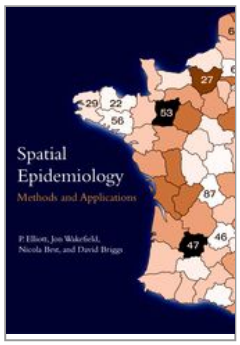


Spatial Epidemiology: Methods and Applications
Paul Elliott, Jon Wakefield, Nicola Best, and David Briggs

Print publication date: 2001
Print ISBN-13: 9780198515326
Published to Oxford Scholarship Online: September 2009
DOI: 10.1093/acprof:oso/9780198515326.001.0001

Bayesian approaches to disease mapping

J. C. Wakefield
N. G. Best
L. Waller

DOI:10.1093/acprof:oso/9780198515326.003.0007

Abstract and Keywords

This chapter examines the underlying assumptions of Bayesian methods for disease mapping and discusses mathematical details. The chapter proceeds as follows. Section 7.2 describes a three-state hierarchical model within which disease mapping data may be viewed. Section 7.3 considers implementation and simulation-based techniques. Section 7.4 provides two illustrative examples of the use of Bayesian disease mapping models. Section 7.5 considers some extensions and alternative approaches to the models presented. Section 7.6 provides a concluding discussion.

Keywords: Bayesian methods, disease mapping, epidemiological studies, spatial epidemiology, mathematical details, simulation-based techniques

7.1 Introduction

Disease mapping may be defined as the estimation and presentation of areal summary measures of health outcomes, and has a long history in epidemiology (see e.g. Chapter 12 and the references in Smans and Esteve 1992). The aims of disease mapping include simple description, hypothesis generation, allocation of health care resources, assessment of inequalities, and estimation of background variability in underlying risk in order to place epidemiological studies in context. Unfortunately, there are well-documented difficulties with the mapping of raw estimates since, for small areas and rare diseases in particular, these estimates will be dominated by sampling variability. The most common summary measure is the standardised morbidity/mortality ratio (SMR) which is defined, for area i,

by Y_i/E_i where Y_i and E_i denote the observed and expected number of counts in area i, respectively. The variance of this estimate is proportional to E_i^{-1} and so for areas with small populations there will be high sampling variability. To overcome this variability it is now commonplace to carry out 'smoothing' of the raw rates via hierarchical modelling. Although this approach may also be justified in a likelihood *random effects* context, or via a minimum mean squared error argument (e.g. Lindley and Smith 1972; Clayton and Kaldor 1987), in this chapter we describe the hierarchical smoothing model from a fully Bayesian perspective.

Good review papers on Bayesian methods for disease mapping have been published by Clayton and Bernardinelli (1992) and Mollié (1996). In this chapter, our aim is to examine more fully the underlying assumptions of the approach and provide more discussion of the mathematical details. The structure of the chapter is as follows. In Section 7.2 we describe a three-stage hierarchical model within which disease mapping data may be viewed. Historically the use of fully Bayesian methods has been hindered by computational consideration and, in Section 7.3, we consider implementation and, in particular, simulation-based techniques. In Section 7.4 we provide two illustrative examples of the use of Bayesian disease mapping models. Section 7.5 considers some extensions and alternative approaches to the models presented here, and Section 7.6 provides a concluding discussion. The two appendices contain more detailed developments of specific aspects.

(p.105) 7.2 Statistical formulation

In general, health outcomes may be available as area-level aggregated *count data*, or each case may have an associated exact location (e.g. from a case-control study), giving rise to *point data*. Count data are more typically used for disease mapping studies (often arising from routinely available sources for example, see Chapter 2); we therefore focus on methods for modelling count data in this chapter. Methods for the estimation and mapping of disease relative risk using point data are described in Chapter 6.

In this section we will describe a three-stage hierarchical model for disease counts. At the first stage we model the observed counts as a function of area-level summaries such as the risk or the relative risk. At the second stage, a joint distribution is specified for the collection of these risks or relative risks, possibly as a function of area-level explanatory variables. These first two stages constitute a generalised linear mixed model (Clayton 1996). The second stage distribution depends on unknown parameters and these are assigned a (hyper) prior distribution at the third stage of the model. The models of this section are closely related to those described in Chapter 11 for ecological correlation studies.

7.2.1 First-stage model

Let Y_{ij} and N_{ij} represent the number of cases and the number of individuals at risk in stratum j, $j = 1, \ldots, J$, and area i, $i = 1, \ldots, n$. As in all epidemiological studies, stratification by known risk factors (e.g. age and sex) is important since different areas will, in general, contain different proportions of individuals within each stratum and ignoring this information may lead to spurious conclusions. For example, Knorr-Held and Besag (1998) analysed lung cancer data in Ohio, USA for 1968–88 and showed that if the age structure of each county is ignored then it appears that the risk for white women is greater than for non-white women. However, when the age structure is accounted for, this conclusion is reversed. In this chapter, we assume that both the cases and the populations at risk are measured without error (though see Appendix I). The effects of relaxing these assumptions are considered more fully in Chapter 3, where the issue of inaccuracies in population counts is considered; and in Best and Wakefield (1999), where errors in both the numerators (cases) and denominators (populations) are considered.

Binomial model

With known populations and for non-infectious diseases, the starting point for analysis is the binomial model

$$Y_{ij} | p_{ij} \sim \text{Bin}(N_{ij}, p_{ij}), \tag{7.1}$$

where p_{ij} is the risk (probability) of disease in area i and stratum j. The maximum likelihood estimates (MLEs) for the stratified area-specific risks are given by $\hat{\rho}_{ij} = Y_{ij}/N_{ij}$ but, in general, the data will be too sparse to obtain robust estimates of each of these $n \times J$ quantities and so some simplifying assumptions are required. It is usual to make the proportionality assumption

$$\frac{p_{ij}}{1 - p_{ij}} = \theta_i \times \frac{p_j}{1 - p_j}, \tag{7.2}$$

(p.106) so that the effect of being in area i is to multiply each of the strata-specific *reference odds* $p_j/(1 - p_j)$ by the common *odds ratio*, θ_i, for that area. In this way we have reduced the number of quantities to estimate per area from J to 1. However, the proportionality assumption is clearly very strong and must be checked. For example, simple graphical plots of $\hat{\rho}_{ij}/(1 - \hat{\rho}_{ij})$ versus $\hat{\rho}_j/(1 - \hat{\rho}_j)$ may be constructed to assess proportionality or, more formally, logistic regression models containing area × stratum interactions may be fitted. If non-proportionality is found then separate analyses of collections of strata within which proportionality holds may be carried out.

The reference odds may be estimated simultaneously with the θ_i (e.g. Clayton 1996), or be fixed using either a set of odds from a reference area (external standardisation), or the overall odds for the study region

$$\frac{\hat{p}_j}{1-\hat{p}_j} = \frac{\sum_i Y_{ij}}{\sum_i (N_{ij} - Y_{ij})},$$

(internal standardisation). In the case where the $\hat{\rho}_j$ are treated as known, the MLEs of the odds ratios θ_i may be estimated via the logistic regression model

$$\text{logit } p_{ij} = \log \theta_i + \hat{\gamma}_j \quad (7.3)$$

where the $\hat{\gamma}_j = \log\{\hat{\rho}_j/(1 - \hat{\rho}_j)\}$ are known offsets. We note that this model does not acknowledge uncertainty in the γ_j, which may be a problem if these quantities are not estimated from extensive data.

In some situations (e.g. hypothesis generation studies, see Chapter 11), the relative risks θ_i will be regressed on a $k \times 1$ vector of area-specific explanatory variables, X_i, via the model

$$\log \theta_i = \alpha + \boldsymbol{X}_i^T \boldsymbol{\beta}, \quad (7.4)$$

where β is a $k \times 1$ vector of regression coefficients. The use of internal standardisation with known offsets, (i.e. Equation (7.3)), requires care since the a priori estimation of γ_j may remove some of the effect of the exposure $\boldsymbol{X}_i$. For example, older individuals may tend to live in areas with large values of X_i. These and other issues are discussed in Breslow and Day (1987, Chapter 4).

The model (7.1)–(7.4) does not acknowledge overdispersion in the data (i.e. $\text{var}(Y_{ij}) > N_{ij} p_{ij}\{1 - p_{ij}\}$) which may have both spatial and non-spatial components and arises from, for example, unmeasured risk factors and inaccuracies in the numerator and denominator. Appendix I develops models for overdispersion via the consideration of unknown risk factors and data anomalies. The models that we describe in the next section explicitly model this overdispersion.

