Cluster detection and risk estimation for spatio-temporal health data

Duncan Lee\textsuperscript{1} and Andrew Lawson\textsuperscript{2}

\textsuperscript{1} School of Mathematics and Statistics, University of Glasgow
\textsuperscript{2} Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Medical University of South Carolina

August 7, 2014

Abstract

In epidemiological disease mapping one aims to estimate the spatio-temporal pattern in disease risk and identify high-risk clusters, allowing health interventions to be appropriately targeted. Bayesian spatio-temporal models are used to estimate smoothed risk surfaces, but this is contrary to the aim identifying a group of areal units that exhibit elevated risks compared with their neighbours. Therefore in this paper we propose a new Bayesian hierarchical modelling approach for simultaneously estimating disease risk and identify high-risk clusters in space and time. Inference for this model is based on Markov chain Monte Carlo (MCMC) simulation, using the freely available \textit{R} package \textit{CARBayesST} that has been developed in conjunction with this paper. Our methodology is motivated by two case studies, the first of which assesses if there is a relationship between Public health District and colon cancer risk in Georgia, while the second looks at the impact of the smoking ban in public places in England on cardiovascular disease risk.

Keywords: Bayesian modelling, Cluster detection, Spatio-temporal risk mapping

1 Introduction

Disease mapping is the area of statistical epidemiology concerned with modelling the spatio-temporal pattern in disease risk. The disease could be infectious or chronic, and is likely to exhibit both spatial and temporal variation. Spatial variation in risk may be caused by similar variation in environment exposures or risk inducing behaviours such as smoking, and can result in substantial health inequalities. For example, a 2014 Office for National Statistics (ONS) report found a differential of nineteen years in the average healthy life expectancy between the richest and poorest neighbourhoods.
Similarly, disease risk also exhibits temporal variation, due to changes in population demographics such as gentrification effects, or the introduction of health legislation such as the ban on smoking in public places in England on 1st July 2007).

Estimation of this spatio-temporal variation allows researchers to address key public health goals, including: (i) forecasting of future disease burden; (ii) detection of high risk disease clusters; (iii) estimation of disease trends; (iv) surveillance of infectious disease epidemics; and (v) identification of risk factors associated with disease. Disease maps are routinely produced by health agencies, including the Cancer e-Atlas from Public Health England (PHE), and the weekly influenza maps produced by the Centres for Disease Control and Prevention (CDC) in the USA. However such maps summarise raw disease rates, which are subject to random fluctuations that can mask the spatio-temporal pattern in disease risk. Furthermore, they do not allow an estimation of the probability that an area exhibits an elevated risk of disease, called an exceedence probability. These problems have led to a range of statistical model development, including Bayesian spatio-temporal models for risk estimation (Bernardinelli et al., 1995, Knorr-Held, 2000 and Lawson, 2013), as well as scan statistics (see Kulldorff et al., 2005) and point process methodology (Diggle et al., 2005) for cluster detection. However, risk estimation and cluster detection have fundamentally different goals, as the former aims to use the spatio-temporal autocorrelation in the data to estimate smoothed maps of disease risk, while the latter aims to identify clusters of areal units that exhibit elevated risks compared with neighbouring areas. Thus typically risk models do not lead directly to inference about clusters, while scan statistics do not result in a map of disease risk for the entire region.

A small number of papers have proposed purely spatial models to simultaneously estimate disease risk and identify clusters, including Gangnon and Clayton (2000), Knorr-Held and Rasser (2000), Green and Richardson (2002), Forbes et al. (2013), Wakefield and Kim (2013), and Anderson et al. (2014). Additionally a two-stage approach was discussed by Charras-Garrido et al. (2013), where the first stage estimates disease risk while the second identifies clusters in the estimated risk surface. However, only Choi and Lawson (2011), Lawson et al. (2012) and Li et al. (2012) have extended these models to the spatio-temporal domain, and have focused on detecting shared latent structures and unusual temporal trends. Therefore this paper proposes a novel class of Bayesian spatio-temporal models that can detect clusters of areal units that exhibit excessively high (or low) disease risk, while utilising the autocorrelation in the data within a cluster to estimate the spatio-temporal pattern in
disease risk. The model is able to detect clusters dynamically, so that both the membership of the cluster and its average risk level can evolve over time. Inference for this model class is based on Markov chain Monte Carlo (MCMC) simulation, and other researchers can use the methodology via the freely available R package CARBayesST, which was developed by the authors in conjunction with this paper.

The methodological development is motivated by two applications, a study of colon cancer incidence in the state of Georgia, USA, and a study of cardiovascular disease risk in England. These two applications are outlined in Section two, together with a review of spatio-temporal disease mapping. Section three presents our methodological contribution, and details the software that has been developed. Section four quantifies the performance of our methodology by simulation, while the results from the two case studies are presented in Section five. Finally Section concludes the paper.

2 Background

2.1 Spatio-temporal disease mapping

The study region is partitioned into \( i = 1, \ldots, N \) contiguous areal units, about which data are collected for \( t = 1, \ldots, T \) consecutive time periods. The observed and expected numbers of disease cases in unit \( i \) during time period \( t \) are denoted by \((y_{it}, e_{it})\) respectively, where the latter is computed using external standardisation and accounts for the heterogeneous population sizes and demographic structures across the \( N \) areal units. The same set of disease rates are used in the standardisation for all time periods, because the temporal trend in risk is then estimated against a common baseline. A Poisson model for the random variable representation of the disease counts \( Y_{it} \) is given by

\[
Y_{it}|e_{it}, \theta_{it} \sim \text{Poisson}(e_{it}\theta_{it}) \quad \text{for} \quad i = 1, \ldots, N \quad \text{and} \quad t = 1, \ldots, T. \tag{1}
\]

