whether the populations living in these areas are similar or different.

3. Case study

3.1 Data

The data we model are publicly available from the Scottish Neighbourhood Statistics (SNS) database (http://www.sns.gov.uk). The study region is the Greater Glasgow and Clyde health board, which contains the city of Glasgow in the east, and the river Clyde estuary in the west. Glasgow is the largest city in Scotland, with a population of around 600,000 people. It is also known to contain some of the poorest people in Europe (Leyland et al. (2007)), and has rich and poor communities that are geographically adjacent. This study region is partitioned into \( n = 271 \) Intermediate Geographies (IG), which were developed specifically for the distribution of small-area statistics, and have a median area of 124 hectares and a median population of 4,239.

The disease data we model are the number of people admitted to acute and psychiatric hospitals in each IG with a main or secondary diagnosis of alcohol related conditions, during the four year period spanning 2001 to 2004. The expected numbers of cases were calculated by external standardisation, using age and sex adjusted rates for the whole of Scotland. These rates were obtained from the Information Services Division (ISD), which is the statistical arm of the National Health Service in Scotland. The simplest measure of disease risk is the standardised incidence ratio, which is presented in Figure 1 as a choropleth map. The Figure shows that the risk of admission to hospital is highest in the heavily deprived east end of Glasgow (east of the study region), as well as along the banks of the river Clyde (the thin white line running south east).
We consider two covariates as potential dissimilarity metrics, both of which have been shown to effect alcohol dependence and abuse rates. The first is a measure of poverty, as numerous studies (see for example Catalano et al. (1993) and Khan et al. (2002)) have shown a link between unemployment and increased alcohol dependence. The variable we consider here is the percentage of people in each IG that are defined to be income deprived, which means they are in receipt of a combination of means-tested benefits. The second variable we consider is ethnicity, because previous studies (for example Grant et al. (2004)) have also shown a link to alcohol dependence. The only variable available to measure ethnicity is the percentage of school children in each IG from ethnic minorities (non-white). While we appreciate this variable is imperfect in many ways (e.g. it relates to children not adults, and it does not differentiate between the different ethnic groups), it is the only measure of ethnicity available for our study.

### 3.2 Results

Both covariates were included in an initial model, and the posterior medians and 95% credible intervals for the regression parameters $\alpha_i$ are displayed in Table 1. Also displayed in Table 1 is $\alpha_{min}$, the threshold value below which the dissimilarity metric does not solely detect any boundaries in the risk surface. Income deprivation has a substantial effect on boundary detection, as the entire 95% credible interval lies above the threshold value $\alpha_{min} = 0.1330$. However, the corresponding interval for ethnicity lies completely below the the ‘no effect’ thresholds, suggesting that it is only income deprivation that determine the locations of the risk boundaries.
Table 1. Summary of the posterior distributions of the regression parameters $\alpha_i$.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Estimate</th>
<th>95% credible interval</th>
<th>$\alpha_{min}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income deprivation</td>
<td>0.2597</td>
<td>(0.2534, 0.2681)</td>
<td>0.1330</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>0.0092</td>
<td>(0.0006, 0.0225)</td>
<td>0.1258</td>
</tr>
</tbody>
</table>

As a result, the model was re-fitted with only income deprivation, and the estimated risk map and boundaries are displayed in Figure 2. The boundaries detected in the risk surface are displayed as solid white lines, and relate to border between regions $(k,j)$ for which the posterior median of $w_{kj}$ equals zero. The study region is split completely in two by the river Clyde (the thin white line running south east), and areas on opposite banks are not assumed to be neighbours. Therefore no boundaries can be detected across the river, which explains the absence of a boundary along its banks.

Figure 2. Estimated risk map with boundaries (white lines).

The risk map shows a similar spatial pattern to the SIR map in Figure 1, although as expected, it is smoother and has less extreme values. The model has detected 252 boundaries in the risk surface, which comprises 36% of the total boundaries in the study region. This suggests that the risk of admission to hospital from alcohol abuse across Greater Glasgow is far from spatially smooth, with a number of high and low risk areas that are geographically adjacent. The level of spatial smoothness also varies across the study region, with the majority of the risk boundaries being observed in the city of Glasgow, rather than in the surrounding areas.

The majority of the estimated risk boundaries appear to correspond to sizeable changes in the risk surface, suggesting that income deprivation appears to be an appropriate dissimilarity metric for detecting such boundaries. However, a few of the boundaries appear to be ‘false positives’, and
correspond to areas having different income deprivation levels but similar risk profiles. The identification of such false boundaries is of interest in their own right, and require a more detailed investigation into why the risk profiles are similar given the vastly different levels of poverty.

4. Conclusions

Here we have outlined an automatic statistical approach to detecting risk boundaries, that does not rely on ad-hoc tuning constants being specified. Our approach also does not suffer from parameter identifiability problems, as the detection of boundaries is achieved using a small number of additional unknown parameters. Our approach is based on the assumption that risk boundaries are more likely to occur between neighbouring populations that are very different, because homogeneous populations are likely to have similar risk profiles. Thus, our model is crucially dependent on the availability of good quality covariate information, which is used to measure the similarity between populations in neighbouring areas.

The alcohol example presented in Section 3 illustrates the importance of having good quality covariate data, as the majority of estimated risk boundaries do correspond to sizeable changes in risk. However, it also identifies some boundaries that do not appear to be real, as well as ignoring others where there is a clear discontinuity in the risk surface. This imperfection could be caused by the omission of other important dissimilarity metrics, the data for which may not be available. Our results also illustrate that the level of spatial correlation is not constant across the study region, and that spatial correlation is to be expected between disease risks in some but not all pairs of neighbouring areas. This in turn suggests that a random effects model that has a single correlation parameter (such as (2)) will be unable to capture the complex correlation structure observed in small-area spatial data of this type.

REFERENCES