With the binomial formulation it is not possible to aggregate the data Y_{ij} across stratum $j = 1,\ldots,J$, since the sum of binomial random variables $Y_i = \Sigma_j Y_{ij}$ is not of convenient form, and in particular is not binomial except in the uninteresting case where $p_{ij} = p_i$. In principle, this lack of data reduction is not a problem. However, it may create computational difficulties due to memory requirements if n and/or J are large and, in practice, numerical estimation problems may also arise if there are large numbers of (i, j) **(p.107)** cells containing zero cases since in this case the likelihood is likely to be very flat and hence contain little information.

Poisson model

For rare diseases we may approximate the binomial distribution (7.1) by the Poisson distribution:

$$Y_{ij} \sim \mathrm{Po}(N_{ij} \times p_{ij}).$$

The proportionality assumption corresponding to (7.2) is then expressed as

$$p_{ij} = \theta_i \times p_j, \tag{7.5}$$

where θ_i now corresponds to the *relative risk* of disease in area i with respect to the reference rate in each stratum. A great advantage of the Poisson approximation is that, when combined with the proportionality assumption (7.5), we may collapse over strata to obtain

$$Y_i \sim \mathrm{Po}(E_i \times \theta_i) \tag{7.6}$$

where $Y_i = \Sigma_j Y_{ij}$ and $E_i = \Sigma_j N_{ij} p_j$ denotes the *expected number* of cases in area i with strata-specific references rates p_j. The MLE $\theta_i = Y_i/E_i$ corresponds to the SMR. In general, the use of SMRs corresponds to *indirect standardisation*. Care must be taken with such an approach (e.g. see Breslow and Day 1987), essentially because, if the proportionality assumption (7.5) is not valid, then inappropriate summary relative risks θ_i will be obtained (see Chapter 9 for a fuller discussion).

For the Poisson model, we may specify a log-linear model for the relative risk as a function of area-specific risk factors $\boldsymbol{X}_i$

$$\log \theta_i = \alpha + \boldsymbol{X}_i^T \boldsymbol{\beta}. \tag{7.7}$$

The use of (7.6) and (7.7) will often suffer from the same problems of overdispersion ($\mathrm{var}(Y_i) < E_i \theta_i$) that were present for the binomial model, see Appendix I.

7.2.2 Second-stage model

In general, and for small areas in particular, the MLEs of odds ratios or relative risks, θ_i, will be highly unstable due to sparse data. To provide more robust estimation one may specify a joint model for $\theta = (\theta_1, \ldots, \theta_n)^T$ which allows the estimate of each θ_i to 'borrow strength' from the remaining estimates $\theta_{i'}$, $i' \neq i$. This is achieved by specifying a multivariate probability distribution for θ. There are many possibilities for this distribution with an important choice being on the form of the variability in the θ_i We may believe that the θ_i vary across the map without spatial pattern (so-called unstructured variability), display spatial dependence (structured variability), or exhibit a combination of the two. In particular, this choice will determine the level of global and local smoothing (corresponding to unstructured and structured variability, respectively) that is carried out. There are various decisions corresponding to each of these possibilities.

(p.108) ***Unstructured variability***

We first consider models that produce global smoothing across the study region. In the case of binomial data and no stratification (i.e. $Y_i|p_i \sim \mathrm{Bin}(N_i, p_i)$), a distribution for p_i that is analytically tractable is the beta distribution (see

Appendix I for details). This case is not of great interest, however, since we would almost always want to stratify the disease counts by age and sex. Such stratification may be carried out with reference odds evaluated via internal or external standardisation, and incorporated in (7.3). A natural approach then is to model the logarithm of the odds ratios, θ_i in (7.2) as

$$\log \theta_i = \alpha + \boldsymbol{X}_i^T \boldsymbol{\beta} + V_i, \tag{7.8}$$

where V_i is the *residual log odds ratio* in area i, relative to the reference region (after adjustment for known stratification risk factors and $\boldsymbol{X}_i$).
The vector $V = (V_1, \ldots, V_n)^T$ is often assumed to arise from the n-dimensional normal distribution

$$\boldsymbol{V} \sim N_n(\boldsymbol{0}_n, \sigma_v^2 \boldsymbol{I}_n),$$

where θ_n denotes the $n \times 1$ vector of zeros, I_n the $n \times n$ identity matrix and $\sigma_v^2 > 0$ controls the between-area variability of the V_i. Various models for overdispersion are discussed in Appendix I but the most natural interpretation of V_i in (7.8) is of an unmeasured risk factor that is common to all individuals in area i, and does not display a spatial pattern. Note that for small V_i, σ_v reflects, approximately, the standard deviation of the residual odds ratios.
The model (7.8) may be used when we have (7.2) with known reference probabilities p_j Knorr-Held and Besag (1998) consider the more general case in which both θ_i and p_j are simultaneously estimated via the model

$$\text{logit } p_{ij} = \alpha + \boldsymbol{X}_i^T \boldsymbol{\beta} + V_i + \gamma_j,$$

with $V_i \sim_{i.i.d.} N(0, \sigma_v^2)$, as before, and $\gamma_j = \text{logit } p_j$.
For the Poisson model (7.6), an analytically tractable second stage distribution for the unstructured variability is the gamma distribution. In the following, Ga(a, b) denotes the gamma distribution with mean a/b and variance a/b^2. This choice is natural since the gamma is conjugate to the Poisson and so the marginal distribution of Y_i can be calculated in closed form as negative binomial. In general, a and b are treated as unknown and the posterior distribution for these parameters is not of closed form (e.g. Clayton and Kaldor 1987). The marginal variance var($y_i|a, b$) can take a variety of forms depending on the gamma formulation that is chosen. Table 7.1 summarises three possibilities.

Case I results in a variance function that is proportional to the mean and is the closest to the conventional quasi-likelihood approach (McCullagh and Nelder 1989) that has been used in spatial epidemiological context by Diggle *et al*. (1997). Case II was used by Clayton and Kaldor (1987) in a disease mapping context and Case III allows the variance to be a quadratic function of the mean. Distinguishing between these cases, via residuals for example, is likely to be

difficult unless the range of the expected number of cases is large (see Wakefield and Morris 1999). When the expected number of cases is not large **(p.109)**

Table 7.1 Marginal variances for the data Y_i following from different gamma specifications in the model $Y_i|\theta_i \sim \text{Po}(E_i\theta_i)$. In each case $\text{E}[\theta_i] = a_i$ and $\text{E}[Y_i|a_i,b] = E_i a_i$ where $a_i = \exp(\alpha + X_i^T\beta)$

Case	Assumption	var(θ_i)	var($Y_i|a_i$,b)
I	$\theta_i \sim \text{Ga}(E_i a_i b, E_i b)$	$a_i/(E_i b)$	$\text{E}[Y_i,|a_i,b](1 + 1/b)$
II	$\theta_i \sim \text{Ga}(a_i b, b)$	a_i/b	$\text{E}[Y_i|a_i,b](1 + E_i/b)$
III	$\theta_i \sim \text{Ga}(b, b/a_i)$	a_i^2/b	$\text{E}[Y_i|a_i,b](1 + \text{E}[Y_i|a_i,b]/b)$

the forms of the variance will not be so different and hence it will be difficult to distinguish between them.

Although the gamma distribution is natural for incorporating unstructured over-dispersion, it does not extend to allowing structured variability with positive spatial correlations (but see Chapter 22 for a variation of this model which does allow for spatial dependence via a gamma *mixture* distribution). Instead, a normal distribution is often used for both the structured and the unstructured components. For the unstructured component we may again assume

$$\log\theta_i = \alpha + \boldsymbol{X}_i^T\boldsymbol{\beta} + V_i, \tag{7.9}$$

where α is an intercept term representing the overall log relative risk of disease in the study region compared to the reference rate, and V_i is the residual log relative risk in area i compared with the study region. As before, $V = (V_1, \ldots, V_n)^{\text{T}}$ is assumed to arise from the n-dimensional normal distribution

$$\boldsymbol{V} \sim N_n(\boldsymbol{0}_n, \sigma_v^2\boldsymbol{I}_n). \tag{7.10}$$

Note that the marginal distribution of Y_i is not available in closed form for the Poisson-log normal model; the mean and variance may be derived, however. For the model given by (7.6), (7.10), and (7.10) this is

$$\text{var}(Y_i|\mu_i,\sigma_v^2) = \text{E}[Y_i|\mu_i,\sigma_v^2]\{1 + \text{E}[Y_i|\mu_i,\sigma_v^2](\exp(\sigma_v^2) - 1)\},$$

where

$$\text{E}[Y_i|\mu_i,\sigma_v^2] = E_i\exp(\mu_i + \sigma_v^2/2),$$

and $\mu_i = \alpha + X_i^T\beta$. Hence the variance is a quadratic function of the mean (Case III in Table 7.1). For the normal distribution, the different assumptions analogous to those in Table 7.1 all lead to the same marginal variance function for the data. We note that the marginal median of Y_i is given by $E_i \exp(\mu i)$.

