特集：疾病地図の推定，疾病の集積性の検出に関する最近の進展

Analysis of small area disease clustering using RJMCMC

Andrew B. LAWSON and Allan B. CLARK

This paper describes a variety of approaches to the analysis of clustering in small area health data. These approaches involve the modelling of clusters in boty shape and size and the approaches use full statistical models in a Bayesian setting, rather than statistical tests.

Included in the review is an account of both case event and count data clustering as well as extensions to space-time clusters.

1 Introduction

The analysis of small area data where the location of residence are known is well documented[24]. However, most of the model based analysis has been based on either clustering around fixed foci[17] or on description of the general pattern of disease[3),7)]. While some good tests of clustering have appeared[33] many of these are not well understood, and are restricted to suggesting cluster locations which are actual events.

The model suggested in this paper can accomplish both of these tasks. In comparison to the Markov Random Field models[3),7)] it does not require the specification of arbitrary neighbourhood structures. In comparison to the clustering detection methods[33] it is not restricted to estimating cluster locations as events. A review of these is provided in Lawson et al[24].

The analysis of clustering in small area health data has attracted increased interest in recent years, see Lawson et al[24]. Both public concern for the existence of 'clusters' of disease and growing interest in the causes of clustering, per se, are partly responsible for this increase. A growing interest in environmental issues both in the general public and the scientific community, has led to interest in clusters related to environmental hazards, e.g. power stations, incinerators, electro-magnetic fields or toxic waste dumping sites. The analysis of clustering in small area health data can be approached in a variety of ways, depending on the purpose of the study.

Two fundamental considerations should first be assessed:

1) Is the clustering in the data of primary interest?

2) Is the clustering of secondary interest, and hence, a nuisance feature?

University of Aberdeen, U.K

In the first case, some detailed aspects of clustering may be of interest, e.g. How likely are n clusters? What is the marginal posterior distribution of the centres of clustering, given n clusters? Are there different scales of clustering supported by the data?

In the second case, clustering tendency is to be estimated, perhaps as part of a background feature of the process but other aspects of the disease process are of major interest. For example, some diseases are known to form clusters at certain scales (e.g. Leukaemias[8)]), but the relation of the disease incidence to putative sources of hazard may be of prime interest. Hence, in this case, clustering is a 'nuisance' background characteristic. A review of these is provided in Lawson and Kulldorff[27].

In this paper, an approach to the analysis of clustering in small area health data is proposed, which can accommodate both of the above cases, via direct modelling of clustering within a more general model framework. The methods used are primarily Bayesian, as considerable use is made of MCMC methods. The methods have considerable generality and can be applied to both case event and counts of cases in arbitrarily-defined regions.

2 Model Development

The data $\mathbf{y}$ and cluster centres $\mathbf{x}$ are spatial point patterns:

$$\mathbf{y}=\{y_1,\cdots\cdots y_m\},\ m>0,\ y_i\in T$$
$$\mathbf{x}=\{x_1,\cdots\cdots x_n\},\ n\geq 0,\ x_i\in U$$

where T is the study window and U is a region which encloses T. By allowing U to differ from T, we allow the possibility of locating putative cluster centres outside the window of observation of the data. This makes some allowance for the edge effect where data could appear in the window but a centre lies outside, i.e. the boundary splits a cluster so that some part of the form is censored. The observed data y in this paper are address locations of cases of disease, observed within T and a fixed time

period. Diseases of interest could be leukaemias, which are thought to cluster weakly[8], or possibly, respiratory disease, such as respiratory cancer, larynx cancer or bronchitis, which could relate to one or more sources of health hazard (e.g. incinerators, waste dump sites etc.). In either case, unobserved heterogeneity in the environment and/or population experiencing the disease events could lead to clustered disease incidence over the window T.

In any analysis of y, the population experiencing the disease events must be considered. The variation of population over space, in its density *and* its propensity to contract a disease (its 'at-risk' structure) can lead to apparent 'clustering' or 'heterogeneity' in y. Hence, to properly assess clustering in such data it is important to account for the spatial variation in the 'at-risk' structure of the population. To achieve this, a variety of approaches can be adopted. The commonest approach is to estimate the 'at-risk' surface either, from the known features of the population, such as age-sex structure or measurements of deprivation or life-style information. These data are usually available for small areas, from national censuses. Or alternatively to use a 'control' disease. The first approach is often termed 'standardisation', when applied to count data in census tracts (see e.g.[16]). It has been applied to the assessment of single 'cluster' of case event data (see[29]). The second approach can be applied where a 'control' disease can be chosen which has a similar 'at-risk' structure to the case disease, but is not known to display clustering. This approach has been used by a variety of workers (see e.g.[29]) to examine possible 'clusters' around putative sources of health hazard. In these cases, the control should not be known to be affected by the hazard, and hence should not 'cluster' near possible sources. In the general clustering case, where no specific environmental cause or factor is hypothesised and can be measured, then the 'control' disease should be known to be free from a clustering tendency.

