17

Bayesian Spatial Analysis

Andrew B. Lawson and Sudipto Banerjee

17.1. INTRODUCTION

Spatially referenced data occur in diverse scientific disciplines including geological and environmental sciences (Webster and Oliver, 2001), ecological systems (Scheiner and Gurevitch, 2001), disease mapping (Lawson, 2006) and in broader public health contexts (Waller and Gotway, 2004). Very often, such data will be referenced over a fixed set of locations in a region of study. These locations can be with regions or areas with well-defined neighbors (such as pixels in a lattice, counties in a map, etc.), whence they are called *areally referenced* or *lattice* data. Alternatively, they may be simply points with coordinates (latitude–longitude, Easting–Northing etc.), in which case they are called *point referenced* or *geostatistical*. Statistical theory and methods to model and analyze such data depend upon these configurations and has enjoyed significant developments over the last decade; see, for example, the books by Cressie (1993), Chilés and Delfiner (1999), Móller and Waagpetersen (2004), Schabenberger and Gotway (2004), and Banerjee *et al.* (2004) for a variety of methods and applications.

With recent advances in computational methods (particularly in the area of Monte Carlo algorithms), it is now commonplace to be able to incorporate spatial correlation as an important modeling ingredient. It is now feasible to fit routinely linear models with a variety of features within a modeling hierarchy. With the implementation of fast algorithms such as Markov Chain Monte Carlo (MCMC), sophisticated models that were previously inaccessible are now within reach allowing us to move beyond the simpler, and often inadequate, descriptive measures for analyzing spatial structure.

Spatial analysis can be viewed in a number of ways. For the statistician, there are two basic approaches to statistical modeling and inference: frequentist or likelihood based inference, and Bayesian inference. Here we focus on the latter approach. Bayesian inference and modeling can be seen as an extension of likelihood methods, but it also has a fundamentally different view of the inferential process.

17.2. NOTATION

The following notation will be used throughout this chapter. A random variate is denoted y_i, for an item in a vector. The vector of these items is $\mathbf{y}$. Often $\mathbf{y}$ will be related to independent variables (such as in a linear model). In that case the matrix of such variables can be defined as X. A linear model can be defined, for a single independent variable $\mathbf{x}_1$ as:

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i.$$

In general, the matrix formulation of the model, where $i = 1, \ldots, n$ will be:

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e} \tag{17.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of the dependent variable, X is an $n \times p$ matrix of p independent predictors (or covariates), $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector of the corresponding slopes and $\mathbf{e}$ is an $n \times 1$ vector of the errors. Often we make distributional assumptions, such as $\mathbf{e} \sim N(\mathbf{0}, \Sigma)$ These expressions imply that the errors are normally distributed with a zero-vector, $\mathbf{0}$, as the mean and a covariance matrix $\boldsymbol{\Sigma}$.

17.2.1. Point-referenced spatial data notation

As we will be dealing with spatial data, we will require some notation specific to such settings. When the referencing is done using coordinates (latitude–longitude, Easting–Northing, etc.) over a domain $\mathcal{D}$, we denote it as $s \in \mathcal{D}$; for instance in two-dimensional domains we have $s \equiv (s_x, s_y)$. The most frequently encountered scenario observes a spatial field measured at a finite set of locations, say $\mathcal{S} = \{s_1, \ldots, s_n\}$. We usually name this a random field, which we denote as $\{w(s) : s \in \mathcal{D}\}$ or simply as $w(s)$ in short. A realization of this random field will be a vector $\mathbf{w} = (w(s_1), \ldots, w(s_n))$.

17.2.2. Health data notation

For health data discussed in this chapter we will confine ourselves (mostly) to examining count data arising within small arbitrary administrative areas (such as census tracts, zip codes, postcodes, counties). Define y_i as the count of disease within the ith small area. Assume that $i = 1, \ldots, m$. For this we need to define a relative risk for the ith region: θ_i. We usually want to make inferences about the relative risk, in any study.

We also usually have available an expected rate for the ith region: e_i. Often the count within the regions will have a Poisson distribution, i.e., $y_i \sim Pois(e_i\theta_i)$.

17.3. LIKELIHOOD AND BAYESIAN MODELS

17.3.1. Likelihood

A random variable X is usually associated with a distribution which governs its behavior. We denote this distribution as $f(x \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a parameter. In general, θ could be a vector of parameters and so is denoted $\boldsymbol{\theta}$. In this case we have $f(x \mid \boldsymbol{\theta})$. When a random sample of values of X are taken $\{x_i, i = 1, \ldots, n\}$ then the likelihood is defined as the joint distribution of the sample values:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}). \tag{17.2}$$

It is assumed that conditional on $\boldsymbol{\theta}$ the sample values are independent. If this were not so, then we would require to take the product of conditional distributions in equation (17.2). When using the frequentist inferential process it is important to base decisions about parameters (estimation of parameter values or confidence intervals) on the likelihood function. Maximum likelihood estimation seeks point estimates of the parameters in $\boldsymbol{\theta}$ by maximising $f(\mathbf{x} \mid \boldsymbol{\theta})$ or $\log f(\mathbf{x} \mid \boldsymbol{\theta})$. Testing and interval estimation is often based on likelihood ratios derived for different values of $\boldsymbol{\theta}$ under different hypotheses. Inference for quantities such as confidence intervals is based on the concept of repeated experimentation, in that probability statements are derived based on properties of repeated sequences of experiments.

17.4. BAYESIAN INFERENCE

Fundamental philosophical differences with the frequentist approach are found when a Bayesian perspective is assumed. First of all, parameters within Bayesian models are assumed to be random variables and hence are governed by distributions themselves. Hence, there is no longer a fixed (true) value for a given parameter. Instead an expected value or other functional of a distribution can be defined. Because parameters have distributions then the likelihood previously defined must be extended to accommodate these distributions.

By modeling both the observed data and any unknown parameter or other unobserved effects as random variables, the hierarchical Bayesian approach to statistical analysis provides a cohesive framework for combining complex data models and external knowledge or expert opinion (e.g., Berger, 1985; Carlin and Louis, 2000; Robert, 2001; Gelman *et al.*, 2004; Lee, 2005) In this approach, in addition to specifying the distributional model $f(\mathbf{y} \mid \boldsymbol{\theta})$ for the observed data $\mathbf{y} = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, we suppose that $\boldsymbol{\theta}$ is a random quantity sampled from a *prior* distribution $p(\boldsymbol{\theta} \mid \lambda)$, where λ is a vector of hyperparameters. Inference concerning $\boldsymbol{\theta}$ is then based on its *posterior* distribution:

$$p(\boldsymbol{\theta} \mid \mathbf{y}, \lambda) = \frac{p(\mathbf{y}, \boldsymbol{\theta} \mid \lambda)}{p(\mathbf{y} \mid \lambda)} = \frac{p(\mathbf{y}, \boldsymbol{\theta} \mid \lambda)}{\int p(\mathbf{y}, \boldsymbol{\theta} \mid \lambda)\, \mathrm{d}\boldsymbol{\theta}}$$

$$= \frac{f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \lambda)}{\int f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \lambda)\, \mathrm{d}\boldsymbol{\theta}}. \tag{17.3}$$

Notice the contribution of both the data (in the form of the likelihood $f(\mathbf{y} \mid \boldsymbol{\theta})$) and the external knowledge or opinion (in the form of the prior $p(\boldsymbol{\theta} \mid \lambda)$) to the posterior. If λ is known, this posterior distribution is fully specified; if not, a second-stage prior distribution (called a *hyper-prior*) may be specified for it, leading to a *fully Bayesian* analysis. Alternatively, we might simply replace λ by an estimate $\hat{\lambda}$ obtained as the value which maximizes the marginal distribution $p(\mathbf{y} \mid \lambda)$ viewed as a function of λ. Inference proceeds based on the estimated posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y}, \hat{\lambda})$, obtained by plugging $\hat{\lambda}$ into equation (17.3). This is called an *empirical Bayes* analysis and is closer to maximum likelihood estimation techniques.