Note, that as pointed out by Wolpert and Ickstadt (1998), the Poisson-log normal model does not aggregate consistently. What this means is that if we specify a lognormal distribution for each of the relative risks and then combine two areas

(say) and specify a **(p.110)** log normal distribution for the relative risk of the combined area, then these distributions are inconsistent (because the sum of log normal distributions is not log normal). This issue is related to the problem of pure specification ecological bias (see Chapter 5) in which risk relationships do not remain constant across levels of aggregation. Provided the user remains aware of these issues, this does not seem such a great disadvantage, however, particularly since a normal second-stage distribution has been observed empirically to provide a good model for log relative risks over a range of aggregations, and is convenient in a number of other respects such as model flexibility and ease of computation.

A number of tests of heterogeneity have been proposed to assess departures from constant relative risk across the map (see Alexander and Cuzick 1992 and Chapter 8). However, we note that heterogeneity in area-level estimates of risk will almost always be present; the question of interest is whether this heterogeneity is of epidemiological significance. As pointed out above (and in Appendix I) the unstructured residual odds ratios or relative risks, $\exp(V_i)$, may be interpreted as corresponding to unknown or unmeasured risk factors that are *shared* by all individuals within area i. Hence between area heterogeneity in risk may indicate the absence of an important ecological (area-level) risk factor from the model (Appendix I). Theoretically, if these risk factors were observed then they could be included in the model and we would no longer need the V_i.

Spatial variability

We now consider the modelling of spatially structured variability in the log relative risks. The first stage model is identical to that with unstructured variability, but we model the log odds ratios (binomial data) or log relative risks (Poisson data) via

$$\log \theta_i = \alpha + \boldsymbol{X}_i^T \boldsymbol{\beta} + U_i \tag{7.11}$$

where U_i, $i = 1, \ldots, n$ denote spatially structured area-specific random effects, in contrast to the unstructured random effects V_i considered previously. The problem is to model the n-dimensional random variable $U = (U_1, \ldots, U_n)^T$, allowing for dependence between U_i and U_j, $i \neq j$. Due to the multitude of possibilities for this dependence, the modelling at this stage is fundamentally more difficult than in the unstructured case. Modelling may proceed either by specifying the *joint distribution* of U, or via the univariate *conditional distributions* $U_i | U_j = u_j, j \neq i, i = 1, \ldots, n$.
First, suppose

$$\boldsymbol{U} \sim N_n\left(\boldsymbol{0}_n, \sigma_u^2 \boldsymbol{\Sigma}\right), \tag{7.12}$$

where $N_n(\cdot,\cdot)$ denotes the n-dimensional normal distribution and Σ is an $n \times n$ positive definite correlation matrix. The parameter $\sigma_u^2 > 0$ controls the overall variance of the U_i. Let $\boldsymbol{Q} = \Sigma^{-1}$ and Q_{ij} denote element (i, j) of this matrix, $i, j = 1, \ldots, n$. As reviewed in Besag and Kooperberg (1995), standard properties of the multivariate normal

distribution (e.g. Johnson and Kotz 1972; Searle *et al*. 1991) produce the set of conditional distributions

$$U_i|U_j = u_j, j \neq i \sim N\left(\sum_{j=1}^{n} W_{ij}u_j, \sigma_u^2 D_{ii}\right), \quad (7.13)$$

(p.111) where $W_{ii} = 0$, $W_{ij} = -Q_{ij}/Q_{ii}$ and $D_{ii} = Q_{ii}^{-1}$. This derivation is described in detail in Appendix II. The specification (7.13) is sometimes referred to as an *autonormal* model (Besag 1974). From the symmetry of $\boldsymbol{Q}$ we have that

$$W_{ij}D_{jj} = W_{ji}D_{ii}. \quad (7.14)$$

From a modelling perspective, use of the joint formulation requires specification of the elements of the covariance matrix $\boldsymbol{\Sigma}$, while use of the conditional formulation reduces to specification of the matrix $\boldsymbol{W}$ of weights W_{ij} and D_{ii} in (7.13). The approaches are related through the relationship $\boldsymbol{Q} = \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{W})$, where $\boldsymbol{D}$ is an $n \times n$ diagonal matrix containing elements D_{ii}, $i = 1, \ldots, n$. As we describe in more detail below, however, convenient choices for $\boldsymbol{W}$ and $\boldsymbol{D}$ do not lead to a joint model that is a well-defined probability density since they lead to a $\boldsymbol{Q}$ that is singular (and consequently the mean of each U_i is undefined and the variances are infinite).

Conditional modelling

Besag (1974) argued that modelling the conditional distributions will often be more straightforward than the joint distribution in problems in which random variables are defined spatially. The majority of approaches incorporating a conditional model for spatial dependence proceed by first specifying a set of spatial weights for use in (7.13). These weights traditionally define a set of 'neighbours' that contribute positive weight to the conditional expectation of U_i with $W_{ij} = 0$ for the remaining regions, and $W_{ii} = 0$. This is in the same spirit as Markovian models in time series. The set of conditional distributions given by (7.13) defines a Markov random field (MRF) model.

In the Gaussian conditional autoregression (CAR) the specification (7.13) with a positive definite $\boldsymbol{Q}$ leads to

$$\boldsymbol{U} \sim N_n(\boldsymbol{0}_n, \sigma_u^2(\boldsymbol{I}_n - \boldsymbol{W})^{-1}\boldsymbol{D}).$$

Cressie and Chan (1989) proposed taking $D_{ii} = E_i^{-1}$ and $W_{ij} = \rho(E_j/E_i)^{1/2}$ for $j \in \partial i$, where ∂i denotes the set of labels of the 'neighbours' of area i, and $W_{ij} = 0$ otherwise. Letting $\boldsymbol{W} = \rho\boldsymbol{C}$, a positive definite $\boldsymbol{Q}$ requires ρ to lie in the interval $(\rho_{\min}, \rho_{\max})$ where $\rho_{\min}^{-1} < 0$ and $\rho_{\max}^{-1} > 0$ are, respectively, the smallest and largest eigenvalues of $\boldsymbol{D}^{-1/2}\boldsymbol{C}\boldsymbol{D}^{1/2}$. If we expect the spatial dependence to be positive, then we may take $\rho \in (0, \rho_{\max})$ The parameter ρ may be interpreted as measuring the strength of spatial dependence in the data since $\text{corr}(U_i, U_j|U_k, k \neq i, j) = \rho$. This interpretation is appealing but the conditional mean is given by

$$\mathrm{E}[U_i] = \frac{\rho}{E_i^{1/2}} \sum_{j \in \partial i} E_j^{1/2},$$

which does not seem a natural choice. Cressie and Chan (1989) also give alternatives in which the spatial dependence depends on the distance between area centroids.
A common MRF model is the *intrinsic Gaussian autoregression* prior considered by Besag *et al*. (1991) and given by

$$U_i | U_j = u_j, j \neq i \sim N\left(\bar{u}_i, \frac{\omega_u^2}{m_i}\right), \quad (7.15)$$

(p.112) where $\bar{u}_i = \frac{1}{m_i}\sum_{j \in \partial i} u_j$ and m_i is the neighbours. Comparison with (7.13) reveals that we have $D_{ii} = m_i^{-1}$ and $W_{ij} = m_i^{-1}$ for neighbouring areas $W_{ij} = 0$ otherwise. This specification seems natural, the conditional mean of U_i is the average of the neighbouring U_j's, but does not yield a positive definite precision matrix $\boldsymbol{Q}$. To see this, note that in the ith row of $\boldsymbol{I} - \boldsymbol{W}$ we have a single one and m_i entries with values $-m_i^{-1}$ and so the row sums are all zero indicating the matrix $\boldsymbol{Q}$ only has rank $n - 1$ and so is not invertible. The variance, ω_u^2, in (7.15), is no longer proportional to a marginal variance (since the latter no longer exists), to emphasise this we have changed our notation from σ_u^2 to ω_u^2, the latter is only interpretable conditionally.
The joint specification corresponding to (7.15) is given by

$$f(\boldsymbol{U}) \propto \exp\left\{-\frac{1}{2\omega_u^2}\sum_{i<j}(U_i - U_j)^2\right\}.$$

It is again clear that the joint distribution does not exist since we may have an arbitrary mean level for each U_i. Besag and Kooperberg (1995) note that, if we take $\mathrm{E}[U_i | U_j = u_j, j \neq i] = \lambda \bar{u}_i$ with $0 < \lambda < 1$, then the joint distribution is well defined. Unfortunately, to obtain reasonable levels of dependence, λ has to be very close to one and hence the *non-stationary* version (7.15) may be preferred (see Chapter 10 for a discussion of non-stationarity). A great advantage of a non-stationary model is that the form of the spatial dependence may vary across the study region.