In what follows we represent this modelling approach by using the first order intensity of the process, the most general from of which is:

$$\lambda(\mathbf{y}|\mathbf{x}) = g(\boldsymbol{y}).\ (1+\sum_{i=1}^{n_x} \mu_i.\ h_i\{\mathbf{y}-x_i;\kappa_i\}).\ \prod_{k=n_x+1}^{n}(1+f(\mathbf{y}-x_k)) \quad (1)$$

In all the applications we examine, $g(\mathbf{y})$ is considered to represent the background 'at-risk' process, the μ_i represents the mean number of points in centre i, the κ_i represents the dispersion around the ith centre, h_i is the cluster distribution function for the ith centre and f is a cluster distribution function for known foci.

Note that we assume a multiplicative link between the population background and cluster distribution functions which implies that any spatial structure modelled in cluster distribution functions will be directly modified by variations in $g(\mathbf{y})$. The alternative of a pure additive link (see e.g.,[5] p142), would imply that spatial structures modelled in the clustering were of fixed size and hence unaffected by the population structure. This would appear to be inappropriate for spatial epidemiological data.

2.1 Prior Distributions and Cluster Structure

In the case of non-focussed clustering, prior distributions must be provided for the components n_x, $\mathbf{x}$ and parameters in $h(\mathbf{y}-\mathbf{x})$. Typically, the number of centres is assumed to have a Poisson (ρ) distribution, while $\mathbf{x}$ could follow a homogeneous Poisson process. However, it is possible to specify joint prior distributions for these parameters, e.g. Strauss distributions. Previous work[20),35)] discusses the theoretical justification for this in non-modulated cluster processes and Cox processes. Alternative specifications for the $\mathbf{x}$ prior distribution (e.g. a Markov inhibition process) can be suggested based on algorithmic considerations (see section 4). The cluster distribution function can take a variety of forms. A commonly used form is

$$h(\mathbf{y}-\mathbf{x}) = \frac{1}{2\pi\kappa} e^{-[||\mathbf{y}-\mathbf{x}||]^2/2\kappa} \quad (2)$$

a radially isotropic Gaussian form with cluster variance κ. However, alternative forms are possible, including non-parametric versions. For example,

$$\hat{h}(\mathbf{y}-\mathbf{x}) = \frac{1}{n_x m h_1 h_2} \sum_{i=1}^{m} k\left(\frac{\mathbf{y}-y_j}{h_1}\right) \cdot \sum_{j=1}^{n_x} k\left(\frac{\mathbf{y}-x_j}{h_2}\right) \quad (3)$$

where $k(.)$ is a kernel function (see e.g.[31]), provides a nonparametric estimator for h.

The possibility of allowing a flexible cluster shape, via density estimation, may be attractive in situations where the exact form of clusters cannot be parameterised. This allows a considerable latitude in the definition of the cluster form while retaining a general model paradigm.

Further alternative approaches can be proposed which provide great flexibility for cluster modelling:

- the use of $\{\mu_i, \kappa_i\}$ for each cluster allows variation in sizes across the field (as in a conventional mixture problem)
- the use of a spatially dependent $\kappa(\mathbf{x})$, cluster variance, can be employed (see e.g.[19]).

Note that in the second alternative, a spatial Gaussian

random field prior distribution can be assumed for $\kappa(\mathbf{x})$. Typically this would be *MVN* $(F\alpha, \sigma^2 I)$, where F is a spatially-dependent design matrix. This allows smooth variation over the study region for the cluster variance. This is considerably more parsimonious than the first alternative.

3 Algorithms

The development of Markov Chain Monte Carlo (MCMC) methods and other iterative simulation tools[34),2)] has allowed the implementation of algorithms which can explore posterior distributions of the spatial problems identified above. Note that for both case and count data, if foci are known then straightforward likelihood models, or conventional aspatial Bayesian models can be applied (see e.g.[6),18),19),23),22)]). If the locations of foci are unknown, then spatial prior distributions must be invoked.

3.1 Incorporation of Background Risk

Before considering the detail of basic algorithms, it is important to examine the issue of how the background risk surface ($g(\mathbf{y})$), can be incorporated in these problems. So far two basic approaches have been proposed:

3.1.1 Profile Likelihood

In early work on case events[29)], the function $g(\mathbf{y})$ was estimated nonparametrically, and inference was made conditional on the fixed of $\hat{g}(\mathbf{y})$, without regard to estimation errors inherent in $\hat{g}(\mathbf{y})$. Diggle[11)] proposed the use of a control disease case event map to provide a density estimate of $g(\mathbf{y})$, while Lawson and Williams[29)] compared the use of a control disease, and expected deaths, to estimate $g(\mathbf{y})$. Both approaches require smoothing of the background risk based on *external* data. An alternative hybrid model which used expected deaths in regions directly was also proposed by[29)]. The use of control disease event maps has a number of disadvantages. In particular, the matching of a control disease to the 'at-risk' group of the cases, while being unrelated to the effect under study, can be difficult. Indeed, Diggle[11)] provides an example of a control disease (respiratory cancer) which is related to the effect under study (air ollution). The use of expected deaths does not suffer from this problem but is usually only available at an aggregated level (e.g. census tract). Other background factors should also be incorporated where known, e.g. deprivation indices. However, these are usually only available at tract level also.