The Bayesian decision-making paradigm improves on the classical approaches to statistical analysis in its more philosophically sound foundation, its unified approach to data analysis, and its ability to formally incorporate prior opinion or external empirical evidence into the results via the prior distribution. Statisticians, formerly reluctant to adopt the Bayesian approach due to general skepticism concerning its philosophy and a lack of necessary computational tools, are now turning to it with increasing regularity as classical methods emerge as both theoretically and practically inadequate. Modeling the θ_is as random (instead of fixed) effects allows us to induce specific (e.g., spatial, temporal or more general) correlation structures among them, hence among the observed data y_i as well. Hierarchical Bayesian methods now enjoy broad application in the analysis of complex systems, where it is natural to pool information across different sources e.g., Gelman *et al.* (2004).

Modern Bayesian methods seek complete evaluation of the posterior distribution using simulation methods that draw samples from the posterior distribution. This sampling-based paradigm enables *exact* inference free of unverifiable asymptotic assumptions on sample sizes and other regularity conditions. A computational challenge in applying Bayesian methods is that for many complex systems, the simulations required to do inference under equation (17.3) generally involve distributions that are intractable in closed form, and thus one needs more sophisticated algorithms to sample from the posterior. Forms for the prior distributions (called *conjugate* forms) may often be found which enable at least partial analytic evaluation of these distributions, but in the presence of nuisance parameters (typically unknown variances), some intractable distributions remain. Here the emergence of inexpensive, high-speed computing equipment and software comes to the rescue, enabling the application of recently developed MCMC integration methods, such as the Metropolis–Hastings algorithm (Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984; Robert and Casella, 2005). Univariate MCMC algorithms are particularly attractive for general purpose implementation, since all that is required is the ability to sample easily from each parameter's complete conditional distribution, namely $p(\theta_i \,|\, \mathbf{y}, \theta_{j \neq i})$, $i = 1, \ldots, k$. The recently developed `WinBUGS` language (`www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`) and the `R` statistical platform (`www.r-project.org`) with its Bayesian packages are promising steps towards a general purpose software package for hierarchical modeling, though it may be insufficiently general in some advanced analysis settings, and in any case more work is needed before it is suitable for routine use by statistical support staff.

Statistical prediction in Bayesian settings is particularly elegant and intuitive. Let $\mathbf{y}_{\text{pred}}$ denote the random variables (they can be a collection) we seek to predict. Then, we simply treat $\mathbf{y}_{\text{pred}}$ as a random variable whose *prior*, conditional upon the parameters, is the data likelihood $f(\mathbf{y} \,|\, \boldsymbol{\theta})$. Then, all predictions will be summarized in the *posterior predictive* distribution:

$$p(\mathbf{y}_{\text{pred}} \,|\, \mathbf{y}) = \int f(\mathbf{y}_{\text{pred}} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}.$$

Once the posterior samples are available from $p(\boldsymbol{\theta} \,|\, \mathbf{y})$, it is routine to draw samples from $p(\mathbf{y}_{\text{pred}} \,|\, \mathbf{y})$ using the principle of *composition*: for each posterior draw of $\boldsymbol{\theta}$, we draw $\mathbf{y}_{\text{pred}}$ from $f(\mathbf{y}_{\text{pred}} \,|\, \boldsymbol{\theta})$. Details of such methods are particularly well explained in the texts by Carlin and Louis (2000) and Gelman *et al.* (2004).

17.4.1. Posterior sampling methods

Practical Bayesian modeling relies upon efficient computation of the posterior distribution of the parameters. As mentioned above, the main computational challenge lies in evaluating the integral in the denominator of equation (17.3). This is especially compounded when $\boldsymbol{\theta}$ is multi-dimensional. Hence, instead of designing multi-dimensional integration routines, even the best of which can easily prove inadequate for several practical settings, we focus upon *sampling* from the posterior distribution, also known as *simulating* the posterior distribution. Once a posterior sample is obtained, all inference summaries (e.g., point estimates and credible intervals) are calculated using the sample. In principle, this strategy works equally well for simpler models where the posterior distribution is a standard family as well as for very complex hierarchical models where the posterior distribution is highly complex. Depending upon the complexity of the posterior distribution, the sampling strategies will vary: with a standard family we can directly draw a random sample, while with complex families more elaborate MCMC algorithms (see below) may be required.

Since the posterior distribution now describes the behavior of the parameters once the data are observed, we work with this distribution for estimation and inference. To obtain estimates of parameters this distribution must be summarized.

A simple example of this type of model in disease mapping is where the data likelihood is Poisson and there is a common relative risk parameter with a single gamma prior distribution:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\mathbf{y} \mid \theta) g(\theta)$$

where $g(\theta)$ is a gamma distribution with parameters α, β, i.e., $G(\alpha, \beta)$, and $L(\mathbf{y} \mid \theta) = \prod_{i=1}^{m} \{(e_i\theta)^{y_i} \exp(-e_i\theta)\}$ bar a constant only dependent on the data. A compact notation for this model is:

$$y_i \mid \theta \sim Pois(e_i\theta)$$

$$\theta \sim G(\alpha, \beta).$$

Here, the posterior distribution is again a Gamma and one can sample from it by simply employing a Gamma random number generator.

Another useful mechanism for posterior simulations when the posterior distribution is not a standard family arises from the principle of *composition*. This essentially observes that the joint posterior distribution of two arbitrary parameter vectors, say $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be expressed as $P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \mathbf{y}) = P(\boldsymbol{\theta}_1 \mid \mathbf{y}) P(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \mathbf{y})$. To obtain samples from the above joint posterior distribution, we first sample $\boldsymbol{\theta}_1^{(j)}$ from the *marginal posterior* distribution $P(\boldsymbol{\theta}_1 \mid \mathbf{y})$ and then sample a $\boldsymbol{\theta}_2^{(j)}$ from the *conditional posterior* distribution $P(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(j)}, \mathbf{y}))$. Repeating this for $j = 1, \ldots, M$ results in a joint posterior sample $(\boldsymbol{\theta}_1^j, \boldsymbol{\theta}_2^{(j)})_{j=1}^M$ of size M. We illustrate this principle below using the linear regression model mentioned in equation (17.1) from a Bayesian perspective. Several other examples can be found in the texts by Carlin and Louis (2000) and Gelman *et al.* (2004).

Let us suppose that we have data y_i from n experimental units, which forms our dependent variable. Suppose also that we have observed p covariates, $x_{1i}, \ldots, x_{pi}$, on the ith individual. Using matrix notations, we write:

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}; \quad \mathbf{e} \sim N(0, \sigma^2 I)$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, X is a $n \times p$ matrix of independent predictors with full column rank (we assume independent columns – so that covariates are not collinear), $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\mathbf{e}$ is the $n \times 1$ vector of uncorrelated normally distributed errors with common variance σ^2.

To construct a Bayesian framework, we will need to assign a prior distribution for $(\boldsymbol{\beta}, \sigma^2)$ in the above model. For illustration, consider the non-informative or *reference* prior distribution for $(\boldsymbol{\beta}, \sigma^2)$:

$$P(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

This is equivalent to a flat or Uniform prior on $(\boldsymbol{\beta}, \sigma^2)$. In hierarchical language we write the Bayesian linear regression model as:

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{0}, \sigma^2 I)$$

$$\boldsymbol{\beta}, \sigma^2 \sim P(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Simple computations (see, e.g., Gelman *et al.*, 2004, Section 14.2) reveal that the marginal distribution $p(\sigma^2 \mid \mathbf{y})$ is a scaled Inv-$\chi^2(n-p, s^2)$ distribution, which is the same as the Inverse-Gamma distribution $IG((n-p)/2, (n-p)s^2/2)$ where:

$$s^2 = \frac{1}{n-p}(\mathbf{y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}})$$

with $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$ being the usual least-squares estimate (also the MLE). The distribution $P(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y})$ is $N(\hat{\boldsymbol{\beta}}, \sigma^2(X^TX)^{-1})$. In fact, here the marginal posterior distribution for $P(\boldsymbol{\beta} \mid \mathbf{y})$ can be derived in closed form as a multivariate-*t* distribution (see, e.g., Robert, 2001) but we outline the sampling-based perspective.