When, as we have specified in (7.11), there is an intercept in the model we require an additional constraint on the prior specification (7.15) to allow identifiablity. Besag and Kooperberg (1995) suggest constraining the U_i to have zero mean and specifying a uniform prior on the whole of the real line for the intercept α. Equivalently, the unconstrained prior (7.15) may be used if we do not include an intercept term in (7.11); we use the latter parameterisation in the applications of Section 7.4.

One difficulty with the conditional approach is that it is often unclear how to choose the weights W_{ij} and the neighbourhood ∂i. To define *neighbours*, a number of authors (e.g. Clayton and Kaldor 1987; Besag *et al*. 1991; Richardson *et al*. 1995; Bernardinelli *et al*. 1997; Waller *et al*. 1997) have taken areas i and j to be neighbours if they share a common boundary. This is reasonable if all regions are of similar size and arranged in a regular pattern (as is the case for pixels in image analysis where these models originated), but is not particularly attractive otherwise. Various other neighbourhood/weighting schemes are possible (e.g. see Cliff and Ord 1981) though such formulations should be considered in the light of the symmetry constraint (7.14). Cressie and Chan (1989) take the neighbourhood structure to depend on the distance between area centroids and determine the extent of the spatial correlation (i.e. the distance within which regions are considered neighbours) via an exploratory analysis using the variogram (Cressie 1993; Chapter 10). Cressie and Chan (1989), Best *et al*. (1999), and Conlon (1999) consider distance-based weights with weights decreasing with increasing inter-centroid distances.

For the Gaussian CAR, if the neighbourhood criterion (i.e. ∂i, $i = 1, \ldots, n$) is specified along with var(U_i) and cov(U_i, U_j) for neighbouring i and j (with $Q_{ij} = 0$ for non-neighbouring i and j), then Besag and Kooperberg (1995) describe a procedure (based on the Dempster (1972) algorithm) by which $\boldsymbol{Q}$ may be determined. For the intrinsic Gaussian autoregression this approach may be modified so that, along with **(p.113)** the neighbourhood criterion, the quantities var ($U_i - U_j$) for neighbouring i and j, are specified. In this case W_{ij} and D_{ii} in (7.13) are then determined. Besag and Kooperberg (1995) report that this approach has been used in the examples of Besag *et al*. (1991), but apart from this the approach has not so far been used in a spatial epidemiological context.

In addition to the choice of spatial weights and neighbourhood structures required for the conditional approach, one could also consider the use of non-Gaussian forms for the conditional distributions. For example, Besag *et al*. (1991) and Best *et al*. (1999) consider the Laplacian distribution which leads to a second-stage spatial model based on the median rather than the mean of neighbouring rates. This may be more appropriate when discontinuities in disease rates are expected between areas.

We finally note that whether the conditional model defines a proper or an improper joint distribution the interpretation of the variances (σ_u^2 or ω_v^2) requires care since they depend on the neighbourhood structure. The specification of equal prior variances is most easily accommodated in the joint specification that we now describe.

Joint modelling

Our description of joint modelling is based on the multivariate normal distribution $N_n(\theta_0, \sigma_u^2\Sigma)$. The $n \times n$ positive definite correlation matrix Σ contains elements Σ_{ij}, $i, j = 1, \ldots, n$ with the off-diagonal terms Σ_{ij}, $i \neq j$ describing the correlation between U_i and U_j (i.e. the residual log odds ratios or residual log relative risks in areas i and j). Various structured forms may be assumed for Σ. A common choice is to assume that the dependence is a function of the distance, d_{ij}, between the population-averaged centroids (say) of areas i and j, that is, $\Sigma_{ij} = f(d_{ij},\phi)$ where ϕ represents a vector of parameters defining the particular structural form chosen. This assumption of *isotropy* is common but can be weakened to allow, for example, directional components. For ease of development, we will limit attention to formulations assuming isotropy. Such joint modelling is described by Raftery and Banfield (1991), and naturally assumes that distance is an appropriate metric for defining spatial associations (Besag *et al*. 1991: 52). An obvious choice for $f(d_{ij},\phi)$ is the family

$$f(d_{ij}, \boldsymbol{\phi}) = \exp\left\{-\left(\frac{d_{ij}}{\phi_1}\right)^{\phi_2}\right\}, \tag{7.16}$$

where $\phi_1 > 0$, $\phi_2 \in (0, 2]$ and $\phi = (\phi_1,\phi_2)$ Note that $\phi_2 = 2$ produces a covariance matrix that has both theoretical and practical drawbacks (see the discussion in Diggle *et al*. 1998). Devine *et al*. (1996) and Wakefield and Morris (1999) use model (7.16) with $\phi_2 = 1$ and investigate the extent of the spatial dependence using a variogram. A number of correlation functions are available as alternatives to (7.16). A two-parameter family that may be preferable (see the discussion in Diggle *et al*. 1998) is the Matérn class (Matérn 1986). In this case we have a scale parameter $\phi_1 > 0$ and a smoothness parameter $\phi_2 < 0$ and

$$f(d_{ij}, \boldsymbol{\phi}) = \frac{1}{2^{\phi_2 - 1}\Gamma(\phi_2)}\left(\frac{d_{ij}}{\phi_1'}\right)^{\phi_2} B\left(\frac{d_{ij}}{\phi_1'}\right),$$

(p.114) where $\phi_1' = \phi_1/(2\sqrt{\phi_2})$ and $B(\cdot)$ is the modified Bessel function of order ϕ_2 Handcock and Stein (1993) use this class in a kriging context.

Combining spatial and unstructured variability

Besag *et al*. (1991) propose to combine unstructured and structured variability via the model

$$\log \theta_i = \alpha + \boldsymbol{X}_i^T\boldsymbol{\beta} + U_i + V_i, \tag{7.17}$$

which they term a *convolution* prior. The U_i and V_i represent spatially structured and unstructured contributions respectively to the log odds ratio or log relative risk, and are assumed to be independent.

7.2.3 Third-stage model

We let ψ denote the parameters of the distributions that we have specified for the random effects $\boldsymbol{U}$ and/or $\boldsymbol{V}$. At the final stage of the model we specify

hyperpriors for these parameters, the intercept β, the regression coefficients β. For α and β, improper uniform or normal priors with large variance are often specified to represent vague beliefs.

Considerable care is required when specifying hyperpriors for ψ. Gamma distributions are typically chosen for the inverses of variances (i.e. σ_v^2, σ_u^2, and ω_{ij}^2), a common choice being Ga(ε,ε:) with ∈ very small (say 10^{-2} or 10^{-3}). However, Kelsall and Wakefield (1999) point out that even a diffuse prior such as this can be highly informative. In particular these priors are not consistent with very small levels of variability in the random effects. As an alternative they suggest using a Ga(0.5, 0.0005) prior for the inverse variance parameters since, in many contexts, this will give a plausible range of relative risks across the map. Mollié (Chapter 15) and Bernardinelli *et al*. (Chapter 16) suggest alternative strategies for choosing the hyperpriors of the conditional model. Experience of choosing appropriate hyperpriors for the parameters of the joint covariance model is limited and sensitivity of the resulting inference to different specifications should be carefully assessed.

7.3 Implementation

We define $\delta = (\delta_1,\ldots, \delta_n)^T$ where $\delta_i = \log \theta_i$, $i = 1, \ldots, n$. We are then interested in the posterior distribution

(7.18)
$$p(\boldsymbol{\delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\delta}) \times p(\boldsymbol{\delta}|\alpha, \boldsymbol{\beta}, \boldsymbol{\psi}) \times p(\alpha, \boldsymbol{\beta}, \boldsymbol{\psi}).$$

This distribution is analytically intractable; in this section we describe how implementation may be achieved.

Posterior estimation can proceed via either an empirical Bayes or fully Bayesian approach. The principle behind empirical Bayes methods is to replace the unknown third-stage hyperparameters α, β, and ψ by point estimates based (say) on the maximised likelihood of the hyperparameters given the data $\hat{\alpha}$, β, and $\hat{\psi}$ Maximisation may be achieved using the EM-algorithm (Dempster *et al*. 1977). The posterior distribution of **(p.115)** the vector of log relative risks, δ, given the data $\boldsymbol{Y}$ and point estimates $\hat{\psi}$, is then considered. Further details may be found in Clayton and Kaldor (1987), Devine *et al*. (1996), and Chapter 15. Devine *et al*. (1996) also describe the use of a constrained empirical Bayes technique that modifies the risk estimates to correct for the fact that the sample variance of the collection of empirical Bayes estimates underestimate the true variance. See Carlin and Louis (1996) for a general discussion of empirical Bayes methods.