3.1.2 Label Modelling

For the special case where a control disease is used, it is possible to use a *bivariate* point process model which directly incorporates the control event locations in the model. By conditioning on the locations of cases and controls, then it is possible to directly model the mark labels on the events (see e.g.[1)]), and thereby the window (T) becomes irrelevant to the inference. In addition, this conditioning can be used to avoid esimation of $g(\mathbf{y})$. Diggle and Rowlingson[12)] (DR) suggested the use of this approach in focussed clustering problems. This leads to a logistic regression formulation of the problem. Note, also that, as a special case of a Markov point process, conditional on the locations, the labels form a binary Markov random field[1)] and the auto-logistic model results. Assuming an Ising model, standard logistic regression methods apply.

3.1.3 A Bayesian Smoothing Model

The label modelling approach above, is only applicable when an appropriate control disease is available. To keep the model approach general, it is important to pursue methods which are not limited to such a specific case. An alternative approach is to regard the smoothing operator in $\hat{g}(\mathbf{y})$ as a sample realisation of possible smoothing values which have a prior distribution. This both allows the incorporation of $\hat{g}(\mathbf{y})$ within the estmation process and allows the exploration of the variation in $\hat{g}(\mathbf{y})$ *in relation* to the other parameter dimensions. This is discussed further in section 3.3.3.

3.2 Basic Point Event Algorithm

In the most general Bayesian formulation of the cluster model, we define the joint posterior distribution of $\{\mathbf{x}, \boldsymbol{\theta}\}$ as

$$P(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\mathbf{x}).\ p(\mathbf{x}).\ g(\boldsymbol{\theta}) \tag{4}$$

where

$$L(\mathbf{y}|\mathbf{x}) = \{\prod_{i=1}^{m} \lambda(y_i|\mathbf{x})\},\ \exp\{-\int_T \lambda(\mathbf{u}|\mathbf{x})\, d\mathbf{u}\} \tag{5}$$

$p(\mathbf{x})$≡prior distribution for $\mathbf{x}$ (Markov inhibition or Uniform) and n_x (Poisson (ρ)),

$g(\boldsymbol{\theta})$≡prior distributions for cluster function parameters.

$$\lambda(y_i|\mathbf{x}) = g(y_i).\ (1+\sum_{j=1}^{n_x} h(y_i - x_j)).\ \prod_{k=n_x+1}^{n} (1+f(y_i - x_k)) \tag{6}$$

where there are n_x unknown and $n-n_x$ known foci. Note that the final fixed-foci term of (6) could also be dependent on covariates related to the individual observations $\{y_i\}$ or random effects[21)].

3.3 The Basic Algorithm

It is convenient to define three sets of parameters for the purpose of the algorithm steps. These sets can be considered as separate components of the sampler design. A two stage sampler proceeds by considering

spatial cluster parameters within an inner iterative sampler conditional on current values of other parameters. The three nested sampling schemes are:

1) spatial cluster (sc) parameters: $\mathbf{x}$, η_x

2) non-spatial (nc) parameters: for example ρ, κ and μ (assuming a Gaussian cluster distribution is used)

3) smoothing parameter(s): for example $g(\mathbf{y})$ may depend on s (a smoothing parameter).

3.3.1 SC parameters

The derivation and properties of the following algorithm are discussed in[20] and[35]. The posterior distribution (4) could be explored by conventional iterative simulation methods, except for the cluster term, where a summation with a random upper limit occurs. This is essentially a mixture problem, and the sc parameters in this problem are best explored by a reversible jump Metropolis-Hastings (MH) sampler[13),14)] involving a mixture kernel. Essentially the joint distribution of $\mathbf{x}$ and η_x must be explored during iteration. This can be achieved by a spatial-birth-death-shift (SBDS) algorithm, where centres are added, deleted or shifted with given probability. A sequence of likelihood ratios can be specified for each case. In general, for a new configuration $\mathbf{x}'$, the posterior density ratio is, conditional on nc and h parameters:

$$PR(\mathbf{x}, \mathbf{x}')=\frac{L(\mathbf{y}|\mathbf{x}')}{L(\mathbf{y}|\mathbf{x})}\cdot\frac{p(\mathbf{x}')}{p(\mathbf{x})} \tag{7}$$