Following the principle of composition sampling, we draw, say for $j = 1, \ldots, M$, $\sigma^{2(j)} \sim IG(n - p/2, (n-p)s^2)$ followed by $\boldsymbol{\beta}^{(j)} \sim N(\hat{\boldsymbol{\beta}}, \sigma^{2j}(X^TX)^{-1})$. This yields our desired posterior sample $(\boldsymbol{\beta}^{(j)}, \sigma^{2(j)})$ with $j = 1, 2, \ldots, M$. Posterior confidence intervals and all inference will again be carried out using these samples.

17.5. HIERARCHICAL MODELS

The idea that the values of parameters could arise from distributions is a fundamental feature of Bayesian methodology and leads naturally to the use of models where parameters arise within hierarchies. In the Poisson-gamma example there is a two level hierarchy: θ has a $G(\alpha, \beta)$ distribution at the first level of the hierarchy and α will have a hyperprior distribution (h_α) as will $\beta(h_\beta)$, at the second level of the hierarchy. This can be written as:

$$y_i \mid \theta \sim Pois(e_i\theta)$$

$$\theta \mid \alpha, \beta \sim G(\alpha, \beta)$$

$$\alpha \mid \nu \sim h_\alpha(\nu)$$

$$\beta \mid \rho \sim h_\beta(\rho).$$

Clearly it is important to terminate a hierarchy at an appropriate place, otherwise one could always assume an infinite hierarchy of parameters. Usually the cut-off point is chosen to lie where further variation in parameters will not affect the lowest level model. At this point the parameters are assumed to be fixed. For example, in the gamma-Poisson model if you assume α and β were fixed then the Gamma prior would be fixed and the choice of α and β would be uninformed. The data would not inform about the distribution at all. However, by allowing a higher level of variation i.e., hyperpriors for α, β, then we can fix the values of ν and ρ without heavily influencing the lower level variation. This allows the data to inform more about the different parameters in the lower levels of the hierarchy.

17.6. MARKOV CHAIN MONTE CARLO METHODS

Markov chain Monte Carlo (MCMC) methods are a set of methods which use iterative simulation of parameter values within a Markov chain. The convergence of this chain to a stationary distribution, which is assumed to be the posterior distribution, must be assessed.

Prior distributions for the p components of θ are defined as $g_i(\theta_i)$ for $i = 1, \dots, p$. The posterior distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$ is defined as:

$$P(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\mathbf{y} \mid \boldsymbol{\theta}) \prod_i g_i(\theta_i). \tag{17.4}$$

The aim is to generate a sample from the posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{y})$. Suppose we can construct a Markov chain with state space $\boldsymbol{\theta}_c$, where $\boldsymbol{\theta} \in \boldsymbol{\theta}_c \subset \Re^k$. The chain is constructed so that the equilibrium distribution is $P(\boldsymbol{\theta} \mid \mathbf{y})$, and the chain should be easy to simulate from. If the chain is run over a long period, then it should be possible to reconstruct features of $P(\boldsymbol{\theta} \mid \mathbf{y})$ from the realized chain values. This forms the basis of the MCMC method, and algorithms are required for the construction of such chains. A selection of recent literature on this area is found in Ripley (1987), Besag and Green (1993), Gelman *et al.* (2004), Gamerman (2000) and Robert and Casella (2005).

The basic algorithms used for this construction are:

1 the Metropolis and its extension, the Metropolis–Hastings algorithm;

2 the Gibbs Sampler algorithm.

17.6.1. Metropolis and Metropolis–Hastings algorithms

In all MCMC algorithms, it is important to be able to construct the correct *transition probabilities* for a chain which has $P(\boldsymbol{\theta} \mid \mathbf{y})$ as its equilibrium distribution. A Markov chain consisting of $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \dots, \boldsymbol{\theta}^t$ with state space Θ and equilibrium distribution $P(\boldsymbol{\theta} \mid \mathbf{y})$ has transitions defined as follows.

Define $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ as a transition probability function, such that, if $\boldsymbol{\theta}^t = \boldsymbol{\theta}$, the vector $\boldsymbol{\theta}^t$ drawn from $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is regarded as a proposed possible value for $\boldsymbol{\theta}^{t+1}$.

17.6.2. Metropolis and Metropolis–Hastings updates

In this case choose a symmetric proposal $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and define the transition probability as:

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')q(\boldsymbol{\theta}, \boldsymbol{\theta}') & \text{if } \boldsymbol{\theta}' \neq \boldsymbol{\theta} \\ 1 - \sum_{\boldsymbol{\theta}''} q(\boldsymbol{\theta}, \boldsymbol{\theta}'')\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'') & \text{if } \boldsymbol{\theta}' = \boldsymbol{\theta} \end{cases}$$

where $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, P(\boldsymbol{\theta}' \mid \mathbf{y})/P(\boldsymbol{\theta} \mid \mathbf{y})\right\}$.

In this algorithm a proposal is generated from $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and is accepted with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$. The acceptance probability is a simple function of the ratio of posterior distributions as a function of $\boldsymbol{\theta}$ values.

The proposal function $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ can be defined to have a variety of forms but must be an irreducible and aperiodic transition function.

Metropolis–Hastings (M–H) is an extension to the Metropolis algorithm where the proposal function is not confined to symmetry and:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{P(\boldsymbol{\theta}' \mid \mathbf{y})\, q(\boldsymbol{\theta}', \boldsymbol{\theta})}{P(\boldsymbol{\theta} \mid \mathbf{y})\, q(\boldsymbol{\theta}, \boldsymbol{\theta}')}\right\}.$$

Some special cases of chains are found when $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ has special forms. For example, if $q(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}', \boldsymbol{\theta})$ then the original Metropolis method arises and further, with $q(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}')$ (i.e., when no dependence on the previous value is assumed) then:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{w(\boldsymbol{\theta}')}{w(\boldsymbol{\theta})}\right\}$$

where $w(\boldsymbol{\theta}) = P(\boldsymbol{\theta} \mid \mathbf{y})/q(\boldsymbol{\theta})$ and $w(.)$ are importance weights. One simple example of the method is $q(\boldsymbol{\theta}') \sim$ Uniform $(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$ and $g_i(\theta_i) \sim$ Uniform $(\boldsymbol{\theta}_{ia}, \boldsymbol{\theta}_{ib})\,\forall i$; this leads to an acceptance criterion based on a likelihood ratio. Hence the original Metropolis algorithm with uniform proposals and prior distributions leads to a stochastic exploration of a likelihood surface. This, in effect, leads to the use of prior distributions as proposals. However, in general, when the $g_i(\theta_i)$ are not uniform this leads to inefficient sampling. The definition of $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ can be quite general in this algorithm and, in addition, the posterior distribution only appears within a ratio as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Hence, the distribution is only required to be known up to proportionality.

17.6.3. Gibbs updates

The Gibbs Sampler has gained considerable popularity, particularly in applications in medicine, where hierarchical Bayesian models are commonly applied (see, e.g., Gilks *et al.* (1993)). This popularity is mirrored in the availability of software that allows its application in a variety of problems (e.g., WinBUGS, MLWin, BACC). This sampler is a special case of the Metropolis–Hastings algorithm where the proposal is generated from the conditional distribution of θ_i given all other $\boldsymbol{\theta}$s, and the resulting proposal value is accepted with probability 1.

More formally, define:

$$q(\theta_j, \theta_j') = \begin{cases} p(\theta_j^* \mid \theta_{-j}^{t-1}) & \text{if } \theta_{-j}^* = \theta_{-j}^{t-1} \\ 0 & \text{otherwise} \end{cases}$$

where $p(\theta_j^* \mid \theta_{-j}^{t-1})$ is the conditional distribution of θ_j given all other $\boldsymbol{\theta}$ values (θ_{-j}) at time $t-1$. Using this definition it is straightforward to show that:

$$\frac{q(\boldsymbol{\theta}, \boldsymbol{\theta}')}{q(\boldsymbol{\theta}', \boldsymbol{\theta})} = \frac{P(\boldsymbol{\theta}' \mid \mathbf{y})}{P(\boldsymbol{\theta} \mid \mathbf{y})}$$

and hence $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1$.