Unfortunately, empirical Bayes methods suffer from a number of limitations. In particular, the estimates of disease risk/relative risk fail to reflect the uncertainty associated with the hyperparameter estimates $\hat{\alpha}$, β, and $\hat{\psi}$ and are thus overprecise. These problems are avoided by implementing a fully Bayesian

approach in which the *joint* posterior distribution (7.18) is investigated. Inference about the relative risks θ requires integration of the joint posterior with respect to (α, β, ψ) and, after this integration is carried out, the posterior uncertainty associated with these hyperparameters is acknowledged.

The intractability of the posterior distribution has led to the use of Markov chain Monte Carlo (MCMC) simulation methods to generate *samples* from the joint posterior distribution. As pointed out by Smith and Gelfand (1992) there is a duality between a probability density function and samples from that density. Given the former we may generate samples, and given the latter we may reconstruct the former. The principle is best understood by imagining a histogram constructed from a set of values sampled at random from a probability distribution: given a large enough sample, the histogram can provide virtually complete information about the distribution from which these samples were drawn. In particular, the mean, variance, percentiles, and other summaries of the distribution can be estimated by calculating the corresponding statistics from the sample. In a disease mapping context we may, for example, calculate the probability that the odds ratio or relative risk in a particular area exceeds a given threshold by counting the number of simulated values which are larger than the threshold.

MCMC algorithms proceed by simulating values of subsets of parameters *conditional* on the remaining parameters. The approach is particularly appealing for hierarchical models in which the conditional independencies indicated in (7.18) may be exploited. In particular, great simplifications of the algorithms may be achieved when we proceed with the conditional spatial specification. In contrast, for the joint model, for unknown ϕ we must, at each iteration evaluate Σ^{-1} and its determinant. Hence, this model requires far more computer time for implementation.

A strategy that avoids this computation within the algorithm is a discretisation of the prior distribution for ϕ (e.g. Kelsall and Wakefield, submitted). In this way, the matrix inversions corresponding to each prior choice may be carried out before beginning the MCMC iterations.

Details of the computational algorithms used for MCMC simulation are provided elsewhere (e.g. Gilks *et al*. 1996; Brooks 1998). These methods are implemented in the WinBUGS statistical software package (Spiegelhalter *et al*. 1996), which includes specific functions to fit conditional and joint models discussed in this chapter; it is this software we use for the examples of Section 7.4.

It is important to be aware of the potential practical problems of using MCMC methods for disease mapping analyses. All MCMC algorithms generate a sequence of *dependent* values which will *eventually* resemble a sample from the required posterior distribution: that is, the frequency with which values in an

interval appear in the sample **(p.116)** is equal to the probability content of the posterior probability content of the interval. However, early samples (called the 'burn-in') should be discarded since they are not representative of the posterior distribution. Various methods exist for determining how many samples to discard, although none are foolproof and most require the benefit of experience and judgement (see Mengersen *et al*. 1999, for a review). The number of samples generated after the discard phase will affect the accuracy of the posterior inference. Although MCMC methods allow estimation of the full posterior, the finite size of the simulation will introduce a degree of approximation error, known as the Monte Carlo standard error. Samples which are large and are nearly independent (i.e. have low autocorrelations between consecutively sampled values) will have a relatively low Monte Carlo standard error. Unfortunately, models used in a mapping context may exhibit high correlations between model parameters and include terms which are only weakly identified; this tends to result in highly autocorrelated samples and hence the MCMC simulation must be run for a large number of iterations in order to generate a sample of sufficient accuracy for posterior inference. Further discussion of these and other potential implementation problems relating to CAR models are discussed by Best *et al*. (1999).

7.4 Illustrative example

In this section we provide a short example to illustrate the practical application of Bayesian methods for disease mapping. Specifically, we compare the conditional and joint approaches to modelling spatial correlation between small-area disease rates. More detailed applications of these methods are provided in Chapters 15 and 16.

The study region of interest comprises the 144 electoral wards of the Mersey and West Lancashire districts of northwest England. We consider incidence of two cancers with contrasting aetiology: (i) lung cancer—the most common tumour, known to be related to smoking and socio-economic deprivation; and (ii) brain cancer—a less common tumour whose aetiology is largely unknown. Observed ward-level counts Y_i, $i = 1,\ldots, 144$, were obtained for each tumour site from the cancer registration data held by the UK Office for National Statistics. For brain cancer, the number of cases per ward ranged from 0 to 17 (median = 6) over the 11-year period 1981–91. For lung cancer, we used data for a single year (1991) to provide a more comparable number of cases per ward (range = 0–60; median = 20). Ward-level population counts by five-year age group and sex were obtained from the 1991 UK census. We note that these examples should be viewed as illustrative only. In particular, the examination of lung cancer rates when no information on smoking behaviour (or an approximate proxy for it) is available, is not an informative epidemiological enterprise.

Cancers are relatively rare diseases, and so it is convenient to assume a Poisson distribution for the observed counts Y_i, as in Equation (7.6). Expected counts E_i for each cancer were internally standardised for age and sex with the risks for each stratum being calculated a priori, as described in Section 7.2.1. Figure 7.1 shows maps of the maximum likelihood estimators $\theta_i = Y_i/E_i$ corresponding to the observed ward-specific SMRs for each cancer. Interpretation is difficult due to the sampling variability that is inherent in such estimates.

Bayesian smoothed estimates of the area-specific relative risks were then estimated using both the conditional and joint modelling approaches described in Section 7.2. **(p.117)**

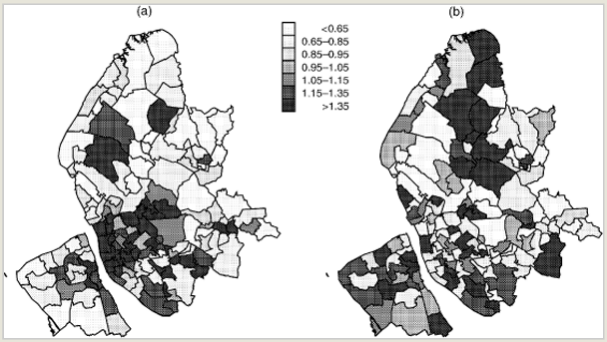


Fig. 7.1 Maximum likelihood estimates of the relative risks (SMRs) for (a) lung cancer; (b) brain cancer.

Under the conditional approach, we fitted the convolution model given by Equation (7.17). We used simple adjacency weights ($W_{ij} = m_i^{-1}$ for $j = \partial_i$ and $W_{ij} = 0$ otherwise) as in (7.15) to specify the conditional distributions of the spatial random effects U_i. Independent Ga(0.5, 0.0005) hyperprior distributions were assumed for the inverse variance parameters ω_u^{-2} and σ_v^{-2}. Under the joint approach, we fitted the second stage model $\log \theta_i = \alpha + U_i$, where the vector of spatial random effects $U = (U_1, \ldots, U_{144})^T$ was modelled using the multivariate normal prior given in (7.12) with $\Sigma_{ij} = \exp(-\lambda d_{ij})$ which corresponds to (7.16) with $\lambda = \phi_1^{-1}$ and $\phi_2 = 1$. We do not include separate unstructured random effects V_i in this model since the parameter λ, which is estimated along with the other model parameters, controls the degree (or lack of), spatial dependence between the random effects. We note, however, that it is feasible to include a set of non-spatial random effects in this model. A Ga(0.01, 0.01) hyperprior distribution was assumed for the inverse variance parameter σ_u^{-2} and a uniform distribution on the range (0.001, 10) was chosen for λ. The upper limit of the latter prior corresponds to a correlation matrix Σ which is approximately equal to the identity matrix since $\Sigma_{ii} = \exp(-10 \times d_{ii}) = 1$ when $d_{ii} = 0$ and $\Sigma_{ij} = \exp(-10 \times d_{ij}) \varepsilon 0$ for all d_{ij}, $i \neq j$ where d_{ij} ranges from 1.55km to 49.5km across the study region. The lower bound of the prior leads to off-diagonal elements of Σ_{ij} in the range $\exp(-0.001 \times 1.55) = 0.998$ to $\exp(-0.001 \times 49.5) = 0.952$, representing very strong spatial dependence.