This ratio is evaluated for x′ within the SBDS algorithm based on an MH criterion. A proposal configuration x′ is accepted with probability

$$A(\mathbf{x}, \mathbf{x}')=\min\left\{1,\ PR(\mathbf{x}, \mathbf{x}').\ \frac{q(\mathbf{x}', \mathbf{x})}{q(\mathbf{x}, \mathbf{x}')}\right\} \tag{8}$$

where $q(\mathbf{x}', \mathbf{x})$ is the proposal distribution for the new state. Often the proposal distribution for a point $\mathbf{u}$ is defined as a function of $h(\mathbf{y}-\mathbf{u})$ itself (e.g. $\frac{1}{m}\sum_{i=1}^{m}h(\mathbf{y}_i-\mathbf{u})$) as simpler uniform proposals can lead to high rejection rates. We use Markov inhibition priors for $\mathbf{x}$, as the peaked nature of the likelihood surface can lead to multiple response, and it is important to propose spatially-separate new $\mathbf{x}$ values to avoid this problem. To this end, the Strauss prior can be used, and is defined for the proposed addition of a point $\mathbf{u}$ as

$$\frac{p(\mathbf{x}\cup\mathbf{u})}{p(\mathbf{x})}=\beta\gamma^{n_R(\mathbf{u})} \tag{9}$$

where β and $0<\gamma<1$ are parameters and $n_R(\mathbf{u})$ counts the number of $\mathbf{x}$ within a distance R of $\mathbf{u}$. Similar ratios can be defined for deaths and shifts.

For the likelihood, (5), the likelihood ratios are:
for addition:

$$\prod_{i=1}^{m}\left[1+\frac{h(y_i-\mathbf{u})}{1+\sum_{j=1}^{n_x}h(y_i-x_j)}\right].\ e^{-\Lambda(T|\mathbf{u})} \tag{10}$$

for deletion:

$$\prod_{i=1}^{m}\left[1-\frac{h(y_i-\mathbf{x}_d)}{1+\sum_{j=1}^{n_x}h(y_i-x_j)}\right].\ e^{\Lambda(T|\mathbf{x}_d)} \tag{11}$$

where $\mathbf{x}_d$ is the point to be deleted;
for shifting:

$$\prod_{i=1}^{m}\left[1+\frac{h(y_i-\mathbf{u})-h(y_i-\mathbf{x}_d)}{1+\sum_{j=1}^{n_x}h(y_i-x_j)}\right].\ e^{\{\Lambda(T|\mathbf{x}_d)-\Lambda(T|\mathbf{u})\}} \tag{12}$$

where

$$\Lambda(T|\mathbf{x})=\int_T\lambda(\mathbf{u}|\mathbf{x})d\mathbf{u}. \tag{13}$$

Note also that it is usual to include a constant rate scale parameter in $\lambda(\mathbf{u}|\mathbf{x})$ (δ say). However, it is possible to condition out δ from the analysis, and this can reduce the parameter dimensionality of the algorithm. To do this the likelihood ratio in (7), can be written in the form, conditional on m :

$$\prod_{i=1}^{m}\left\{\frac{1+\sum^{\mathbf{x}'}h(\mathbf{y}_i;\mathbf{x}')}{1+\sum^{\mathbf{x}}h(\mathbf{y}_i;\mathbf{x})}\right\}\cdot\left\{\frac{\Lambda_{\mathbf{x}}(T|\mathbf{x})}{\Lambda_{\mathbf{x}'}(T|\mathbf{x}')}\right\}^{m}, \tag{14}$$

where δ is removed from $\lambda(\mathbf{u}|\mathbf{x})$ and $\Lambda(.)$. It is straightforward to derive the equivalent ratios to (10)-(12), for this case.

3.3.2 NC parameters

The parameters of the cluster distribution function, and other prior distributions can be treated conventionally. In most cases here, we assume that n_x has a Poisson (ρ) prior distribution. This parallels the assumptions which specify a Poisson Cluster Process in ordinary point process models[10),19)]. It is also possible to assume a prior distribution for ρ, and a Gamma distribution is often used. We have no strong prior reason to assume any other distribution than a uniform indifference prior on a suitable range (usually $\leq m$).

The cluster distribution parameters (μ, κ), based on model (2), are also assumed to have uniform indifference priors. The sampler steps used for ρ, μ and κ differ depending on whether a Gibbs or MH step is simple to implement. A Gibbs step is straightforward for ρ, whereas to implement a Gibbs step for κ or μ requires an optimisation step (to obtain ml estimates), and in these cases an MH step is used.

3.3.3 Smoothing Parameters

The function $g(y)$ In previous work $g(y)$ has been esti-

mated nonparametrically, or has been conditioned out of the analysis (see section 3.1). It is possible to incorporate the estimation of $g(y)$ within the MCMC algorithm. We regard the smoothing parameter (s) of a smoothing operation which estimates $g(y)$, as a random quantity. We can update the parameter s just like any other parameter in the MCMC algorithm. In order to specify the Bayesian model completely we need to define a prior distribution for s. A natural choice, motivated by work on Gaussian Mixtures[9], is the inverse gamma distribution.

$$p(s|\alpha,\ \beta)=\frac{\alpha^{\beta}}{\Gamma(\beta)}\ s^{-\beta-1}\exp\left(-\frac{\alpha}{s}\right) \tag{15}$$

The hyperparameters (α, β) can be assigned non-informative hyperpriors. This approach is discussed further in Lawson and Clark[26].