17.6.4. M–H versus Gibbs algorithms

There are advantages and disadvantages to M–H and Gibbs methods. The Gibbs Sampler provides a *single* new value for each $\boldsymbol{\theta}$ at each iteration, but requires the evaluation of a conditional distribution. On the other hand the M–H step does not require evaluation of a conditional distribution but does not guarantee the acceptance of a new value. In addition, block updates of parameters are available in M–H, but not usually in Gibbs steps (unless joint conditional distributions are available). If conditional distributions are difficult to obtain or computationally expensive, then M–H can be used and is usually available.

In summary, the Gibbs Sampler may provide faster convergence of the chain if the computation of the conditional distributions at each iteration are not time consuming. The M–H step will usually be faster at each iteration, but will not necessarily guarantee exploration. In straightforward hierarchical models where conditional distributions are easily obtained and simulated from, then the Gibbs Sampler is likely to be favored. In more complex problems, such as many arising in spatial statistics, resort may be required to the M–H algorithm.

17.6.5. *Special methods*

Alternative methods exist for posterior sampling when the basic Gibbs or M–H updates are not feasible or appropriate. For example, if the range of the parameters is restricted then slice sampling can be used (Robert and Casella, 2005, Ch. 7; Neal, 2003). When exact conditional distributions are not available but the posterior is log-concave then adaptive rejection sampling algorithms can be used. The most general of these algorithms (ARS algorithm; Robert and Casella, 2005, pp. 57–59) has wide applicability for continuous distributions, although they may not be efficient for specific cases. Block updating can also be used to effect in some situations. When generalized linear model components are included then block updating of the covariate parameters can be effected via multivariate updating.

17.6.6. *Convergence*

MCMC methods require the use of diagnostics to assess whether the iterative simulations have reached the equilibrium distribution of the Markov chain. There are a wide variety of methods now available to assess convergence of chains within MCMC. algorithms (ARS algorithm; Robert and Casella, 2005, pp. 57–59) provide recent reviews. The available methods are largely based on checking the distributional properties of samples from the chains.

17.7. MODEL GOF MEASURES

It is inevitable that our statistical analysis will entail the fitting and comparison of a variety of models. For this purpose, we will need to attend to issues concerning model adequacy and model comparison. To compare between the different models and perhaps help us choose those that provide better fits, we will use the Deviance Information Criteria (DIC) (Spiegelhalter *et al.*, 2002) as a measure of model choice. The DIC has nice theoretical properties for a very wide class of likelihoods since it provides an estimate of goodness-of-fit and for model complexity and is particularly convenient to compute from posterior samples. This criterion is the sum of the Bayesian deviance (a measure of model fit) and the (effective) number of parameters (a penalty for model complexity). It rewards better fitting models through the first term and penalizes more complex models through the second term, with lower values indicating favorable models for the data. The deviance, up to an additive quantity not depending upon the parameters $\boldsymbol{\theta}$, is simply minus twice the log-likelihood, $D(\boldsymbol{\theta}) = -2 \log f(\mathbf{y} \mid \boldsymbol{\theta})$, where $f(\mathbf{y} \mid \boldsymbol{\theta})$ is the first stage likelihood for the respective model. The Bayesian deviance is the posterior mean, $\overline{D(\boldsymbol{\theta})} = E_{\boldsymbol{\theta} \mid \mathbf{y}}[D(\boldsymbol{\theta})]$, while the effective number of parameters is given by $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$. The DIC is then given by $\overline{D(\boldsymbol{\theta})} + p_D$ and is easily computed from the posterior samples.

We also often use predictive fits to assess model performance using the posterior predictive distributions. We will employ the posterior predictive loss approach (Gelfand and Ghosh, 1998) to identify models providing the best fit. The actual computations are very similar to the predictive paradigm discussed towards the end of Section 17.2. Here, for any given model, if $\boldsymbol{\theta}$ is the set of parameters, the posterior predictive distribution of a *replicated* data set is given by:

$$P(\mathbf{y}_{\text{rep}} \mid \mathbf{y}) = \int P(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta} \mid \mathbf{y})\, \mathrm{d}\boldsymbol{\theta}$$

where $P(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$ has the same distribution as the data likelihood. Replicated data sets from the above distribution are easily obtained by simulating a replicated data set from the above distribution. Preferred models will perform well under a decision-theoretic *balanced loss function* that penalizes both departure from corresponding observed values (lack of fit), as well as from what we expect the replicates to be (variation in replicates). Measures for these two criteria are evaluated as $G = (\mathbf{y} - \boldsymbol{\mu}_{\text{rep}})^T(\mathbf{y} - \boldsymbol{\mu}_{\text{rep}})$ and $P = \text{tr}\,(\text{Var}\,(\mathbf{y}_{\text{rep}}) \mid \mathbf{y})$, where $\boldsymbol{\mu}_{\text{rep}} = E[\mathbf{y}_{\text{rep}} \mid \mathbf{y}]$ is the posterior predictive mean for the replicated data points, and P is the trace of the posterior predictive dispersion matrix for the replicated data; both of these are easily computed from the samples drawn. Gelfand and Ghosh (1998) suggest using the score $D = G + P$ as a model selection criterion, with lower values of D indicating better models.

Using these formal statistical methods, we will be able to enhance the accuracy of the outputs of computer models, compare between them to validate an underlying scientific hypothesis and provide predictions of complex systems.

17.8. UNIVARIATE SPATIAL PROCESS MODELS

17.8.1. Ingredients of a Gaussian process

As briefly mentioned in the Introduction, modeling of point-referenced spatial data typically proceeds from a spatial random field $\{w(s) : s \in \mathcal{D}\}$, where $\mathcal{D}$ is typically an open subset of $\Re^d$ where d is the dimension; in most practical settings $d = 2$ or $d = 3$. We say that a random field is a *valid* spatial process if for an any finite collection of sites $\mathcal{S} = \{s_1, \ldots, s_n\}$ of arbitrary size, the vector $\mathbf{w} = (w(s_1), \ldots, w(s_n))$ follows a well-defined joint probability distribution.

For the practical spatial modeller, the most common specification is a *Gaussian Random Field* (GRF) or a *Gaussian Process* (GP), which additionally specifies that $\mathbf{w}$ follows a multivariate normal distribution. To be more specific, we write $w(s) \sim GP(\mu(s), C(\cdot))$ which is a Gaussian Process with a mean function $\mu(s)$, i.e., $E[w(s)] = \mu(s)$, and a *covariance function* $\text{Cov}(w(s), w(s')) = C(s, s')$. This specifies the joint distribution for a collection of sites $s_1, \ldots, s_n$ as $\mathbf{w} \sim N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu(s_i))_{i=1}^n$ is the corresponding $n \times 1$ mean vector and $\Sigma_{\mathbf{w}} = [C(s_i, s_j)]$ is the $n \times n$ covariance matrix with (i, j)th element given by $C(s_i, s_j)$.

Clearly the covariance function cannot be just any function: it needs to ensure that the resulting $\Sigma_{\mathbf{w}}$ matrix is symmetric and positive definite. Symmetry is guaranteed as long as $C(s, s')$ is symmetric in its arguments, while functions that ensure the positive-definiteness are known as positive definite functions. The important characterization of such functions, at least from a modeler's perspective, says that a real-valued function is a valid covariance function if and only if it is the characteristic function of a symmetric random variable (this is derived from a famous theorem due to Bochner). Further technical details about positive definite functions can be found in Cressie (1993), Chilés and Delfiner (1999) and Banerjee *et al.* (2004).