Here \( \theta_{it} \) is the risk of disease in areal unit \( i \) during time period \( t \), and log-linear models are typically specified for \( \theta_{it} \). The first model for this type of data was proposed by [Bernardinelli et al. (1995)](Bernardinelli et al. 1995), who represented \( \{\theta_{it}\} \) by a set of spatially varying linear time trends. However, linear trends are not always appropriate, so the most common model in spatio-temporal disease mapping was proposed by [Knorr-Held (2000)](Knorr-Held 2000), and is based on a flexible main effects plus interaction decomposition given by:
\[
\ln(\theta_{it}) = \delta + \phi_i + \theta_i + \alpha_t + \beta_t + \gamma_{it}, \quad (2)
\]

\[
\delta \sim \text{N}(0, 100),
\]

\[
\phi_i | \phi_{-i}, W, \tau^2_\phi \sim \text{N} \left( \frac{\sum_{j=1}^N w_{ij} \phi_j}{\sum_{j=1}^N w_{ij}}, \tau^2_\phi \right),
\]

\[
\theta_i | \tau^2_\theta \sim \text{N} \left( 0, \tau^2_\theta \right),
\]

\[
\alpha_t | \alpha_{t-1}, \tau^2_\alpha \sim \text{N} \left( \alpha_{t-1}, \tau^2_\alpha \right),
\]

\[
\beta_t | \tau^2_\beta \sim \text{N} \left( 0, \tau^2_\beta \right),
\]

\[
\gamma_{it} | \gamma_{-it}, \tau^2_\gamma \sim \text{N} \left( m_{it}, \tau^2_\gamma v_{it} \right).
\]

The spatial \((\phi_k + \theta_k)\) and temporal \((\alpha_t + \beta_t)\) main effects are a convolution of independent \((\theta_k, \beta_t)\) and autocorrelated \((\phi_k, \alpha_t)\) processes. Spatial autocorrelation is modelled by an intrinsic conditional autoregressive (ICAR, Besag et al. 1991) model with a binary \(N \times N\) neighbourhood matrix \(W\), where \(w_{ij} = 1\) if areal units \((i, j)\) share a common border and is zero otherwise. Thus the conditional expectation is the mean of the random effects in neighbouring areal units, inducing spatial smoothness into \(\phi_k\). Temporal autocorrelation is modelled by a first order random walk, where the conditional expectation of \(\alpha_t\) is \(\alpha_{t-1}\). Four different models were proposed by Knorr-Held (2000) for the interaction term \(\gamma_{it}\), which respectively assume independence, purely spatial autocorrelation, purely temporal autocorrelation and spatio-temporal autocorrelation. The first of these has \((m_{it} = 0, v_{it} = 1)\) and further details are given by Knorr-Held (2000). This model does not have an inbuilt mechanism for delineating clusters of areas at elevated or reduced risk compared with their neighbours during the model fitting. However, Charras-Garrido et al. (2013) utilise a post-hoc approach in a purely spatial setting, by applying a Bayesian regularization for Gaussian mixtures (Fraley and Raftery, 2007) clustering algorithm to the fitted risk surface \(\{\hat{\theta}_{it}\}\).

2.2 Case study 1 - Colon cancer risk in Georgia, USA

The state of Georgia, USA contains \(N = 159\) counties, whose health services are collectively administered by 18 Public Health Districts (PHD, (http://dph.georgia.gov/public-health-districts). These districts have autonomy in administering health care, and here we investigate if the risks of colon cancer incidence (International Classification of Disease (ICD) codes C18) show any evidence of clustering by PHD. The data are yearly counts of incidence between 2001 and 2010, and are mostly low but skewed to the right, with a median count of 4, and a maximum of 136. The standardised incidence
ratio (SIR) is an exploratory measure of disease risk, and for county \( i \) during time period \( t \) is given by
\[
SIR_{it} = \frac{y_{it}}{e_{it}}.
\]
An SIR of one represents an average baseline risk, while a value of 1.2 corresponds to a 20% elevation in disease risk. The SIR is displayed for the first (2001) and last (2010) years of the study period in the top row of Figure 1, which in addition to skewness, also shows substantial changes in the SIR between the two years (the maximum pairwise correlation in SIR in different years is 0.47). The white lines on the SIR maps depict the 18 Public Health Districts (PHD) in Georgia.

2.3 Case study 2 - Cardiovascular disease risk in England

A ban on indoor smoking in public places was introduced in England on 1st July 2007, and the goal of this study is to assess whether the ban had any effect in reducing cardiovascular disease risk. The data for this study are quarterly counts of admissions to hospital due to cardiovascular disease (ICD codes I00-I99) for the set of \( N = 323 \) local authorities in mainland England for \( T = 36 \) quarters from 2003 to 2011 inclusive. The expected counts for all quarters were computed using age and sex specific disease rates for England in 2003, the first year of the study. The SIR for the first (2003 quarter 1) and last (2011 quarter 4) time periods are displayed in the bottom panel of Figure 1, and show definite clusters of high-risk areas in Greater Manchester (mid to north of the study region) and Tyneside (far north east of the study region). The clusters are also temporally stable, with a mean correlation of 0.95 between the spatial SIR surfaces at different time periods. The existence of temporal trends in risk are hard to see from the figure, and we return to this question in Section 5.

3 Methodology

The spatio-temporal cluster detection and risk estimation models developed in this paper are described below, where the first subsection describes the base clustering model, while the second subsection extends it to allow within cluster spatial autocorrelation. The final two subsections provide a critique of model comparison techniques commonly used in the disease mapping and clustering literature, and describe the software package developed in conjunction with this paper to fit the models proposed.

3.1 Clustering risk model

The pure clustering model represents the set of risks \( (\theta_{1t}, \ldots, \theta_{Nt}) \) for the \( N \) areal units during time period \( t \) by at most \( G \) distinct risk levels or classes \( (\lambda_{1t}, \ldots, \lambda_{G_t}) \). These risk classes are ordered as \( \lambda_{1t} < \lambda_{2t} < \ldots < \lambda_{Gt} \), which helps mitigate against the label switching problem common in mixture
Figure 1: Maps showing the SIR for colon cancer incidence in Georgia in 2001 and 2010 (top panel) and cardiovascular disease in England in 2003 quarter 1 (Q1) and 2011 quarter 4 (Q4) (bottom panel). For the latter the Easting and Northing coordinates are in kilometres.
models. The assignment of areal unit $i$ to a risk class during time period $t$ is controlled by $Z_{it}$, where $Z_{it} \in \{1, \ldots, G\}$. This leads to the first level risk model:

$$Y_{it}|e_{it}, \theta_{it} \sim \text{Poisson}(e_{it}\theta_{it}) \quad \text{for } i = 1, \ldots, N \text{ and } t = 1, \ldots, T,$$

$$\ln(\theta_{it}) = \lambda_{t,Z_{it}}.$$  

We term $G$ the maximum number of risk classes because $\{Z_{it}\}$ are allowed to take values in the set $\{1, \ldots, G\}$ but are not forced to, meaning that some of these classes may be empty. Here $G$ is fixed in the model rather than treating it as a parameter to be updated using, for example, a reversible jump MCMC algorithm [Knorr-Held and Rasser 2000]. This approach is chosen for simplicity, both in computational terms to obtain results in a relatively short time frame, but also in terms of interpretation for epidemiologists. Instead, we choose the maximum number of risk classes $G$ to be overly large, and implement a constraint to penalise models with too many risk classes. The simulation study in Section 4 shows this approach works well, and is able to identify a number of different cluster structures in data with less than $G$ classes, including a risk surface that has no clusters of areas exhibiting elevated risk.