Model fitting was carried out using MCMC simulation methods implemented in the WinBUGS software. Two separate chains starting from different initial values were run for each model. Convergence was checked by visual examination of

‘time series’ style plots of the samples for each chain, and by computing the Gelman and Rubin (1992) **(p.118)**

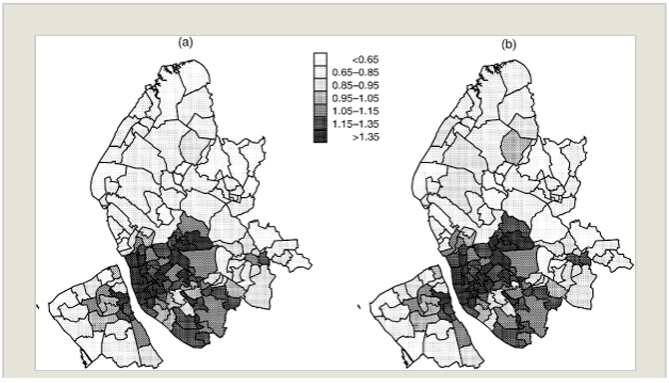
Table 7.2 Posterior means (95% credible intervals) of the variance components for each model and cancer site

	Lung cancer		Brain cancer	
	Conditional	Joint	Conditional	Joint
Unstructured	0.007	–	0.003	–
variance, σ_v^2	(0.0002, 0.037)		(0.001, 0.017)	
Spatial conditional	0.202	–	0.0008	–
variance, ω_u^2	(0.096, 0.323)		(0.0001, 0.003)	
Spatial marginal	–	0.148	–	0.015
variance, $\bar{\sigma}_u^2$		(0.072, 0.347)		(0.003, 0.042)
Distance	–	0.236	–	4.956
decay, λ		(0.078, 0.509)		(0.202, 9.739)

diagnostic based on the ratio of between to within chain variances for each model. On this basis, the first 3000 samples of each simulation were discarded as ‘burn-in’; each chain was run for a further 20 000 iterations, and posterior estimates were based on pooling the 2 × 20 000 samples for each model. This gave Monte Carlo standard errors < 1% of the posterior standard deviation for all parameters except the variance components σ_v^2 and ω_u^2 for which the Monte Carlo standard errors were about 8% of the standard deviation.

Table 7.2 shows the posterior mean and 95% credible intervals for the variance components for each dataset and model. Comparison of the equivalent parameter estimates for each cancer site shows that there is considerably more variability in the relative risks of lung cancer than of brain cancer. This excess variability is largely attributed to spatially structured effects: under the conditional model, the variance of the spatial components, ω_u^2, is over two orders of magnitude greater for lung cancer than for brain cancer, while the unstructured variance, σ_v^2, is negligible for both cancers. Under the joint model, the overall variance parameter σ_u^2 is one order of magnitude greater for lung cancer than for brain cancer, and λ is much smaller, indicating greater variability in the random effects among ward-level risk of lung cancer compared with brain cancer. In terms of interpretation we may calculate the distance at which the correlations drop to 0.5 which is given by log 2/λ. For lung and brain cancers, respectively, the distances are approximately 2.9 km and 0.14 km indicating far greater spatial dependence for the former.

Figures 7.2(a) and (b) show the smoothed estimates of the relative risk of lung cancer, θ_i, for the conditional convolution model and the joint model, respectively. Corresponding maps for brain cancer are shown in Figs 7.3(a) and (b). Although the estimates for lung cancer look similar under both models, the estimates for brain cancer show a different spatial pattern under the conditional model. This pattern may be explained by noting the following: (i) the conditional model is non-stationary; (ii) the posterior distribution for the conditional spatial variance parameter ω_u^2 is close to zero for brain cancer, suggesting that the risk in any given ward is very similar to that in neighbouring wards; (iii) the observed SMRs for brain cancer tend to be higher in the southwest (median = 1.1, interquartile range = 0.90 – 1.30) versus the rest of the study region **(p.119)**



(median = 0.85, interquartile range = 0.68–1.21); (iv) the River Mersey separates wards in the southwest from the rest of the region and thus acts as a boundary under the adjacency-based conditional weighting scheme (i.e. the 21 wards southwest of the Mersey are not considered to be neighbours of any wards across the river). This combination of **(p.120)** factors effectively leads to the risk estimates in all wards southwest of the river being smoothed towards one 'local' mean, while the risk estimates of all wards northeast of the river are smoothed towards a different (and lower) 'local' mean. We found that nearly identical estimates of relative risk were obtained by fitting a model with unstructured random effects only, but allowing a separate intercept term for wards to the southwest and for wards to the northeast of the river.

Fig. 7.2 Posterior mean relative risk of lung cancer, θ_i, estimated using (a) the conditional formulation; (b) the joint formulation.

Fig. 7.3 Posterior mean relative risk of brain cancer, θ_i, estimated using (a) the conditional formulation; (b) the joint formulation.

The southwest to northeast trend in relative risk is less pronounced for lung cancer under the conditional model because the risks in the two regions are more comparable. The phenomenon does not occur for either cancer under the joint formulation since the distance-based correlation structure allows cross-river dependence of the U_i.

This example clearly demonstrates that considerable care is needed when specifying, estimating, and interpreting Bayesian disease mapping models! Further analyses of these data are presented in Chapter 8.

7.5 Extensions and alternative approaches

7.5.1 Spatio-temporal models

Let $\theta_i(t)$ denote the log relative risk/log odds ratio in area i at time t. Then Bernardinelli *et al*. (1995) propose the following model:

$$\log \theta_i(t) = \alpha + \delta_{i1} + \beta t + \delta_{i2} t.$$

They allow δ_{i1} and δ_{i2} to be both either structured or unstructured random effects. This model therefore allows for a temporal trend that is allowed to vary across areas. In order to investigate whether regional spatial patterns change over time, Waller *et al*. (1997) consider an extension, nesting spatial effects within time (i.e. the model includes a set of n unstructured random effects V_{it}

and n spatially structured random effects U_{it} for each time period t). The prior variances may also vary with time period. Using such a formulation, Waller *et al*. (1997) found increasing residual clustering in annual lung cancer mortality for counties in Ohio, USA, across the years 1968–88. Knorr-Held and Besag (1998) analyse the same data, adding age-adjustment and using a binomial model in their analysis, and summarise spatial pattern across time.

Knorr-Held and Besag (1998) note that models with independent time and space effects offer ease in interpretation (U_i and V_i offer adjustments to relative risk, while U_{it} and V_{it} offer adjustments to relative risk within time period), but Bernardinelli *et al*. (1995) stress that space-time interactions may be expected in disease incidence data. Knorr-Held (1999) has recently implemented a model incorporating spatio-temporal interactions with a temporally evolving spatial structure.

7.5.2 Non-parametric mixture models

Various authors have proposed a mixture model approach to disease mapping (e.g. Clayton and Kaldor 1987; Schlattmann *et al*. 1996). The basic idea is that the population under investigation consists of an unknown number of homogeneous sub-regions with different levels of risk. Within each sub-region, the disease counts are assumed to follow a Poisson distribution, whilst the distribution of risk parameters across sub-regions is **(p.121)** specified as a non-parametric mixture with an unknown number of components. Each component has an uncertain mixing weight and associated relative risk parameter. Knorr-Held and Raßer (1999) use a fully Bayesian approach based on a reversible jump MCMC algorithm (Green 1995). Unlike the models described in Section 7.2, such non-parametric models make no assumptions concerning the form of spatial or unstructured variation in disease risk, and are qualitatively useful in that they provide a method of classifying areas into one of a small number of risk categories which can be advantageous for display/exploratory purposes. However, it is not possible to incorporate the uncertainty in the mixing distribution in inference via maximum likelihood approaches and Bayesian non-parametric models are subject to difficulties with both implementation and prior specification.

7.5.3 Methods based on Poisson processes

Another approach to modelling aggregate counts of events (cases and non-cases) in geographical areas is to view the events as realisations of a heterogeneous Poisson process integrated over small areas (e.g. Wakefield and Elliott 1999; Section 6.3.2 in Chapter 6). In this case the expected number of cases is given by

$$\int_{A_i} \lambda_1(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

where A_i denotes the ith area, x is the spatial location in A_i and $\lambda_1(x)$ is the intensity function of the process that generates the cases. The latter is usually modelled as

$$\lambda_1(\boldsymbol{x}) = \rho(\boldsymbol{x})\lambda_0(\boldsymbol{x})$$

where $\lambda_0(x)$ is an intensity function representing the population density and $\rho(x)$ can be specified as a function of spatially referenced covariate effects. This approach is very natural since one is modelling the underlying risk surface, rather than the discrete set of risks corresponding to the areas that are arbitrarily imposed by the data collection procedure.

Best *et al*. (Chapter 22) describe a related model in which the intensity function is modelled explicitly as a function of spatial location and environmental covariates. Kelsall and Wakefield (submitted) assume that the underlying logarithm of the relative risk is modelled by a Gaussian process in continuous space and then derive the correlation between areas i and j, $i \neq j$, calculating the average correlation between locations within each of the areas. They choose a correlation function that is a cubic function of distance but many others are possible (Wackernagel 1998).

7.6 Concluding remarks

There are a number of unresolved issues in the hierarchical modelling of disease risk across geographical areas, and in the display and interpretation of disease maps. Modelling the spatial dependence in particular is a difficult problem since there are few areas and the form may change across the study area. The form of the spatial dependence should, if **(p.122)** possible, be related to potential exposures. For example, distance-based models may be more likely to be more realistic for air pollution that varies smoothly.