Cluster distribution smoothing If the cluster distribution function is estimated nonparametrically as in (3), then the appropriate smoothing constants (h_1, h_2) must also be included in the iterative estimation scheme. A procedure similar to that suggested in (15) can be used and replaces the steps for μ and κ.

3.4 The Label Modelling Algorithm

The alternative to direct estimation of $g(\mathbf{y})$ in profile likelihood or the Bayesian smoothing model, is the label modelling (DR) approach. In this case we can define a conditional probability of a case event at $\mathbf{y}$, out of a case/control disease bivariate realisation, as :

$$P(\mathbf{y})=\frac{g(\mathbf{y})\delta.(1+\sum_{j=1}^{n_x}h(\mathbf{y}-x_j)).\prod_{i=1+n_x}^{n}(1+f(d_{yx_i}))}{g(\mathbf{y})+g(\mathbf{y}).\delta(1+\sum_{j=1}^{n_x}h(\mathbf{y}-x_j)).\prod_{i=1+n_x}^{n}(1+f(d_{yx_i}))}$$
$$=\frac{\delta.(1+\sum_{j=1}^{n_x}h(\mathbf{y}-x_j)).\prod_{i=1+n_x}^{n}(1+f(d_{yx_i}))}{1+.\delta(1+\sum_{j=1}^{n_x}h(\mathbf{y}-x_j)).\prod_{i=1+n_x}^{n}(1+f(d_{yx_i}))} \tag{16}$$

Where d_{yx_i} , denotes the distance from $\mathbf{y}$ to the ith cluster centre. Note that this conditional model can be derived from a full bivariate competing risk model for cases and controls. i.e.

$$\Pr\ (\text{case at } \mathbf{y})=\lambda_1(\mathbf{y}|\mathbf{x}).e^{-\int_T\Sigma\ell\lambda_i(\mathbf{u}|\mathbf{x})d\mathbf{u}}$$
$$\Pr\ (\text{event at } \mathbf{y})=\sum_{i}^{2}\lambda_i(\mathbf{y}|\mathbf{x}).e^{-\int_T\Sigma\ell\lambda_i(\mathbf{u}|\mathbf{x})d\mathbf{u}}$$

where λ_i is the intensity of the relevant effect (λ_1 for cases) (see e.g.[28]).

The likelihood of m cases and n controls is

$$L=\prod_{i=1}^{m}P(y_i).\prod_{j=m+1}^{m+n}(1-P(y_i)) \tag{17}$$

This conditional approach can also be used to replace $L(\mathbf{y}|\mathbf{x})$ in (5) by (17) within the main cluster algorithm. This leads to comparable ratios for the SBDS algorithm.

The nc parameter sampler can be constructed as for the basic algorithm and it is also possible to use a non-parametric estimate of $h(\mathbf{y}-\mathbf{x})$ in this situation (see e.g.[25]).

4 Count Modelling

Small area data is often available only as counts of cases wihtin arbitray regions (usually census tracts). Hence, a considerable literature has grown around the analysis of such data. While methods have been developed to test for global (e.g.[30),36]) and focussed culstering of counts (e.g.[15),32),4),17]) little attention has been paid to the modelling of non-focussed clustering of count data. The methods applied to case event data can be applied here. Given the conditional independence assumption, then the counts in disjoint regions (say n_i) are independently Poisson distributed with integrated intensity given by

$$\Lambda(A_i|\mathbf{x})=\int_{A_i}g(\mathbf{u}).(1+\sum_{i=1}^{n_c}\mu_i.\ h_i\{\mathbf{u}-x_i;\kappa_i\})$$
$$\prod_{k=n_x+1}^{n}(1+f(\mathbf{u}-x_k))d\mathbf{u}$$

where A_i denotes the ith region. If we only observe the events in specific subregions then it is likely that we shall only observe the controls at that scale. However, hybrid algorithms are available when we observe the processes on different scales[29]. Denote the control count (usually called the expected count) in the ith region by m_i. With this notation we make the approximation

$$\Lambda(A_i|\mathbf{x})=m_i.\int_{A_i}(1+\sum_{i=1}^{n_c}\mu_i.\ h_i\{\mathbf{u}-x_i;\kappa_i\})$$
$$\prod_{k=n_x+1}^{n}(1+f(\mathbf{u}-x_k))d\mathbf{u}$$

Conditional on N, the total number of cases (i.e. $N=\sum_{i=1}^{p}n_i$), the likelihood for p regions is

$$L(\mathbf{n}|\mathbf{x},\boldsymbol{\theta})=\prod_{i=1}^{p}\left[\frac{\Lambda(A_i|\mathbf{x})}{\sum_{i=1}^{p}\Lambda(A_i|\mathbf{x})}\right]^{n_i} \tag{18}$$

We can directly use the basic point process algorithms and replace the likelihood ratios with those based on (18). The sampling algorithms can be modified to accommodate this case.