To complete the specification of the clustering model prior distributions are required on the mean risk levels $\{\lambda_{tj}\}$ and the class indicators $\{Z_{it}\}$. *A-priori*, one would expect the mean risk in each class to evolve smoothly over time, which is achieved by putting first order random walk priors on the set $(\lambda_{1j}, \ldots, \lambda_{Tj})$ for each risk class $j = 1, \ldots, G$. However, the cluster means follow the ordering constraint $\lambda_{t1} < \lambda_{t2} < \ldots < \lambda_{tG}$ across the $G$ groups for each time period, leading to the prior model:

$$\lambda_{tj}|\lambda_{t-1,j} \sim N(\lambda_{t-1,j}, \sigma^2)I(\lambda_{tj} \in [\lambda_{t,j-1}, \lambda_{t,j+1}]) \quad \text{for } j = 1, \ldots, G, \ t = 2, \ldots, N,$$

$$\lambda_{1j} \sim \text{Uniform}(\lambda_{1,j-1}, \lambda_{1,j+1}) \quad \text{for } j = 1, \ldots, G,$$

$$\sigma^2 \sim \text{Inverse-Gamma}(a, b).$$

At time period 1 one a flat prior is specified on the closed interval obeying the same ordering constraint, and in the above notation ($\lambda_{00} = -\infty, \lambda_{t,G+1} = \infty$). The vector of class indicators for area $i$ for all $T$ time periods are denoted by $Z_i = (Z_{i1}, \ldots, Z_{iT})$, and are also modelled as temporally autocorrelated, as one would expect the risk in an area to evolve smoothly in time. This results
in preferring models where $Z_{it} = Z_{i,t-1}$, which suggests a Markov model for each $Z_i$. Independent Markov models are specified for each areal unit $i$, because areal units at opposite ends of the study region are likely to have similar risks. Examples include two wealthy suburban areas on opposite sides of a city, or two largely rural local authorities in England in the north and south of the country. Therefore the joint prior distribution for $(Z_1, \ldots, Z_N)$ is decomposed as:

$$f(Z_1, \ldots, Z_N) = \prod_{i=1}^N f(Z_i) = \prod_{i=1}^N \left[ f(Z_{i1}) \prod_{t=2}^T f(Z_{it}|Z_{i,t-1}) \right],$$

and the introduction of spatially correlated risks is discussed in the next subsection. Each $Z_{it} \in \{1, \ldots, G\}$, and we propose discrete distributions $f(Z_{i1})$ and $f(Z_{it}|Z_{i,t-1})$ given by

$$P(Z_{i1}) = \frac{\exp(-\beta(Z_{i1} - G^*)^2)}{\sum_{r=1}^G \exp(-\beta(r - G^*)^2)},
\alpha, \beta \sim \text{Uniform}(0, M)$$

Here, temporal autocorrelation in the class indicators is controlled by $\alpha$, and $\alpha = 0$ corresponds to no temporal autocorrelation as in this case there is no dependence of $Z_{it}$ on $Z_{i,t-1}$. The second term in the Markov model (5) is a constraint imposed to guard against a potential lack of identifiability that arises because $G$ is chosen to be overly large. Consider the case where $G = 3$ but the data contain only one risk class (i.e. a background risk level close to one), then the model will fit equally well if $Z_{it} = 1, 2, 3$ as long as all the class indicators $Z_{it}$ are the same. Therefore we set $G^* = (G+1)/2$, which represents the middle of the class indicators, so $\beta$ controls the strength of a constraint that favours models whose class indicators are close to the middle class $G^*$. Thus $\beta$ plays no epidemiologically meaningful role, and the term $\beta(Z_{it} - G^*)^2$ is included in (5) as a constraint to improve the convergence of the MCMC algorithm. However, in an initial testing of the model $\beta$ tended to be hard to identify, so here we set $\beta = 1$, which leads to good model performance in the simulation study. The clustering only model proposed here is given by (3), (4) and (5). This model induces temporal autocorrelation in the risks $\theta_{it}$ via the temporally autocorrelated priors on the class means $\lambda_tj$ and the class indicators $Z_{it}$, but does not induce any spatial smoothness on the risks. Furthermore, during time period $t$ two areas that have the same cluster indicator have identical risks, where as it is more realistic that the risks within a class will be spatially autocorrelated and similar but not identical. Therefore the next subsection
extends this purely clustering model to allow for spatial autocorrelation.

3.2 Extending the cluster model to incorporate spatial autocorrelation

Conditional autoregressive priors applied to a set of random effects are the most commonly used mechanism for inducing spatial autocorrelation in areal unit disease data, and reviews are given by Wakefield (2007) and Lee (2011). The latter argues that the CAR prior proposed by Leroux et al. (1999) is the most appealing out of those commonly used from both theoretical and practical standpoints, and it is thus the prior used here. The extended clustering model with CAR induced spatial autocorrelation proposed here is given by

\[
Y_{it}|e_{it}, \theta_{it} \sim \text{Poisson}(e_{it}\theta_{it}) \quad \text{for } i = 1, \ldots, N \text{ and } t = 1, \ldots, T, 
\]

\[
\ln(\theta_{it}) = \lambda _{it}z_{it} + \phi_{it},
\]

\[
\phi_{it}|\phi_{-it} \sim N \left( \frac{\rho \sum_{j=1}^{N} w_{ij} \phi_{jt}}{\sum_{j=1}^{N} w_{ij} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^{N} w_{ij} + 1 - \rho} \right),
\]

\[
\tau^2 \sim \text{Inverse-Gamma}(a,b),
\]

\[
\rho \sim \text{Uniform}(0,1),
\]

which is combined with (4) and (5) to complete the model specification. Here \( \rho \) is a spatial autocorrelation parameter, and \( \rho = 1 \) corresponds to the ICAR prior proposed by Besag et al. (1991) for strong spatial smoothing, while if \( \rho = 0 \) the random effects are independent with a mean of zero and a constant variance of \( \tau^2 \). Thus this model utilises a single set of random effects with a spatial autocorrelation parameter, where as model (2) uses two sets of random effects fixed at the extreme cases of \( \rho = 0, 1 \). In this model there is a separate random effects surface \( \phi_t = (\phi_{1t}, \ldots, \phi_{Nt}) \) for each time point, although the \( T \) surfaces are controlled globally by time invariant parameters \( (\rho, \tau^2) \).