Since it is common for spatial data to consist of single observations from a site, we often need to assume *stationary* or *isotropic* processes for ensuring estimable models. Stationarity, in spatial modeling contexts, refers to the setting when $C(s, s') = C(s - s')$; that is, the covariance function depends upon the separation of the sites. Isotropy goes further and specifies $C(s, s') = C(\|s - s'\|)$, where $\|s - s'\|$ is the distance between the sites. Furthermore, we will parametrize the covariance function as $C(s - s') = \sigma^2 \rho(s - s')$, where $\rho(s - s')$ is called a *correlation function* and σ^2 is a spatial variance parameter. In particular, we will use the the isotropic exponential correlation function $\rho(d, \phi) = \exp(-\phi d)$, with $d = \|s - s'\|$.

17.8.2. Bayesian spatial regression and kriging

There is an expanding literature on modeling point-referenced spatial data. The most common setting assumes a response or dependent variable $Y(s)$ observed at a generic location s, referenced by a coordinate system (e.g., UTM or lat–long), along with a vector of covariates $\mathbf{x}(s)$. One seeks to model the dependent variable in a spatial regression setting such as:

$$Y(s) = \mathbf{x}^T(s)\boldsymbol{\beta} + w(s) + \varepsilon(s). \qquad (17.5)$$

The residual is partitioned into a spatial process, $w(s)$, capturing residual spatial association, and an independent process, $\varepsilon(s)$, also known as the *nugget* effect, modeling pure errors that are independently and identically distributed as $N(0, \tau^2)$, where τ^2 is a measurement error variance or micro-scale variance. The key to incorporating spatial association is by modeling $w(s)$ as a Gaussian Process with spatial variance σ^2 and a valid correlation function $\rho(\cdot, \boldsymbol{\xi})$ with ξ representing parameters that quantify correlation decay and smoothness of the resulting spatial surface.

When we have observations, $\mathbf{y} = (Y(s_1), \ldots, Y(s_n))$, from n locations, we treat the data as a partial realization of a spatial process, modeled through $w(s)$. Hence, $w(s) \sim GP(0, \sigma^2\rho(\cdot, \phi))$, is a zero-centered Gaussian Process with variance σ^2 and a valid correlation function $\rho(d, \phi)$, which depends upon inter-site distances ($d_{ij} = \|s_i - s_j\|$) and a parameter ϕ quantifying correlation decay. Also, we assume $\varepsilon(s)$ are i.i.d. $N(0, \tau^2)$. Inferential goals include estimation of regression coefficients, spatial and nugget variances, and the strength of spatial association through distances. Likelihood-based inference proceeds from the distribution of the data, $\mathbf{y} \sim N(X\boldsymbol{\beta}, \Sigma)$, with $\Sigma = \sigma^2 R(\phi) + \tau^2 I$, where X is the covariance matrix and $R(\phi)$ is the correlation matrix with $R_{ij} = \rho(d_{ij}, \phi)$. See Cressie (1993) for details, including maximum-likelihood and restricted maximum-likelihood methods, and Banerjee *et al.* (2004) for Bayesian estimation.

Statistical prediction (kriging) at a new location s_0 proceeds from the conditional distribution of $Y(s_0)$ given the data $\mathbf{y}$. Collecting all the model parameters into $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \phi, \nu)$, we note that

$$E[Y(s_0) \mid \mathbf{y}] = \mathbf{x}(s_0)^T\boldsymbol{\beta} + \boldsymbol{\gamma}^T\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) \qquad (17.6)$$

$$\mathrm{Var}[Y(s_0) \mid \mathbf{y}] = \sigma^2 + \tau^2 - \boldsymbol{\gamma}^T\Sigma^{-1}\boldsymbol{\gamma} \qquad (17.7)$$

where $\boldsymbol{\gamma} = (\sigma^2\rho(\phi; d_{01}), \ldots, \sigma^2\rho(\phi; d_{0n}))$ and $d_{0j} = \|s_0 - s_j\|$. Classical prediction computes the BLUP (Best Linear Unbiased Predictor) by substituting maximum-likelihood estimates for the above parameters. A Bayesian solution first computes a posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{y})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\xi})$ is the collection of all model parameters and then computes the posterior predictive distribution $P(Y(s_0) \mid \mathbf{y})$ by marginalizing over (averaging over) the posterior distribution, $\int P(Y(s_0) \mid \mathbf{y}, \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \mathbf{y})$.

A Bayesian framework is convenient here, driving inference assisted by proper and moderately informative priors on the weakly identified correlation function parameters. For example, for the smoothness parameter in the Matérn covariance, ν, we can follow Stein (1999) in assuming that the data cannot distinguish $\nu = 2$ and $\nu > 2$, which suggests placing a Unif$(0, 2)$ prior on ν. Usually a MCMC algorithm is required to obtain the joint posterior distribution of the parameters, but again there are different strategies to opt for. For example, we may work with the marginalized likelihood as above, $\mathbf{y} \mid \boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \sigma^2 H(\phi) + \tau^2 I)$, or we may add a hierarchy with spatial random effects, $\mathbf{w} = (w(s_1), \ldots, w(s_n))$:

$$\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{w} \sim N(X\boldsymbol{\beta} + \mathbf{w}, \tau^2 I)$$
$$\mathbf{w} \sim N(\mathbf{0}, \sigma^2 R(\phi)).$$

In either framework, a Gibbs sampler may be designed, with embedded Metropolis or slice-sampling steps, to obtain the marginal posterior distribution (see, e.g., Banerjee *et al.*, 2004). Much more complex hierarchical models have been discussed extensively in the spatial literature but, irrespective of their complexity, they mostly fit into the template we outlined above.

When we want to capture spatial and temporal associations, modeling is accomplished by envisioning a spatial process evolving through time. The literature in spatiotemporal models is quite rich (see, e.g., Cressie, 1993; Banerjee *et al.*, 2004, and the references therein). Essentially, modeling proceeds from a spatiotemporal process $w(s, t)$ in the above context, where s denotes the location, and t denotes time. Of course, appropriate assumptions on the covariance function associated with $w(s, t)$ have to be made. A popular covariance specification for spatiotemporal models is separability, which models spatiotemporal correlation functions as a product of a purely spatial and a purely temporal covariance function. These and other more general specifications may be found in Banerjee and Johnson (2006).

17.8.3. Illustration

Interest lies in predicting the relative density of eastern hemlock across the Bartlett Experimental Forest. Basal area per hectare[1] of all tree species was estimated at each of 438 forest inventory plots distributed across the domain of interest. The response variable is the fraction of estimated eastern hemlock basal area per hectare. Covariates include elevation and six spring and fall Tasseled Cap spectral components that were derived from Landsat satellite images (Kauth and Thomas, 1976).

A spatial regression model (as in equation (17.5)) was fitted to the data. We employed flat priors for the regression estimates $\boldsymbol{\beta}$ and, based on estimates from initial descriptive analyses including variograms (see, e.g., Banerjee *et al.*, 2004), we used inverted-gamma $IG(2, 0.01)$ for both the spatial variance σ^2 and the measurement error variance τ^2. The maximum distance between inventory plots is 4834.81 meters, so a uniform prior on ϕ was set so that the effective range was less than 3000 meters. Using these priors an MCMC algorithm was devised to obtain posterior samples. Gibbs updates were used for the regression parameters $\boldsymbol{\beta}$ while Metropolis updates were employed for spatial variance components (σ^2, τ^2) and the spatial range parameter ϕ.

The CODA package in R (`www.r-project.org`) was used to diagnose convergence by monitoring mixing, Gelman–Rubin diagnostics, autocorrelations, and cross-correlations. Analysis was based on three chains of 11,000 samples each. The first 1,000 samples were discarded from each chain as a part of burn-in. Subsequent parameter estimation and analysis used the remaining 30,000 (10,000 × 3) samples.

Table 17.1 presents the 95% central credible intervals for the parameter estimates based upon the posterior samples. All six covariates are significant and perhaps explain some of the spatial variation in the data, as is indicated by the spatial variance σ^2 being smaller than the measurement error variance τ^2. The spatial range is calculated as the distance beyond which the correlation function drops below 0.05; for the exponential correlation function this is approximately $3/\phi$. Finally Figure 17.1 displays an image plot of the estimated response surface overlaid with contours of the estimated spatial random effects (the $w(s)$s). The random effects serve to offset the spatially varying density of the response surface.