5 Spatio-temporal Clustering

So far we have only discussed clustering in space. In this section we shall extend the above model to deal with space-time (spatio-temporal) clustering. In what follows we describe three different types of clustering each being indetified by its persistence properties. The

analysis of spatio-temporal data is more complicated than the spatial case, with edge effects arising in a number of different forms, and a larger variety of possible sampling spproaches which could lead to different modelling strategies.

The edge effect problem can take a variety forms due to possible lead times in detecting a disease, e.g. breast cancer. This of course has a consequence for defining the temporal location and how coarse one can split time up. If the outcome is death then, of course, this problem does not exist. Alternatively, there may be cases occurring outside the spatial study region during the temporal study period and these will be censored.

Putting these problems aside for the moment, we can imagine a number of possible clustering structures. The clustering structures that are included in our model are

1. The disease may cluster in time throughout the whole spatial region. We shall call this a temporal cluster
2. The disease may cluster in space throughout the whole temporal study period. We shall call this a spatial cluster
3. The disease may cluster locally in both time and space. We shall call this a spatio-temporal cluster

Each of these can be seen in figure 1.

We shall assume that these cluster terms can be separably defined, and so we define three different and additive functions of the three cluster centre types. We assume an intensity function of the form

$$\lambda(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \boldsymbol{\theta}) = g(\mathbf{y}).\ \{1 + \alpha_1 \sum_{i=1}^{nsc} h_1(\mathbf{y}^s - x_{1i}) + \alpha_2 \sum_{i=1}^{ntc} h_2(\mathbf{y}^t - x_{2i}) + \alpha_3 \sum_{i=1}^{nstc} h_3(\mathbf{y} - x_{3i})\}$$

where y^s is the spatial coordinate of $\mathbf{y}$, y^t is the temporal coordinate of $\mathbf{y}$, $\mathbf{x}_1 = \{x_{1i}\}_{i=1}^{nsc}$ are the spatial cluster centres, $\mathbf{x}_2 = \{x_{2i}\}_{i=1}^{ntc}$ are the temporal cluster centres and $\mathbf{x}_3 = \{x_{3i}\}_{i=1}^{nstc}$ are the spatio-temporal cluster centres and $\boldsymbol{\theta}$ is a vector of parameters that specify the cluster distribution functions (h_1, h_2 and h_3). A series of weights $\{\alpha_1, \alpha_2, \alpha_3\}$ are also included within the formulation, to allow for different contributions from the different components.

The basic point event algorithms can be applied to this intensity with only minor modifications.

5.1 Spatio-Temporal Example : Scottish Birth Abnormalities

We examined the application of the sampler to the distribution of congenital birth abnormalities within post code units in the Tayside region of central Scotland for the period 1991-1995. The distribution of the total abnormalities is not orderly within post code units and multiple events can occur, and so we resort to a count model (18) for the unit spatio-temporal area. We have used total births for the post code sectors as a control for the population background.

The temporal standardized mortality ratio (SMR) is plotted in figure 2. This clearly suggests that the rate of birth abnormalities is varying with time with two large peaks after 18 months and 53 months.

The spatial SMR is plotted is figure 3. The figure indicates a large amount of clustering in the centre of the map and is suggestive of some spatial trend.

We fitted the model and the results (not shown) indicate two temporal clusters at months 15 and 53, and one spatial cluster centre in the south-east. A nonparametric estimate of the rate in space-time is given in figure 4 and indicates some localized temporal clustering.