Alternative approaches allowing for spatial autocorrelation have been proposed for areal unit data, including geostatistical models (Kelsall and Wakefield, 2002), geographically weighted regression (Young et al., 2009), spline-based smoothing models (Ugarte et al., 2010), simultaneous autoregressive models (Wall, 2004) and moving average or convolution models (Best et al., 2005). We consider the latter here as an alternative to (6), which is given by
\[ Y_{it}|e_{it}, \theta_{it} \sim \text{Poisson}(e_{it}\theta_{it}) \quad \text{for } i = 1, \ldots, N \text{ and } t = 1, \ldots, T, \quad (7) \]

\[
\ln(\theta_{it}) = \lambda_{t,Z_{it}} + \sum_{j=1}^{N} f(d_{ij}|\rho)X_{jt},
\]

\[ X_{jt} \sim N(0, \tau^2), \]

\[ \tau^2 \sim \text{Inverse-Gamma}(a,b), \]

\[ \rho \sim \text{Uniform}(0, P), \]

again in conjunction with (4) and (5). Here \( f(d_{ij}|\rho) \) is a scaled Gaussian kernel of distance \( d_{ij} \) between two area’s centroids defined by \( f(d_{ij}|\rho) = \exp(-d_{ij}^2/(2\rho))/\sum_{k=1}^{N} \exp(-d_{ik}^2/(2\rho)) \), except that \( f(d_{ii}|\rho) = 0 \). As with the CAR model (6), \( \rho \) acts as a spatial smoothness parameter, with larger values of \( \rho \) leading to spatially smoother surfaces. In preliminary analyses we also implemented a Simultaneous autoregressive or observation driven model of the form

\[ \ln(\theta_{it}) = \lambda_{t,Z_{it}} + \sum_{j=1}^{N} f(d_{ij}|\rho)[\ln(Y_{jt}/e_{j}) - \lambda_{t,Z_{jt}}], \]

but as it performed worse than both the CAR and convolution models we do not discuss it further.

3.3 Model comparison and assessment in disease mapping

There are many metrics for measuring the fit or appropriateness of a statistical model to a data set, and we describe the ones we use in this paper below.

3.3.1 Overall measures of model fit

Numerous measures of model fit have been proposed in the statistical literature, many of which trade off the fit to the data against the number of effective parameters. The most popular of these criteria in a Bayesian setting is the Deviance Information Criterion (DIC, [Spiegelhalter et al. 2002]), which is based on \( D(y, \Theta) = -2\ln(f(y|\Theta)) \), the deviance of the data \( y \) given the set of parameters in the model \( \Theta \). The DIC is given by

\[ \text{DIC} = \bar{D}(y, \Theta) + p_D(y, \Theta), \]

where \( \bar{D}(y, \Theta) \) is the average deviance computed over the posterior distribution of \( \Theta \). The second term is the effective number of parameters in the model estimated by \( p_D(y, \Theta) = \bar{D}(y, \Theta) - D(y, \hat{\Theta}), \)
where $\hat{\Theta}$ is an estimate of $\Theta$ such as the posterior mean. Models with smaller DIC values are preferred. An alternative is a predictive measure of model fit called the conditional predictive ordinate (CPO, [Pettit 1990]), which for observation $y_{it}$ is given by

$$\text{CPO}_{it} = f(y_{it}|\mathbf{y}_{-it}) = \int_{\Theta} f(y_{it}|\Theta) f(\Theta|\mathbf{y}_{-it}) d\Theta,$$

where $\mathbf{y}_{-it}$ denotes all data points except $y_{it}$. Following [Congdon 2005] this can be approximated based on $M$ MCMC samples as

$$\hat{\text{CPO}}_{it} = \left[ \frac{1}{M} \sum_{j=1}^{M} \frac{1}{f(y_{it}|\Theta^{(j)})} \right]^{-1},$$

where $\Theta^{(j)}$ is the $j$th posterior sample. Then a summary measure of the CPO for all $NT$ data points is called the log marginal predictive likelihood (LMPL), and is given by

$$\text{LMPL} = \sum_{i=1}^{N} \sum_{t=1}^{T} \ln(\hat{\text{CPO}}_{it}),$$

and models with larger LMPL values are preferred.

### 3.3.2 Exceedence probabilities

Posterior Exceedence Probabilities (PEP) are a popular tool in disease mapping to quantify the extent to which a single area exhibits an elevated risk, and a comprehensive study on their utility is given by [Richardson et al. 2004]. For area $i$ during time period $t$ the PEP is defined as

$$\text{PEP}_{it}(c) = P(\theta_{it} > c|\mathbf{y}),$$

the posterior probability that the risk $\theta_{it}$ is greater than a threshold level $c$, which can be easily computed from $M$ posterior samples obtained from MCMC simulation. PEP are local measures of exceedence of a specified risk level $c$ for each area, rather than being overall measures of model fit as described above. PEP can be applied to both real and simulated data, and in the latter a better fitting model would have a lower $\text{PEP}_{it}(c)$ for an area that has a baseline risk close to one and higher $\text{PEP}_{it}(c)$ for an area in a high-risk cluster.
3.3.3 Measures for simulated data

Simulated data are often used to test model accuracy, and has the advantage that the true risk surface that generated the noisy disease counts is known. In this setting the accuracy with which a model estimates the risk surface can be quantified by its Root Mean Squared Error (RMSE), which for our spatio-temporal data is given by

\[
RMSE = \sqrt{\frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} (\theta_{it} - \hat{\theta}_{it})^2},
\]

where \((\theta_{it}, \hat{\theta}_{it})\) are the true and estimated risks respectively. RMSE is a commonly used measure of model quality, as it trades off the bias of an estimate against its variance. In the clustering context considered in this paper, it is of interest to quantify the accuracy with which a model can estimate the cluster structure in the data. Suppose the true spatio-temporal risk surface has \(H\) groups or classes, \(C^t_\theta = \{C^t_1, \ldots, C^t_H\}\), where as a model estimates that the risk surface is clustered into \(F\) classes \(C^e_\theta = \{C^e_1, \ldots, C^e_F\}\). Then the Rand index \([\text{Rand} \text{, 1971}]\) quantifies the agreement between the two cluster structures, and is given by

\[
\text{Rand} = \frac{a + b}{\binom{NT}{2}}.
\]

Here \(a\) is the number of pairs of risks \((\theta_{it}, \theta_{jr})\) that are classified in the same cluster under both cluster structures, while \(b\) is the number of pairs of risks that are classified in different clusters under both cluster structures. Intuitively, the Rand index measures the percentage agreement between the two cluster structures, with a value of one indicating the structures are identical.