The strategy for modelling spatial dependence depends on the sensitivity of inference to the choice of form assumed and, to an extent, on the aim of the analysis. In some studies the form of the spatial dependence may be of interest in itself since it may suggest the type of exposure that is (at least partially) responsible for the variability in relative risk. In ecological studies the estimate of the regression coefficient is of principle importance and so the residual spatial dependence is a nuisance parameter (though a sensitivity analysis should be carried out). There may also be confounding between the exposure and unmeasured risk factors that are being picked up by the spatial random effects.

Summarising the results of a mapping study will, in general, not be straightforward and a number of summaries may be presented. For example, maps showing point estimates of the relative risks, and the precision of such estimates, are informative. The posterior probability of the exceedance of a threshold of interest will also often be useful. As can be seen from the examples presented here, it will often be necessary to run a variety of models with differing assumptions to explore the robustness of any particular inference. Where inconsistencies are found, these need to be explained. Too often a single map, or a small number of maps are presented (with potentially great visual

impact) among the potentially large number that could be selected. Choice of cut-points, colours, and shading schemes are also important (Smans and Esteve 1992; Chapter 14). These aspects of presentation must be considered and addressed with great care.

As always in spatial epidemiological studies the necessity for high quality data should not be forgotten, and proposed modelling approaches should acknowledge the possible existence of data anomalies.

Appendix I: models for overdispersion

In Section 7.2.2, random effects were introduced in order to acknowledge that there may be both errors in the numerator and denominator data, and risk factors that are unmeasured. These models lead to excess variation in the observed counts. In this section we discuss this overdispersion in more detail. Bernardinelli *et al*. (1995: 2436) motivate the use of random effects by stating that, 'A cluster size bigger than the area size leads to a [spatially structured] *clustering* model, while a cluster size smaller than the area size leads to a *heterogeneity* model'. In this section we expand on this statement, and attempt to make it more precise, by considering various modelling scenarios.

Binomial case

We begin with the binomial model and let Y_i and N_i denote the number of cases and the population at risk in area i, $i = 1, \ldots, n$. If the risk is constant within each area, and independent between areas, then we have $Y_i|p_i \sim_{i.i.d.} \text{Bin}(N_i, p_i)$ This model follows immediately when there is no variability in risk within the area but an alternative derivation has been described by Knorr-Held and Besag (1998). suppose that within each area, there are K strata corresponding to different risk factors, and that, for area i, the disease probability in stratum k is given by p_{ik}, with the probability of an individual falling in stratum k being ν_{ik}, $k = 1, \ldots, K$. If it is assumed that the collection of stratum **(p.123)** counts within each area $(N_{i1}, \ldots, N_{iK})$follow a multinomial distribution $\text{Mult}_K(N_i, \nu_i)$ where $\nu_i = (\nu_{i1}, \ldots, \nu_{iK})^T$, then (without knowledge of which stratum individuals are contained within) the response of each of the N_i individuals in area i is Bernoulli with constant probability of disease $p_i = \Sigma_K \nu_{ik} p_{ik}$ Each of the outcomes is independent (due to the multinomial sampling) and so Y_i is $\text{Bin}(N_i, p_i)$. This multinomial formulation is appropriate provided individuals in the K risk groups are randomly distributed within the area. When we have clustering of risk factors within each area (as would be the case if the stratum referred to genetic risk factors, say), then dependence is induced which will increase the variance and hence invalidate the binomial assumption.

We now consider the use of random effects in the binomial model. Specifically suppose that $p_i \sim_{i.i.d.} f(q, \tau^2)$ where $f(\cdot,\cdot)$ represents a probability density function with $\text{E}[P_i] = q$ and $\text{var}(P_i) = \tau^2 q(1 - q)$. This leads to a marginal distribution for Y_i that is no longer binomial but for which

$$\mathrm{E}[Y_i] = N_i q \quad \text{and} \quad \mathrm{var}(Y_i) = N_i q(1-q)\sigma_i^2,$$

where $\sigma_i^2 = 1 + \tau^2(N_i - 1)$ (which is greater than one if $N_i > 1$). Note also that marginally we have cov (Y_i, $Y_{i'}$) = 0 for $i \neq i'$ since p_i and $p_{i'}$ are independently drawn from f. With regard to the above quote of Bernardinelli *et al*, this model, literally interpreted, is based on *each area* corresponding to a *single cluster*. We may therefore interpret such a model as accounting for unmeasured area-level covariates that do not display spatial structure.

As an aside we note that McCullagh and Nelder (1989: 125) develop a model for over-dispersion that is closely related to that just described. Letting $Y = \Sigma_i$, Y_i and $N = \Sigma_i N_i$ denote the total number of cases and the total population of the study region, their model gives E[Y] = Nq and var(Y) = $Nq(1 - q)\sigma^2$ where $\sigma^2 = 1 + \tau^2 (c - 1)$ with $c = \Sigma_i N_i^2/N \geq 1$. To obtain the marginal distribution of Y_i a specific form is required for f. The conjugate choice is the beta distribution Be(α,β) with $q = \alpha/(\alpha + \beta)$ and $\tau^2 = (\alpha + \beta + 1)^{-1}$ This choice allows the probabilities Pr($Y_i = y$) = E_p[Pr($Y_i = y|p$)] to be evaluated analytically and leads to a beta-binomial marginal distribution. Again, such a model may be interpreted as accounting for unmeasured, spatially unstructured area-level covariates. If these unknown covariates are spatially correlated, or are associated with spatial regions larger than the areas $i = 1, \ldots, n$ used in the analysis, an alternative form is required for f which allows for spatial dependence (for example joint or conditional models, specified on the logistic scale, see Section 7.2.2).

Note that none of the above models explicitly arise from a scenario involving a cluster size *smaller* than the area size. Since we would not expect clusters to follow areas boundaries that are defined (usually) for administrative reasons, random effects models should be viewed as rough approximations to the 'true' data-generating mechanism.

Poisson case

We proceed as with the binomial case. Consider $Y_i|\theta_i \sim_{i.i.d.}$ Po ($E_i \theta_i$) and suppose that known risk factors have been accounted for within the expected numbers. We then assume that the relative risks θ_i are drawn from a distribution $f(\phi, \tau'^2)$ with E[θ_i] = ϕ and var(θ_i) = $\phi\tau'^2$. This leads to E[Y_i] = $E_i\phi$ and $\mathrm{var}(Y_i) = E_i\phi\sigma_i^2$ where $\sigma_i^2 = 1 + \tau'^2 E_i$. As described in Section 7.2.2, the conjugate choice for f is the gamma distribution. The above description is consistent with Case II of Table 7.1 with $\phi = \exp(\alpha)$ (i.e. no area-level covariates) and $\tau'^2 = b^{-1}$. Similar developments are possible for the other two cases **(p.124)** in Table 7.1. Again the interpretation is that individuals within each area consist of a single cluster with a cluster-specific relative risk. As before, cluster sizes larger than the area size may be approximated by choosing a form for f incorporating spatial dependence.

We now give another development of random effects models in which we directly show how such models may be interpreted in terms of an unobserved covariate. suppose that the 'true' model is given by

$$Y_i|\theta_i \sim \text{Po}(E_i\theta_i) \tag{7.19}$$

with

$$\log\theta_i = \alpha' + Z_i\beta \tag{7.20}$$

and where Z_i is a risk factor whose levels are independently distributed across the areas of the study region via some distribution f. If we do not observe Z_i then we may view the random effects that we introduce as acting as surrogates for the unmeasured covariate. For example, suppose that f corresponds to a normal distribution with μ_z and variance σ_z^2. Then suppose we assume the model

$$\log\theta_i = \alpha + V_i \tag{7.21}$$

with $V_i \sim_{i.i.d.} N(0,\sigma_v^2)$. Then this model corresponds exactly to (7.19) and (7.20) with $\alpha = \alpha' + \beta\mu_z$, $\sigma_v^2 = \beta^2\sigma_z^2$ and $V_i = \beta(Z_i - \mu_z)$, directly relating the random effect to an unmeasured risk factor. Obviously spatially dependent random effects may be justified in exactly the same way by assuming that the exposures across areas have spatial pattern.

Finally, we show how the random effects may be interpreted as accounting for data anomalies. We begin with denominator errors. Let E_i' denote the 'true' expected number and E_i the observed expected number in area i, and assume that the inaccuracies in area i may be modelled using the simple Berkson errors-in-variables model (see Chapter 3):

$$\log E_i' = \log E_i + V_i$$

with $V_i \sim_{i.i.d.} N(0,\sigma_v^2)$. Finally, we assume that there is no between-area variability in the relative risks so that $\log\theta_i = \alpha$ Then we again obtain a model equivalent to (7.19) and (7.21) that is Y_i follows the distribution Po($E_i e^{\alpha} e^{V_i}$) with $\text{Po}(E_i e^{\alpha} e^{V_i})$ In a similar vein, exp(V_i) may be viewed as accounting for numerator errors in area i (e.g. case under-or overascertainment). For both types of error the assumption of independent errors across areas may be reasonable. In practice the random effects are accounting for the combined effect of both unmeasured risk factors and for numerator and denominator data anomalies.