17.9. BAYESIAN MODELS FOR DISEASE MAPPING

In previous sections we have alluded to a simple Poisson model for disease counts. In fact, this is the basic model often assumed for small area counts of disease (in tracts, zip codes, counties, etc.). We consider two data resolutions here. First we consider case event data where, within a suitable study region (W), realization of cases arises. The locations of cases are usually residential addresses. These form a spatial point process. Often data is not available at this level of spatial resolution and aggregation to larger spatial units occurs. Aggregated counts of disease are often more readily available (e.g., from

Table 17.1 Parameter estimates for the model covariates elevation and spring and fall Tasseled Cap spectral components. Lower table provides parameter estimates for error terms σ^2 and τ^2, spatial range ϕ, and associated effective range

Parameter	*Estimate: 50% (2.5%, 97.5%)*
Intercept	−0.262 (−0.954, 0.387)
ELEV	−0.002 (−0.002, −0.001)
SPR-TC1	0.007 (0.001, 0.013)
SPR-TC2	−0.007 (−0.011, −0.003)
SPR-TC3	0.011 (0.006, 0.015)
FALL-TC1	−0.007 (−0.011, −0.003)
FALL-TC2	0.008 (0.004, 0.011)
FALL-TC3	−0.004 (−0.008, −0.001)
σ^2	0.009 (0.005, 0.016)
τ^2	0.014 (0.012, 0.018)
ϕ	0.002546 (0.001325, 0.005099)
Effective range (meters)	1178.448 (588.301, 2264.629)

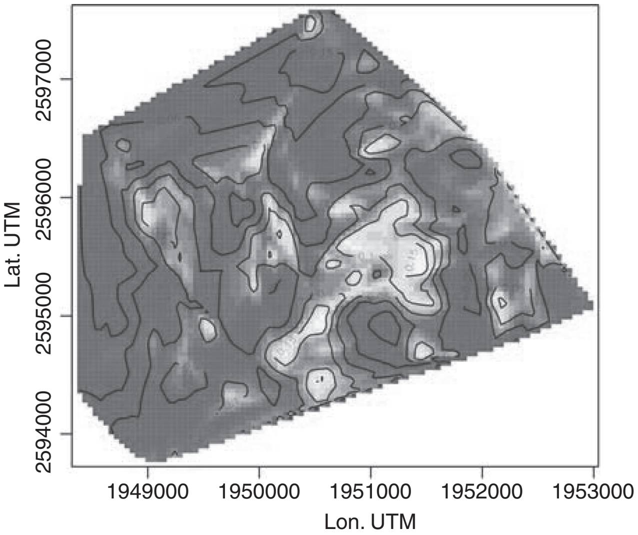

Figure 17.1 Contour lines of estimated spatial random effects overlayed on an image plot of estimated relative density of eastern hemlock. Note, the random effects serve to offset the spatially varying density of eastern hemlock.

official government sources). Hence, the second common data type is disease count data within small areas. These small areas are arbitrary with respect to the disease process (such as census tracts, counties, postcodes) and form a sub-division of the study region. In what follows we will briefly consider case event data, but will concentrate discussion on the more commonly available count data type.

17.9.1. Case event data

Assume we observe within a study region (W), a set of m cases, with residential addresses given as $\{s_i\}, i = 1, \dots , m$. Figure 17.2 displays an example of such data: larynx cancer incident case addresses for a fixed time period (see Lawson, 2006, Ch 1 for discussion). Here the random variable is the *spatial location*, and so we must employ models that can describe the distribution of locations. Often the natural likelihood model for such data is a heterogeneous Poisson Process (PP). In this model, the distribution of the cases (points) is governed by a first-order intensity function. This function, $\lambda(s)$ say, describes the variation across space of the intensity (density) of cases. This function is the basis for modeling the spatial distribution of cases. we denote this model as:

$$\mathbf{s} \sim \mathbf{PP}(\lambda(\mathbf{s})).$$

The likelihood associated with this model is given by:

$$L = \prod_{i=1}^{m} \lambda(s_i) \exp\{-\int_W \lambda(\mathbf{u})\, \mathrm{d}\mathbf{u}\}$$

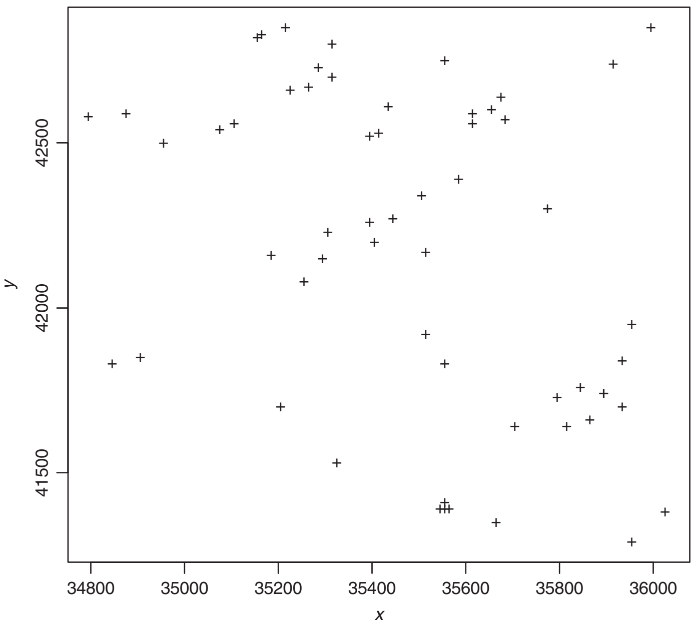

Figure 17.2 Larynx cancer incident case address locations in NW England (1974–1983).

where $\lambda(s_i)$ is the first-order intensity evaluated at the sample locations $\{s_i\}$. This likelihood involves an integral of $\lambda(\mathbf{u})$ over the study region.

In disease mapping studies, usually the variation in disease relates closely to the underlying population that is *at risk* for the disease in question. This is known as the *at risk* background. Hence any definition of the intensity of cases must make allowance for this effect. Any areas where there are lots of *at risk* people are more likely to yield cases and so we must adjust for this effect. Often the intensity is specified with a multiplicative link between these components:

$$\lambda(s) = \lambda_0(s)\lambda_1(s \mid \boldsymbol{\theta}).$$

Here the *at risk* background is represented by $\lambda_0(s)$ while the modeled excess risk of the disease is defined to be $\lambda_1(s \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. In modeling we usually specify a parametric form for $\lambda_1(s \mid \boldsymbol{\theta})$ and treat $\lambda_0(s)$ as a nuisance effect that must be included. Usually some external data is used to estimate $\lambda_0(s)$ nonparametrically (leading to profile likelihood). This data relates to the local population density. Alternatively, if the spatial distribution of a *control disease* is available (see Lawson and Cressie (2000) for more details), then the problem can be reformulated as a binary logistic regression where $\lambda_0(s)$ drops out of the likelihood. Denote the control disease locations as $\{s_j\}, j = m + 1, \ldots , m + n$, and with $N = n + m$, a binary indicator function can be defined:

$$y_i = \begin{cases} 1 & \text{if } i \in 1, \ldots , m \\ 0 & \text{otherwise} \end{cases}$$

$$\forall i, i = 1, \ldots , N$$

and the resulting likelihood is just given by:

$$L(\mathbf{s} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{[\lambda_1(s_i)]^{y_i}}{1 + \lambda_1(s_i)}.$$

By conditioning of the joint set of cases and controls the population effect is removed and does not require estimation.