3.4 Software and implementation

An add-on package to the statistical software R \([\text{R Core Team} \text{, 2013}]\) called \textit{CARBayesST} has been developed in conjunction with this paper to implement the models discussed here, which is one of the first R packages to implement spatio-temporal models for areal unit data. The package is freely available to download from \textit{http://cran.r-project.org/}, and can fit the three models proposed here, namely the cluster only model (given by (3) - (5), the cluster model with CAR induced spatial autocorrelation given by (4), (5) and (6), and the cluster model with convolution induced spatial autocorrelation given by (4), (5) and (7). Additionally, the software can also implement the model proposed by \textit{Knorr-Held} \textit{2000} with independent interaction effects and given by (2).
4 Model assessment via simulation

This section presents a simulation study that compares the relative performances of the clustering models proposed here against commonly used spatio-temporal disease mapping models. Specifically, the model proposed by Knorr-Held (2000) with independent interactions given by (2) is compared against the three clustering models described in Section 3, namely the cluster only model given by (3), (4) and (5), the cluster model with CAR spatial smoothing given by (6), (4) and (5), and the cluster model with convolution spatial smoothing given by (7), (4) and (5). Each of these models is fitted with a maximum of $G = 5$ risk classes, where as the simulated data exhibit either one or two classes. In what follows the model of Knorr-Held (2000) is referred to as Knorr-Held, while the three clustering models are denoted as Clust-Only, Clust-CAR, Clust-Conv respectively. All analyses in this study were conducted using the CARBayesST software.

4.1 Data generation and study design

Simulated data are generated for the $N = 159$ counties comprising the state of Georgia, USA for a period of $T = 10$ years, which is the study region for the first case study outlined in Section 2.2. Disease counts are generated from model (1), where the size of the expected numbers $\{e_{it}\}$ is varied in this study to assess their impact on model performance. The log-risk surface is generated from a multivariate Gaussian distribution, with a piecewise constant mean (for clustering) and a spatially smooth variance matrix. The latter induces smooth spatial variation into the log-risks, and is defined by an exponential correlation model. The clusters are displayed by the black shaded regions in the right panel of Figure 2, which displays sample realisations of the risk surface with clusters and smoothly varying background risks under a number of different scenarios (see below). The cluster structure was chosen to contain different cluster shapes, including a singleton cluster, a long thin cluster, a block cluster and a cluster with a hole in the middle. We consider five different scenarios for this study, which are outlined below.

1. **No clustering** - The null scenario where the risk surface exhibits a smoothly varying baseline risk close to one (risks between 0.7 and 1.3) with no clustering, as illustrated in the left panel of Figure 2. This scenario tests whether the methods false identify clusters when none are present.

2. **Medium temporally consistent clustering** - The risk surface contains clusters with average risks of 2, as illustrated in the middle panel of Figure 2. This scenario tests whether the methods can identify clusters whose risk is elevated slightly above the background level.
Figure 2: Maps showing example realisations of the spatial surface used in the simulation study. The left panel shows the background risk close to one, while the middle and right panels respectively show medium and high risk clusters.

3. **High temporally consistent clustering** - The risk surface contains clusters with average risks of 3, as illustrated in the right panel of Figure 2. This scenario tests whether the methods can identify clusters whose risk is vastly elevated above the background level.

4. **Medium temporally inconsistent clustering** - The risk surface contains no clusters during time periods 1 to 3 or periods 8 to 10, and the realisations look like those under scenario 1. However, clusters with an average risk of 2 appear during time periods 4 to 7, which follow the template shown in the middle panel of Figure 2. This scenario tests the models ability to detect slightly elevated clusters that are not temporally consistent.

5. **High temporally inconsistent clustering** - The risk surface contains no clusters during time periods 1 to 3 or periods 8 to 10, and the realisations look like those under scenario 1. However, clusters with an average risk of 3 appear during time periods 4 to 7, which follow the template shown in the right panel of Figure 2. This scenario tests the models ability to detect high-risk clusters that are not temporally consistent.

The five scenarios above are repeated with expected numbers of disease cases $e_{it} \in [10, 25], [50, 100], [150, 250]$, which allows us to examine model performance for diseases with different prevalences. Inference for each model is based on 10,000 MCMC samples, which were generated following a burn-in period of a further 10,000 samples. Convergence was visually assessed to have been reached after 10,000 samples by viewing trace plots of the parameters for a number of simulated data sets.
4.2 Results

Two hundred data sets are generated under each of the fifteen combinations of scenario and disease prevalence \( (e_{it}) \), and the results are displayed in Table 1. The top panel of the table displays the RMSE (see 9) of the estimated spatio-temporal risk surfaces from each model, and a number of key themes emerge. Firstly, the Clust-CAR variant of the three clustering models typically has the lowest RMSE, although the differences from the Clust-Conv variant are not large (at most 0.017 higher). Both these models perform better in terms of RMSE than the Clust-Only model, which is due to the latter assuming the risk surface is piecewise constant at each MCMC iteration, while the former allow for within class variation. If there are no clusters in the risk surface (scenario 1) then the Clust-CAR model performs similarly to the Knorr-Held model in terms of RMSE, with values that differ by less than 0.011. The RMSE values for scenarios 2 and 3 suggest that the Knorr-Held model is generally preferable over the Clust-CAR model, which is likely to be because the former assumes a constant spatial main effect for each time period (excepting the interaction term), facilitating a temporal borrowing of strength in the estimation, in common with the way the data were generated. In contrast the clustering models assume separate spatial surfaces for each time period. This differential in performance is reversed for scenarios 4 and 5, in which clusters are only present for 40% of the time periods. In these scenarios the Clust-CAR model consistently outperforms the Knorr-Held model, which is due to its lack of a common spatial surface across all time periods.