6 Conclusions

In this paper we have described a wide range of

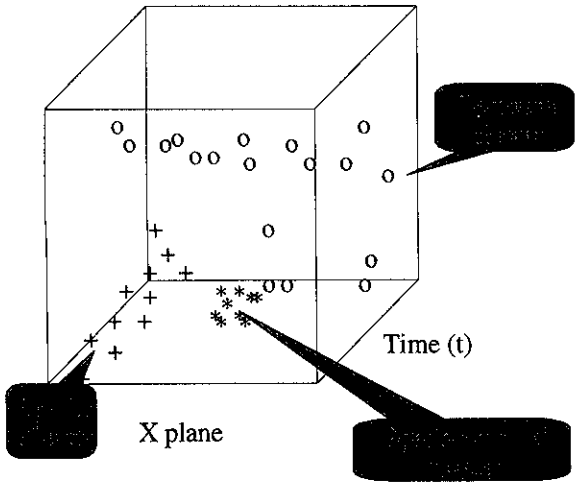

Figure 1

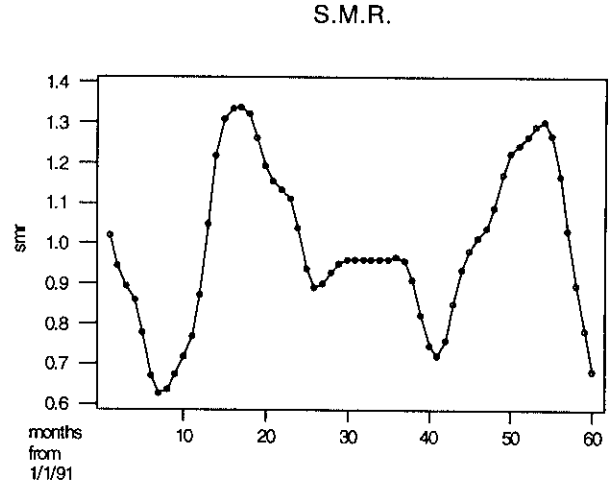

Figure 2

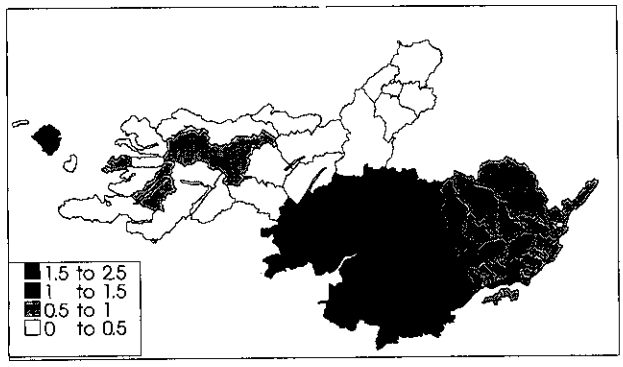

Figure 3

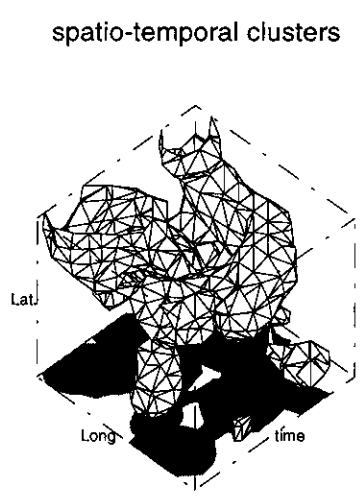

Figure 4

approaches to cluster modelling, all of which depend on the use of RJMCMC to allow the exploration of the joint posterior distribution of $\{\eta_x, x_i\}$, the number and locations of the cluster centres. A wide range of applications have been discussed and we have demonstrated the flexibility of the approach to cluster detection. These methods are computer intensive, but given the recent advances in computer technology we do not fell that this is a major drawback.

We have demonstrated the model on a spatio-temporal data set in Scotland. It is hoped that we can include covariates in this model to allow for trend and differing levels of deprivation. The results have been promising and reflect the general spplicability of the approach.

While one can never replace a carefully designed case-control study with observational data, the development of realistic models for clustering can suggest the need for further study. The inclusion of covariates into such models is of prime importance, indeed in our formulation the cluster terms represent unmeasured covariates. However, covariates are usually hard to obtain and one often has to result to crude measures such as Kafadar-Tukey urbanization index or deprivation indices as surrogates. It is hoped that this information will be available in the future.

The model presented here may have uses in the development of a spacetime surveillance alarm system. The divelopment of such a system would be of major public health importance and is a long term research aim.

References

1) A. Baddely and J. Møller. Nearest-neighbour Markov point processed and random sets. *International Statistical Review*, 57: 89-121, 1989.
2) J. Besag and P. J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, 55: 25-37, 1993.
3) J. Besag, J. York, and A. Molli'e. Bayesian image estoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43: 1-59, 1991.
4) J. Bithell and R. Stone. On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. *Journal of Epidemiology and Community Health*, 43: 79-85, 1989.
5) N. Breslow and N. Day. *Statistical Methods in Cancer Research, volume 2 : The design and analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon, 1987.
6) D. Clayton and L. Bernardinelli. Bayesian methods for mapping disease risk. In P. Elliott, J. Cuzick, D. English, and R. Stern, editors, *Geographical and Environmental Epidemiology : Methods for Small-Area Studies*. Oxford University Press, 1992.
7) D.G. Clayton and J. Kaldor. Empirical Bayes estimates of agestandardised relative risks for use in disease mapping. *Biometrics*, 43: 671-691, 1987.
8) J. Cuzick and M. Hills. Clustering and clusters-summary. In G. Draper, editor, *Geographical epidemiology of chidhood leukaemia and non-hodgkin lymphomas in Great Britain 1996-1983*, pages 123-125. HMSO, London, 1991.
9) J. Diebolt and C. Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society B*, 56: 363-375, 1994
10) P. Diggle. *Statistical Analysis of Spatial Point Patterns*. Academic Press, London, 1983.
11) P. Diggle A Point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society A*, 153: 340-363, 1990.
12) P. Diggle and B. Rowlingson A conditional approach to point process modelling of raised incidence *Journal of the Royal Statistical Society A*, 157: 433-440, 1994.
13) C. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Jour Statist.*, 21: 84-88, 1994.
14) P. J. Green. Reversible jump MCMC computation and