17.9.2. Parametric forms

Often we can define a suitable model for excess risk within $\lambda_1(s)$. In the case where we want to relate the excess risk to a known location (e.g., a putative source of pollution) then a distance-based definition might be considered. For example:

$$\lambda_1(s) = \rho \exp\{\mathbf{F}(s)\boldsymbol{\alpha} + \gamma d_s\} \qquad (17.8)$$

where ρ is an overall rate parameter, d_s is a distance measured from s to a fixed location (source) and γ is a regression parameter, $\mathbf{F}(s)$ is a design vector with columns representing spatially-varying covariates, and $\boldsymbol{\alpha}$ is a parameter vector. The variables in $\mathbf{F}(s)$ could be site-specific or could be measures on the individual (age, gender, etc.). In addition this definition could be extended to include other effects. For example we could have:

$$\lambda_1(s) = \rho \exp\{\mathbf{F}(s)\boldsymbol{\alpha} + \eta\nu(s) + \gamma d_s\} \qquad (17.9)$$

where $\nu(s)$ is a spatial process, and η is a parameter. This process can be regarded as a random component and can include within its specification spatial correlation between sites. One common assumption concerning $\nu(s)$ is that it is a random field defined to be a spatial Gaussian process.

In the intensity (17.8), all the variables can be estimated using maximum likelihood. However when a Bayesian approach is assumed then all parameters have prior probability distributions and so we would need to consider sampling the posterior distribution given by:

$$P_1(\boldsymbol{\alpha}, \eta, \gamma \mid \mathbf{s}) \propto L(\mathbf{s} \mid \boldsymbol{\alpha}, \eta, \gamma) \cdot P_0(\boldsymbol{\alpha}, \eta, \gamma)$$

where $P_0(\boldsymbol{\alpha}, \eta, \gamma)$ is the joint prior distribution of the parameters. Assuming independent prior distributions for each parameter component, i.e., $P_0(\boldsymbol{\alpha}, \eta, \gamma) = g_{\alpha_1}(\alpha_1) \cdot g_{\alpha_2}(\alpha_2) \cdot g_{\alpha_3}(\alpha_3) \ldots g_\eta(\eta) \cdot g_\gamma(\gamma)$, this model can be sampled via standard MCMC algorithms. In intensity (17.9), the spatial component $\nu(s)$ would have a spatially correlated prior distribution and so a Bayesian approach would be natural.

17.9.3. Count data

Often only count data is available within a set of small areas. Denote y_i as the count of disease within the ith small area where $i = 1, \ldots, p$. As in the case of case event data we need to allow for the at risk population in our models. This can usually be easily achieved for count data since *expected rates* or *counts* can be obtained or calculated for small areas. For example, age × sex standardized rates for census tracts, postal zones, or zip codes are often available from government sources. Denote these rates as e_i, $i = 1, \ldots, p$. Also, in our model we want to model the *relative risk* of disease via the parameter θ_i, $i = 1, \ldots, p$. The relative risk will be the focus of modeling and it is usually assumed that the $\{e_i\}$ are fixed.

The simplest model for such data is a Poisson log linear model where:

$$y_i \sim Poiss(e_i\theta_i).$$

In addtion the relative risk θ_i is usually modeled with a log link for positivity. A simple example could be:

$$\log \theta_i = \alpha_0,$$

a constant. This model represents constant area-wide risk and often the null hypothesis aasumed by many researchers is that $\alpha_0 = 0$, so that $\theta_i = 1$. This represents the situation where the underlying rate or count generates the risk directly (i.e., $y_i \sim Poiss(e_i)$). This would be applicable if there were no excess risk in the study area. Of course this is seldom reality and it is the alternative hypotheses where θ_i have some spatilal structure that is of interest in modeling.

Some examples of models currently adopted for different applications can be instructive:

Putative health hazard assessment

Usually in these applications some measure of the association between small area counts and a fixed location or locations is to be made. This association could be via distance or directional measures. For example, define the distance from the ith small area centroid to the source as d_i and the angle as ψ_i. A log linear model for risk related to a source might be of the form:

$$\log \theta_i = \alpha_0 + \alpha_1 d_i + \alpha_2 \cos(\psi_i - \mu_0)$$

$$+ \alpha_3 \sin(\psi_i - \mu_0) + \Gamma_i.$$

Here, the directional component is summarized by the cosine and sine terms in relation to a mean angle parameter (μ_0), while the distance component is assumed to be log-linearly related to risk. The final term Γ_i is meant to repesent unattributed extra variation in risk. This could include random effect terms, such as:

$$\Gamma_i = u_i + v_i$$

where each term could represent different aspects of the extra variation. For example, u_i is often defined to have a correlated prior distribution (and is called correlated or structured heterogeneity (CH)), whereas v_i is often assumed to represent uncorrelated heterogeneity (UH). The prior distributions assumed for these terms are commonly:

$$v_i \sim N(0, \tau_v)$$

$$(u_i \mid \cdots) \propto \frac{1}{\sqrt{\beta}} \exp\left\{-\sum_{j\in\partial_i} w_{ij}(u_i - u_j)^2\right\}$$

where $w_{ij} = 1/2\beta \; \forall i, j$. The neighborhood ∂_i is assumed to be the areas with common boundary with the ith area. The second of these prior distributions assumes dependence between neighboring areas. This distribution is termed a conditional autoregressive (CAR) prior distribution. It is an example of a Markov random field. Note that in this definition the parameter β controls the spatial smoothness (or correlation) of the component.

The posterior distribution can be specified as follows:

$$P(\mathbf{u}, \mathbf{v}, \beta, \tau_v, \boldsymbol{\alpha} \mid \mathbf{y}) \propto \mathbf{L}(\mathbf{y} \mid \boldsymbol{\theta})$$

$$\times \mathbf{f}_1(\mathbf{u})\mathbf{f}_2(\mathbf{v})\mathbf{f}_3(\boldsymbol{\alpha})f(\beta)f(\tau_v)$$

where $\mathbf{f}_1(\mathbf{u})$ is the CAR prior distribution, $\mathbf{f}_2(\mathbf{v})$ is a zero mean normal distribution,

$\mathbf{f}_3(\boldsymbol{\alpha})$ is the joint prior distribution for the regression parameters, $f(\beta)$ and $f(\tau_\nu)$ are prior distributions for the remaining parameters. Note that β and τ_ν are hyperparameters and they have prior distributions as could any hyperparameters within the other prior distributions $(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)$. The prior distributions for regression parameters are often assumed to be independent and each parameter is often assumed to have a zero mean normal prior distribution.

Disease map reconstruction

Often the main aim of modeling disease incidence is simply to provide a good estimate of disease risk. This can be specified as the relative risk within each region (θ_i). Hence the aim is to provide an accurate estimate of the true underlying risk within the map. Much recent work has been focussed on this area of concern, and many models and approaches have been developed (see, e.g., Banerjee *et al.*, 2004, section 5.4; Lawson, 2006, Chapter 8.0, Lawson (2008)). Typically a log linear model with random effects is defined:

$$\log\,\theta_i = \alpha_0 + \Gamma_i \quad \text{where} \quad \Gamma_i = u_i + \nu_i.$$

Here the u_i, ν_i terms are CH and UH defined as above. This is often called the convolution model and was originally proposed by Besag *et al.* (1991). This model has proved to be very robust against mis-specification of the risk, although it can also over-smooth rates. Lawson *et al.* (2000), Best *et al.* (2005) and Hossain and Lawson (2006) have provided recent simulation-based evaluations of a range of methods in this area.

Ecological analysis

This area of focus arises when the risk within a small area is to be related to a covariate or covariates usually measured at the aggregate level. Often the main issue relates to making individual level inference from aggregate data. Aggregation or averaging induces biases in estimation of parameters for models (see, e.g., Wakefield, 2004). The *modifiable areal unit problem* (MAUP) is an example of an aggregation-related inference problem. Another problem that can arise is the *misaligned data problem (*MIDP*)*. This arises when the spatial resolution of covariates is different from the outcome variable. The classic example of this would be modeling cancer outcomes at zip code level and relating these to groundwater uranium measured at point locations (wells). A fuller discussion of these issues can be found in Banerjee *et al.* (2004). In general the type of model assumed is often of the form:

$$\log\,\theta_i = x_i^T\beta + z_i^T\xi$$

where x_i^T is a row vector of fixed covariate values for the ith small area and β is a corresponding parameter vector, and z_i^T is a row vector of random effects and ξ a unit vector.