The bottom panel of Table 1 quantifies the models’ abilities to identify clusters of areas exhibiting elevated risks, compared with other areas exhibiting background risk levels close to one. The cluster/class that area \( i \) is in during time period \( t \) is summarised by the posterior median of \( Z_{it} \) for each of the three clustering models. In contrast, the Knorr-Held model does not have an in-built mechanism for automatically partitioning the areas into risk classes, and thus we apply the posterior classification approach described in Charras-Garrido et al. (2012) and Charras-Garrido et al. (2013). This approach applies a Gaussian mixture model to the posterior median risk surface (Fraley and Raftery, 2007) to obtain a cluster structure, where again the number of mixture components is chosen by the model. The accuracy of the estimated class structure and the true structure is quantified by the Rand Index (see 10), where a perfect clustering gives a value of 1 and two completely random clusters gives a value of 0.5. The mean Rand index over the 200 simulated data sets in each scenario is presented in Table 1.
The table shows that the posterior classification approach in conjunction with the Knorr-Held model is unable to correctly identify the cluster structure in any of the fifteen combinations of scenario and disease prevalence ($e_{it}$), with mean Rand indices ranging between 0.641 and 0.918. This means that in scenario 1 this approach identifies clusters exhibiting elevated risks when none are present. In contrast, the three clustering models show almost complete accuracy in this scenario, with mean Rand indices ranging between 0.997 and 1.000. The best of the three clustering models is the Clust-Only model, and it shows greatest improvements over the other two clustering models in Scenarios 2 and 4 where the clusters are only slightly elevated above the background level. This is likely to be because the clustering ($\lambda_{it,z_{it}}$) and spatial smoothing ($\phi_{it}$ or $\sum_{j=1}^{N} f(d_{ij} | \rho) X_{jt}$) components of the Clust-CAR and Clust-Conv models are competing to account for the spatial variation in the data, and thus some elevated clusters are wrongly assigned to the spatial smoothing component rather than the clustering component. The Clust-Only model does not have a spatial smoothing component and hence does not suffer from this problem. However, when the clusters have a high-risk of three (Scenarios 3 and 5) all the clustering models perform well, because the variation in the risk surface is not smooth and hence the clusters are explained by the clustering component. Finally, the table shows that correct cluster identification is more difficult for rare diseases (small $e_{it}$), and only the Clust-Only model produces reliable results. This is because in this situation the true spatial structure is dominated by Poisson noise, making cluster detection more difficult.

5 Results from the case studies

5.1 Modelling

The four models compared in the simulation study are applied to the Georgia and England case studies, namely the Knorr-Held, Clust-Only, Clust-CAR and Clust-Conv models. Inference for each model is based on 30,000 McMC samples, which were generated by running three Markov chains for 20,000 iterations and discarding the first 10,000 of each as a burn-in period. In common with the simulation study, convergence was assessed visually using trace plots of the McMC samples for selected parameters. The three cluster models were fitted with $G = 5$, although a sensitivity analysis showed changing this did not largely effect the results. The overall fit of each model to each data set is summarised in Table 2, which displays the DIC and LMPL metrics as discussed in Section 3.
Table 1: Table showing the root mean square error (RMSE, top panel) for the risk surface estimated by each of the four models, and the Rand index (bottom panel) as a measure of the clustering ability of the models.

| Scenario | c_{it} | Knorr-Held | Clust-Only | Clust-CAR | Clust-Conv |
|----------|--------|------------|------------|-----------|------------|
| **RMSE** |        |            |            |           |            |
| 1        | [10, 25]| 0.062      | 0.085      | 0.073     | 0.071      |
|          | [50, 100]| 0.051      | 0.085      | 0.050     | 0.057      |
|          | [150, 250]| 0.043      | 0.084      | 0.040     | 0.048      |
| 2        | [10, 25]| 0.097      | 0.285      | 0.184     | 0.201      |
|          | [50, 100]| 0.064      | 0.125      | 0.112     | 0.118      |
|          | [150, 250]| 0.050      | 0.096      | 0.052     | 0.055      |
| 3        | [10, 25]| 0.106      | 0.237      | 0.231     | 0.248      |
|          | [50, 100]| 0.069      | 0.113      | 0.073     | 0.072      |
|          | [150, 250]| 0.053      | 0.110      | 0.049     | 0.057      |
| 4        | [10, 25]| 0.133      | 0.176      | 0.131     | 0.144      |
|          | [50, 100]| 0.094      | 0.105      | 0.093     | 0.102      |
|          | [150, 250]| 0.066      | 0.088      | 0.047     | 0.051      |
| 5        | [10, 25]| 0.176      | 0.177      | 0.175     | 0.191      |
|          | [50, 100]| 0.109      | 0.095      | 0.072     | 0.064      |
|          | [150, 250]| 0.071      | 0.094      | 0.044     | 0.053      |
| **Rand** |        |            |            |           |            |
| 1        | [10, 25]| 0.669      | 1.000      | 1.000     | 1.000      |
|          | [50, 100]| 0.710      | 1.000      | 1.000     | 1.000      |
|          | [150, 250]| 0.720      | 0.995      | 1.000     | 1.000      |
| 2        | [10, 25]| 0.885      | 0.739      | 0.727     | 0.727      |
|          | [50, 100]| 0.842      | 0.961      | 0.746     | 0.774      |
|          | [150, 250]| 0.823      | 0.996      | 0.979     | 0.997      |
| 3        | [10, 25]| 0.852      | 0.924      | 0.733     | 0.745      |
|          | [50, 100]| 0.845      | 0.999      | 0.981     | 1.000      |
|          | [150, 250]| 0.847      | 0.987      | 0.997     | 0.998      |
| 4        | [10, 25]| 0.748      | 0.878      | 0.878     | 0.878      |
|          | [50, 100]| 0.811      | 0.970      | 0.879     | 0.881      |
|          | [150, 250]| 0.838      | 0.983      | 0.984     | 0.997      |
| 5        | [10, 25]| 0.646      | 0.950      | 0.878     | 0.879      |
|          | [50, 100]| 0.815      | 0.996      | 0.972     | 0.997      |
|          | [150, 250]| 0.843      | 0.970      | 0.983     | 0.987      |