Appendix II: properties of the normal distribution

suppose we have $U = (U_1,\ldots, U_n)^T$ and assume

$$\boldsymbol{U} \sim N_n\left(\boldsymbol{\theta}_n, \sigma_u^2\boldsymbol{\Sigma}\right),$$

(p.125) where $N_n(\cdot,\cdot)$ denotes the n-dimensional normal distribution, $\boldsymbol{e}_n$ an $n \times 1$ vector of ones, $\sigma_u^2 > 0$ and Σ is an $n \times n$ positive definite matrix. Let $\boldsymbol{Q} = \Sigma^{-1}$ and suppose now that $\boldsymbol{U}^T = (\boldsymbol{U}_1, \boldsymbol{U}_2)$ where $\boldsymbol{U}_1$ is $m \times 1$ ($1 \le m < n$) and $\boldsymbol{U}_2$ is $(n - m) \times 1$. Then (e.g. Searle *et al*. 1991) we have

(7.22)
$$\boldsymbol{U}_2 | \boldsymbol{U}_1 = \boldsymbol{u}_1 \sim N_{n-m}(-\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\boldsymbol{U}_1, \sigma_u^2 \boldsymbol{Q}_{22}^{-1}),$$
where
$$\boldsymbol{Q} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}.$$
We now explicitly consider a disease mapping context and use (7.22) to derive the CAR model given by (7.13). We first let $\boldsymbol{W} = \{W_{ij}, i, j = 1, \ldots, n\}$ denote the matrix of weights and $\boldsymbol{D}$ an $n \times n$ diagonal matrix containing elements D_{ii}, $i = 1, \ldots, n$, such that $\boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{W})$ is symmetric and positive definite. Writing $\boldsymbol{Q} = \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{W})$ we then have

(7.23)
$$\boldsymbol{U} \sim N_n(\boldsymbol{0}_n, \sigma_u^2 \boldsymbol{Q}^{-1}).$$
Taking $\boldsymbol{U}_2$ to be U_i and $\boldsymbol{U}_1$ to be the $(n - 1) \times (n - 1)$ matrix obtained from $\boldsymbol{U}$ by deleting the ith row and the ith column, we may use (7.22) to yield (7.13), as required. From (7.22) it also follows that the partial correlation between U_i and U_j is given by $\text{corr}(U_i, U_j | U_k = u_k, k \neq i, j = \text{sgn}(W_{ij})(W_{ij} W_{ji})^{1/2}$.

Acknowledgement

This work was supported, in part, by an equipment grant from The Wellcome Trust (0455051/Z/95/Z).

References

Bibliography references:

Alexander, F. and Cuzick., J. (1992). Methods for the assessment of disease clusters. In *Geographical and environmental epidemiology: methods for small-area studies* (P. Elliott, J. Cuzick, D. English, and R. Stern, ed.), 238–50. Oxford University Press.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, **14**, 2433–43.

Bernardinelli, L., Pascutto, C., Best, N. G., and Gilks, W. R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine*, **16**, 741–52.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, **82**, 733–46.

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

Best, N. G. and Wakefield, J. C. (1999). Accounting for inaccuracies in population counts and case registration in cancer mapping studies. *Journal of the Royal Statistical Society A*, **162**, 363–82.

Best, N. C., Arnold, R. A., Thomas, A., Waller, L. A., and Conlon, E. M. (1999). Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith ed.), 131–56, Oxford University Press.

(p.126) Breslow, N. E. and Day, N. E. (1987). *Statistical methods in cancer research*. Vol. II: *The analysis of cohort studies*, IARC Scientific Publications 82. International Agency for Research on Cancer, Lyon.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69–100.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*, Chapman and Hall, London.

Clayton, D. G. (1996). Generalised linear mixed models. In *Markov chain Monte Carlo in practice* (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter ed.), 279–301. Chapman and Hall, London.

Clayton D. G. and Bernardinelli L. (1992). Bayesian methods for mapping disease risk. In *Geographical and environmental epidemiology: methods for small-area studies* (P. Elliott, J. Cuzick, D. English, and R. Stern ed.), 205–20. Oxford University Press.

Clayton, D. G. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, **43**, 671–82.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models and applications*. Pion, London.

Conlon, E. M. (1999). *Estimation and flexible correlation structures in spatial hierarchical models of disease mapping*. Unpublished PhD thesis, Division of Bio statistics, School of Public Health, University of Minnesota.

Cressie, N. A. C. (1993). *Statistics for spatial data* (rev. edn). Wiley, New York.

Cressie, N. and Chan, N. H. (1989). Spatial modelling of regional variables. *Journal of the American Statistical Association*, **84**, 393–401.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, **28**, 157–75.

Dempster, A. P., Laird. N. M., and Rubin. D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Devine, O. J., Louis, T. A., and Halloran, M. E. (1996). Identifying areas with elevated disease incidence rates using empirical Bayes estimators. *Geographical Analysis*, **28**, 187–99.

Diggle, P. J., Morris, S. E., Elliott, P., and Shaddick, G. (1997). Regression modelling of disease risk in relation to point sources, *Journal of the Royal Statistical Society, Series A*, **160**, 491–505.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Applied Statistics*, **47**, 299–350.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Chapman and Hall, New York.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–32.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, **35**, 403–10.

Johnson, N. L. and Kotz, S. (1972). *Distributions in statistics: continuous multivariate*. Wiley, New York.

Kelsall, J. E. and Wakefield, J. C. (1999). Discussion of 'Bayesian models for spatially correlated disease and exposure data', by Best *et al*. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith ed.), **151**, Oxford University Press.

Kelsall, J. E. and Wakefield, J. C. (submitted). Modelling spatial variability in disease risk. Submitted to *Journal of the American Statistical Association*.

Knorr-Held, L. (1999). Bayesian modelling of inseparable space-time variation in disease risk. University of Munich Institute of Statistics Technical Report.

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, **17**, 2045–60.

Knorr-Held, L. and Raßer, G. (1999). Bayesian detection of clusters and discontinuities in disease maps. University of Munich Institute of Statistics Technical Report.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

Matérn, B. (1986). *Spatial variation* (2nd edn). Springer, Berlin.

McCullagh, P. and Nelder, J. A. (1989). *Generalised linear models* (2nd edn). Chapman and Hall, London.

Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostic: A 'review'. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, ed), 415–40. Oxford University Press. (See also http://www.maths.qut.edu.au/mengersen/McDiag)

(p.127) Mollié, A. (1996). Bayesian mapping of disease. In *Markov chain Monte Carlo in practice* (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter ed.), 359–79. Chapman and Hall, London.

Raftery, A. E. and Banfield, J. D. (1991). Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics (discussion of Besag, York and Mollie). *Annals of the Institute of Statistical Mathematics*, **43**, 32–43.

Richardson, S., Montfort, C., Green, M., Draper, G., and Muirhead, C. (1995). Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Statistics in Medicine*, **14**, 2487–501.

Schalttmann, P., Dietz, E., and Bohning, D. (1996). Covariate adjusted mixture models and disease mapping with the program Dismap Win. *Statistics in Medicine*, **15**, 919–29.

Searle, S. R., Casella, G., and McCulloch, C. E. (1991). *Variance components*. Wiley, London.

Smans, M. and Esteve, J. (1992). Practical approaches to disease mapping. In *Small area studies in geographical and environmental epidemiology* (J. Cuzick, P. Elliott, D. English, and R. Stern ed.), 141–50. Oxford University Press.

Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *The American Statistician*, **46**, 84–8.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996). *BUGS: Bayesian inference using Gibbs sampling, Version 5.0*. Medical Research Council Biostatistics Unit, Cambridge.

Stern, H. S. and Cressie, N. (1999). Small-area and point-level Bayesian models for inference on extremes in disease maps. In *Disease mapping and risk assessment for public health*, (A. B. Lawson, D. Boehning, E. Lesaffre, A. Biggeri, J. F. Viel, and R. Bertollini ed.), 63–84. Wiley, Chichester.

Wackernagel, H. (1998). *Multivariate geostatistics* (2nd edn). Springer, New York.

Wakefield, J. C. and Elliott, P. (1999). Issues in the statistical analysis of small-area health data. *Statistics in Medicine*, **18**, 2377–99.

Wakefield, J. C. and Morris, S. E. (1999). Spatial dependence and errors-in-variables in environmental epidemiology. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, ed.), 657–84, Oxford University Press.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, **92**, 607–17.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–67.