Bayesian model determination. *Bimetrika*, 82 : 711-732, 1995.
15) M. Hills and F. Alexander. Statistical methods used in assessing the risk of disease near a source of possible environmental pollution : a review. *Journal of the Royal Statistical A*, 152 : 353-363, 1989.
16) H. Inskip, V. Beral, P. Fraser, and P. Haskey. Methods for ageadjustment of rates. *Statistics in Medicine*, 2 : 483-493, 1983.
17) A.B. Lawson. On the analysis of mortality events around a prespecified fixed point. *Journal of the Royal Statistical Society A*, 156 : 363-377, 1993.
18) A.B. Lawson. On using spatial Gaussian priors to model heterogeneity in environmental epidemiology. *The Statistician*, 43 : 69-76, 1994. Proceedings of the Proctical Bayesian Statistics Conference.
19) A.B. Lawson. Markov chain Monte Carlo methods for putative pollution source problems in environmental epidemiology. *Statistics in Medicine*, 14 : 2473-2486, 1995.
20) A.B. Lawson. Markov chain Monte Carlo methods for spatial cluster processes. In *Computer Science and Statistics : Proceedings of the Interface*, volume 27, pages 314-319, 1996.
21) A.B. Lawson. Cluster modelling of disease incidence via mcmc methods. *Journal of Statistical Planning and Inference*, 1997. submitted.
22) A.B. Lawson. Some spatial statistical tools for pattern recognition. In A. Stein, F.W.T.P. de Vries, and J. Schut, editors, *Quantitative Approaches in Systems Analysis*, volume 7, pages 43-58. C. T. de Wit Graduate School for Production Ecology, 1997.
23) A.B. Lawson, A. Biggeri, and C. Lagazio. Modelling heterogeneity in discrete spatial data models via MAP and MCMC methods. In A. Forcina, G. Marchetti, R. Hatzinger, and G. Galmacci, editors, *Proceedings of the 11th International Workshop on Statistical Modelling*, pages 240-250. Graphos, Citta di Castello, 1996.
24) A.B. Lawson, D. Böhning, E. Lessafre, A. Biggeri, J.-F. Viel, and R. Bertollini, editors. *Disease Mapping and Risk Assessment for Public Health.* Wiley, 1999.
25) A.B. Lawson and A. Clark. Markov chain Monte Carlo methods for clustering in case event and count data in spatial epidemiology. In E. Halloran and J. Greenhouse, editors, *Statistics and Epidemiology : Environment and Clinical Trials.* Springer verlag, New York.
26) A.B. Lawson and A. Clark. Markov chain Monte Carlo methods for putative sources of hazard and general clustering. In A. B. Lawson, D. Böhning, E. Lesaffre, A. Biggeri, J.-F. Viel, and R. Bertollini, editors, *Disease Mapping and Risk Assessment for Public Health.* Wiley, 1999.
27) A.B. Lawson and M. Kulldorff. A review of cluster detection methods. In A. B. Lawson, D. B'ohning, E. Lesaffre, A. Biggeri, J.-F. Viel, and R. Bertolline, editors, *Disease Mapping and Risk Assessment for Public Health.* Wiley, 1999.
28) A.B. Lawson and F. Williams. Spatial competing risk modelling : multidiseaase outcomes from a single putative pollution source.
29) A.B. Lawson and F. Willams. Armadale : a case study in environmental epidemiology. *Journal of the Royal Statistical Society A*, 157 : 285-298, 1994.
30) R.F. Raubertas. Spatial and temporal analysis of disease occurrence for detection of clustering. *Biometrics*, 44 : 1121-1129, 1988.
31) K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Jour. of the Amer. Statist. Assoc.*, 85 : 617-624, 1990.
32) R. Stone. Investigations of excess environmental risks around putative sources : statistical problems and a proposed test. *Statististics in Medicine*, 7 : 649-660, 1988.
33) T. Tango. A class of tests for detecting 'general' and 'focussed' clustering of rare diseases. *Statistics in Medicine*, 14 : 2323-2334, 1995.
34) M. A. Tanner. *Tools for Statistical Inference.* Springer verlag, New York, 3rd edition, 1996.
35) M. van Lieshout and A. Baddeley. Markov chain Monte Carlo methods for clustering of image features. In *Proceedings 5th IEEE International Conference on Jmage Processing and its Applications*, number 410 in IEEE Conference Publication, pages 241-245. IEEE Press, 1995.
36) A. Whittemore, N. Friend, B. Brown, and E. Holly. A test to detect clusters of disease. *Biometrika*, 74 : 631-635, 1987.

本論文は，小地域健康マップにおける地域疾病集積性の解析における多様なアプローチの紹介である．その中には，仮説検定，Bayesian 流の集積性の形状，大きさのモデルも含まれている．また，市町村別の頻度データ（count data）だけでなく，被対象者の住所地が具体的に調査できたデータ（case data）の解析方法も考察し，更に，時間−空間集積性まで広げているので興味深い．