Table 2: Summary of the overall fit to the data of each model, as measured by the DIC and LMPL. The top panel of the table relates to the Georgia colon cancer case study, while the bottom panel relates to the England cardiovascular data.

| Metric       | Knorr-Held | Clust-Only | Clust-CAR | Clust-Conv |
|--------------|------------|------------|-----------|------------|
| **Georgia data** |            |            |           |            |
| DIC (n.eff)  | 6919.4 (125.0) | 7705.5 (4.0) | 7156.1 (373.3) | 7174.3 (380.1) |
| LMPL         | -3336.1    | -3848.8    | -3240.2   | -3246.8    |
| **England data** |            |            |           |            |
| DIC (n.eff)  | 101,435.4 (2339.9) | 190,608.7 (66.71) | 110,769 (9788.5) | 115,486.0 (9152.2) |
| LMPL         | -48566.6   | -95217.6   | -47,423.6 | -50,155.8 |
5.2 Case study 1 - Georgia colon cancer study

As shown in Table 2, the DIC and LMPL metrics differ in their assessment of the best fitting model, as the DIC is minimised by the Knorr-Held model while the LMPL is maximised by the Clust-CAR model. This difference is likely to be because the DIC penalises models with larger effective numbers of parameters, and the effective numbers here are 125 (Knorr-Held) and 373 (Clust-CAR) respectively. This reduction is due to the fact that the Knorr-Held model assumes a common spatial surface for all time periods (which is dominating the interaction terms for these data), where as the Clust-CAR model assumes a separate surface for all time points. Additionally, given that the data contain $N = 159$ spatial units and $T = 10$ time periods, these numbers suggest both models have induced substantial spatial and temporal smoothing on their respective random effects.

The posterior median risk surfaces for 2001 from the Knorr-Held and Clust-CAR models are displayed in the top row of Figure 3 and show substantial differences between the spatial patterns. The correlation between these risk surfaces is only 0.58, while their correlations with the raw SIR values are 0.50 (Knorr-Held) and 0.70 (respectively). The low correlation between the risks estimated from the Knorr-Held and Clust-CAR models is because the former assumes a common spatial pattern for all time points where as the Clust-CAR model does not, and the SIR data suggest a common spatial pattern is unrealistic. The figure also shows the estimated risks are substantially smoother than the raw SIR values, ranging between 0.591 and 1.912 compared with the SIR which ranges between 0 and 7.24. This suggests that the vast majority of the spatial pattern in the SIR was due to Poisson sampling variation as a result of small counts of cases, rather than true spatial variation in risk. It also suggests that there are no high-risk clusters in the data, a fact that is underlined by all three cluster models estimating only a single risk class for these data.

Therefore instead we illustrate the use of exceedence probabilities to visualise areas that exhibit small but elevated risks of disease compared with the baseline level. The bottom panel of Figure 3 shows the estimated posterior exceedence probability from the Clust-CAR model for $c = 1$ (left) and $c = 1.5$ (right), and illustrates two interesting points. Firstly, the right panel shows that only seven areas have a substantial probability of exceeding a risk of 1.5, and four of these consist of a cluster of areas in the north east of the state. This set of areas is the only ones in the study region that could be considered a cluster, but even then the level of elevated risk is low. Secondly, none of the maps in Figure 3 show any evidence of a relationship between the 18 Public Health Districts in Georgia and...
risk, suggesting that there does not appear to be a disparity between the PHDs that is effecting colon cancer risk.

5.3 Case study 2 - England cardiovascular study

As shown in Table 2, the DIC and LMPL metrics again differ in their assessment of the best fitting model, as the DIC is minimised by the Knorr-Held model while the LMPL is maximised by the Clust-CAR model. The preference for the Knorr-Held model is likely due to the consistency of the spatial risk pattern across the $T = 36$ time periods, and the vastly smaller number of effective parameters it uses. The main results for this study are displayed in Figure 4, where the top left panel displays the estimated risk surface for the first time period (2003 quarter 1). The figure shows the risk surface exhibits a similar spatial pattern to the raw SIR values in Figure 1, where high-risk clusters are again evident. The Clust-CAR model suggests there are three risk classes in the data, the baseline level containing 86% of the local authorities, an elevated risk class containing 13% of the areas and a highly elevated class containing just two areas with extreme risks. These three classes are shown for the first quarter in 2003 in the top right panel of Figure 4 and match up with the risk surface well as expected. The elevated areas are typically in the north of England, with most of the south being in the baseline risk class.

The bottom panel in Figure 4 displays the posterior median (solid line) and 95% credible intervals (dashed lines) for the average risk in each class (e.g. $\exp(\lambda_t, Z_{it})$), and the vertical line denotes the beginning of the smoking ban in July 2007. The figure shows that the baseline risk class exhibits very little temporal trend, with average risks of 0.912 and 0.920 for the first and last time period. In contrast, the other classes exhibit sizeable decreases in risk, with the elevated risk class dropping from 1.47 to 1.36 over the nine year period, while the highly elevated class drops from 2.02 to 1.83. These results suggest that the risk of cardiovascular disease has indeed dropped for the high-risk groups over the study period, but that the decline appears to be starting in 2010, two and a half years after the smoking ban. This reduction may be a lagged effect of the smoking ban, and further research is required before a causal link can be established.

6 Discussion

This paper has proposed one of the first models for simultaneously estimating disease risk and identifying high-risk clusters, and is accompanied by freely available software. Existing approaches either
Figure 3: The maps in the top row display the estimated risks for 2001 from the Knorr-Held and Clust-CAR models, while the bottom row shows estimated exceedence probabilities from the Clust-CAR model for $c = 1$ (left) and $c = 1.5$ (right).
Figure 4: Maps showing the posterior median risk (top left) and the posterior median risk class (top right) for 2003 quarter 1, where for the latter light grey areas are in the baseline risk class, while the mid grey and black areas are in the elevated and highly elevated classes. The Easting and Northing coordinates are in kilometres. The bottom panel displays the temporal trend in the mean of these three risk classes over time.
estimate a spatio-temporal smoothed risk surface without cluster detection, or utilise testing based paradigms to identify clusters without estimating the risk surface for the entire study region. Some work has combined these facets into a unified model in a purely spatial context, but the extension to a spatio-temporal setting throws up a number of additional modelling challenges that we have addressed here.

A number of general themes emerged from our simulation study. First, all models perform better as disease prevalence increases (increasing $e_{it}$), both in terms of risk estimation (via RMSE) and the ability to correctly detect elevated risk clusters (via the Rand index). This is because the assumed mean of $y_{it}$ is $\theta_{it}e_{it}$, so when $e_{it}$ is small Poisson sampling variability dominates any true spatio-temporal variation in $\theta_{it}$. Second, the model of Knorr-Held (2000) performs better in terms of risk estimation than the cluster models if the data exhibit a common spatial pattern for all time periods (scenarios 1 to 3), because that is the main assumption underlying that model. Conversely, when the risk structure is highly non-separable (scenarios 4 and 5), then the cluster models exhibit improved estimation. Third, a post-hoc approach to clustering performs uniformly poorly throughout, whereas the clustering models typically perform well even though the number of classes, $G$, is not fixed at the same as the number in the data. All clustering models perform uniformly well when the data contain either no or high-risk clusters, while the Clust-Only model performs best if the clusters are only elevated slightly above the baseline risk. This is because the two components in the linear predictors of the Clust-CAR and Clust-Conv models are competing for the variation in the data, leading to the clustering structure sometimes being wrongly captured by the spatial autocorrelation component.

In common with the disease mapping literature, our clustering models do not include covariates, so that the risk classes relate to the risk in an area and not the residual risk after adjusting for the covariates. This allows us to classify groups of areas at elevated risk, so public health interventions can be appropriately targeted. However, classifying the residual risk surface after adjusting for covariates would aid in the identification of unknown etiologic covariates, as the residual class structure could be examined to look for similarity with spatially varying covariates. Covariate information could also be used as predictors for the cluster allocation model ($Z_{it}$), and both will be investigated in future work.

Another avenue for future work is illustrated by the simulation and Georgia case studies, which show the difficulties in cluster detection and risk estimation for rare diseases. Both the Knorr-Held
and cluster models smooth the extreme raw SIR values towards the null risk of one, as the spatial smoothing dominates the SIR data values. Thus further work is required to develop spatio-temporal disease mapping models specifically designed for low count data $y_{it}$. Another future direction will extend the cluster models to consider multiple diseases simultaneously, which throws up additional modelling challenges. One key challenge will be how to extend the discrete Markov model for $Z_{it}$, so that it can account for between disease correlations.

References

Anderson, C., D. Lee, and N. Dean (2014). Identifying clusters in Bayesian disease mapping. *Biostatistics* 15, 457–469.

Bernardinelli, L., D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine* 14, 2433–2443.

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

Best, N., S. Richardson, and A. Thomson (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 14, 35–59.

Charras-Garrido, M., D. Abrial, and J. de Goer (2012). Classification method for disease risk mapping based on discrete hidden Markov random fields. *Biostatistics* 13, 241–255.

Charras-Garrido, M., L. Azizi, F. Forbes, S. Doyle, N. Peyrard, and D. Abrial (2013). On the difficulty to delimit disease risk hot spots. *Journal of Applied Earth Observation and Geoinformation* 22, 99–105.

Choi, J., and A. B. Lawson (2011). Evaluation of bayesian spatial-temporal latent models in small area health data. *Environmetrics* 22, 1008–1022.

Congdon, P. (2005). *Bayesian models for categorical data* (1st ed.). John Wiley and Sons.

Diggle, P., B. Rowlingson, and T. Su (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* 16, 423–434.

Forbes, F., M. Charras-Garrido, L. Azizi, S. Doyle, and D. Abrial (2013). Spatial risk mapping for rare disease with hidden Markov fields and variational EM. *Annals of Applied Statistics* 7, 1192–1216.
Fraley, C. and A. Raftery (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification 24*, 155–181.

Gangnon, R. and M. Clayton (2000). Bayesian detection and modeling of spatial disease clustering. *Biometrics 56*, 922–935.

Green, P. and S. Richardson (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association 97*, 1055–1070.

Kelsall, J. and J. Wakefield (2002). Modeling Spatial Variation in Disease Risk. *Journal of the American Statistical Association 97*, 692–701.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine 19*, 2555–2567.

Knorr-Held, L. and G. Rasser (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics 56*, 13–21.

Kulldorff, M., R. Heffernan, J. Hartman, R. Assuncao, and F. Mostashari (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine 2*, 216–224.

Lawson, A. B. (2013). *Bayesian Disease Mapping: hierarchical modeling in spatial epidemiology* (2 ed.). New York: CRC Press.

Lawson, A. B., J. Choi, B. Cai, M. M. Hossain, R. Kirby, and J. Liu (2012). Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data. *Journal of Agricultural, Biological and Environmental Statistics 17*, 417–441.

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

Leroux, B., X. Lei, and N. Breslow (1999). *Estimation of disease rates in small areas: A new mixed model for spatial dependence*. *Statistical Models in Epidemiology, the Environment and Clinical Trials, Halloran, M and Berry, D (eds)*, pp. 135–178. Springer-Verlag, New York.

Li, G., N. Best, A. Hansell, I. Ahmed, and S. Richardson (2012). BaySTDetect: detecting unusual temporal patterns in small area data via bayesian model choice. *Biostatistics 13*, 695–710.

Office for National Statistics (2014). Inequality in healthy life expectancy at birth by national deciles of area deprivation: England, 2009-11.
Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society Series B* 52, 175–184.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rand, W. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* 66, 846–850.

Richardson, S., A. Thomson, N. Best, and P. Elliott (2004). Interpreting Posterior Relative Risk Estimates in Disease Mapping Studies. *Environmental Health Perspectives* 112, 1016–1025.

Spiegelhalter, D., N. Best, B. Carlin, and A. Van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* 64, 583–639.

Ugarte, M., T. Goicoa, and A. Militino (2010). Spatio-temporal modelling of mortality risks using penalized splines. *Environmetrics* 21, 270–289.

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

Wakefield, J. and A. Kim (2013). A Bayesian model for cluster detection. *Biostatistics* 14, 752–765.

Wall, M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference* 121, 311–324.

Young, L., C. Gotway, J. Yang, G. Kearney, and C. DuClos (2009). Linking health and environmental data in geographical analysis: It’s so much more than centroids. *Spatial Spatio-temporal Epidemiology* 1, 73–84.