Surveillance

With recent concerns over bioterrorism (Fienberg and Shmueli, 2005; Sosin, 2003; Lawson and Kleinman, 2005), the focus of disease surveillance has become important. Essentially this focus concerns the monitoring of disease incidence with a view to detecting aberrations or unusual incidence events. This often requires the monitoring of large scale databases of health information. In addition, the focus of the monitoring could be a range of effects. There could be a need to find clusters of disease on maps or change points in time series or some mixture of these effects in space–time. Detection of change in multiple time and spatial series is the focus. This is a challenging area that requires the use of fast computational algorithms and novel spatial-sequential inference. In essence, a range of models found in equations (17.1)–(17.3) above may need to be examined simultaneously in this analysis.

17.9.4. Example

Here we examine briefly an example of relative risk estimation. The example consists of the South Carolina incidence of congenital anomalies deaths by county for 1990. This has also been examined in Chapter 6 of Lawson *et al*. (2003). Figure 17.3 diplays the standardised mortality ratio for this disease for 1990. We are concerned to estimate the true relative risk underlying these county rates. To achieve this we propose a log linear model for the risk in each area. Hence we assume the likelihood:

$$y_i \sim Poiss(e_i \theta_i)$$

and then a log linear model of the form

$$\log \theta_i = \alpha_0 + \Gamma_i \quad \text{where} \quad \Gamma_i = u_i + v_i.$$

The two effects have the following prior distributions:

$$u_i \sim CAR(\overline{u}_{\delta_i}, \tau/n_{\delta_i})$$

where δ_i is the neighborhood of the ith area, $\overline{u}_{\delta_i}$ is the mean of u_i in the neighborhood, and n_{δ_i} is the number of neighbors, τ is the variance, and

$$v_i \sim N(0, \kappa)$$

where κ is the variance. Now α_0 is assumed to have a uniform prior distribution on a large range, while the τ and κ are variances and their inverses (precisions: $1/\tau, 1/\kappa$) have gamma prior distributions with fixed parameters (shape: 0.5, scale: 0.0005). There is some debate currently about how informative such hyperprior distributions are (see, e.g., Gelman, 2005). In fact it is always recommended that sensitivity to prior assumptions be examined in any application. The Bayes estimate of the relative risk is the posterior expected value of relative risk for

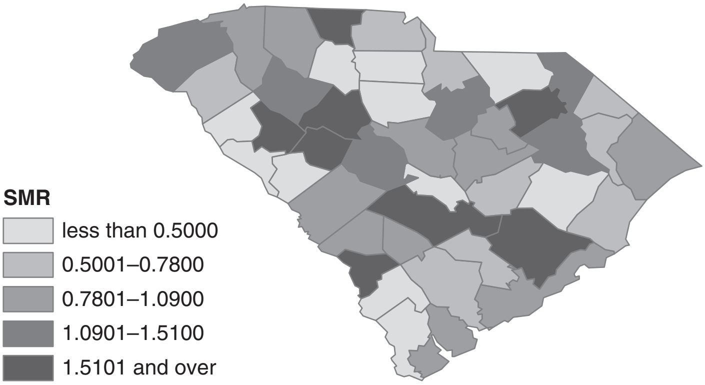

Figure 17.3 Congenital anomalies deaths, standardized mortality ratio, South Carolina, 1990.

each region. This can be obtained from a posterior sample by averaging the converged sample output. The estimates of the relative risk for the congential abnormalities data are displayed in Figure 17.4. The posterior probability of $\theta_i > 1$ over the whole map is shown in Figure 17.5 Note that this quantity can be used to assess whether ther are any areas of ‘significant’ risk elevation on the map. For more details of this example see Lawson *et al*. (2003: chapter 6).

17.10. SOFTWARE FOR BAYESIAN MODELING

Posterior sampling is the commonest approach to Bayesian inference. There is now a range of software that can peform this task. The best known of these is the free software WinBUGS (downloadable from www.mrc-bsu.cam.ac.uk/bugs/). This package employs both Gibbs sampling and Metropolis–Hastings updating methods for a

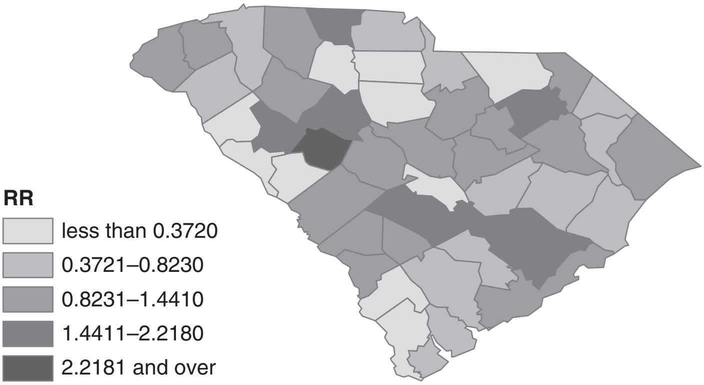

Figure 17.4 Posterior expected relative risk estimates for the congenital abnormalities data for South Carolina, 1990.

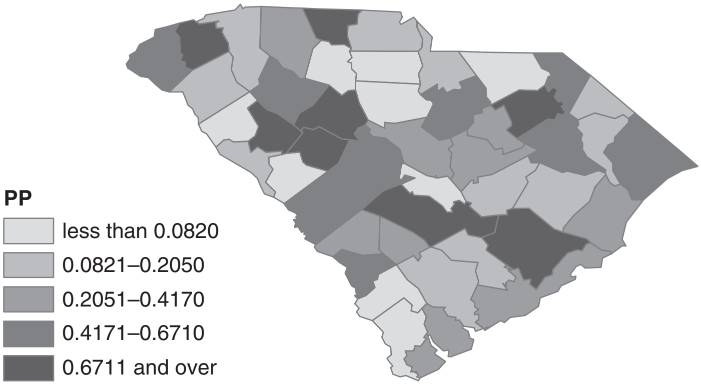

Figure 17.5 Posterior probability of exceedance ($P_r(\theta_i > 1)$) for the South Carolina congenital abnormalities data.

wide range of models. The package also has a wide range of online runnable examples and has a GIS tool called GeoBUGS that allows mapping of small area data and parameter estimates, as well as spatial modeling of various kinds. Bayesian Kriging and both CAR and multivariate CAR models can be fitted using this package. Facilities also exist within R (e.g. packages such as bayesm, geoR, geoRglm, MCMCpack, mCmC, spBayes etc.) and MATLAB (spatial statistics toolbox) to perforn MCMC computations for Bayesian spatial models.

ACKNOWLEDGMENTS

Portions of this research were based upon data generated in long-term research studies on the Bartlett Experimental Forest, Bartlett, NH, funded by the U.S. Department of Agriculture, Forest Service, Northeastern Research Station. The authors would especially like to thank Marie-Louise Smith in the USDA Forest Service Northeastern Research Station for sharing a data set and Andrew Finley in the Department of Forest Resources at the University of Minnesota for help with the statistical computations.

NOTE

[1] Basal area is the cross-sectional area of a tree at 1.37 meters from the ground. Basal area per hectare is the sum of all the basal area per tree in the hectare.

REFERENCES

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. London: Chapman and Hall/CRC Press.

Banerjee, S. and Johnson, G.A. (2006). Coregionalized Single- and Multi-resolution Spatially-varying Growth Curve Modelling with Applications to Weed Growth. *Biometrics*, 61, 617–625.

Berger, J.O. (1985). *Bayesian Decision Theory*. New York: Springer Verlag.

Besag, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, **55**: 25–37.

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

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

Carlin, B.P. and Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. London: Chapman and Hall/CRC Press.

Chen, M., Shao, Q. and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer Verlag.

Chilés and Delfiner (1999). *Geostatistics: Modelling Spatial Uncertainty*, p. 43. New York: Wiley.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, revised edition. New York: Wiley.

Fienberg, S. and Shmueli, G. (2005). Statistical issues and challenges associated with rapid detection of bio-terrorist attacks. *Statistics in Medicine*, **24**: 513–529.

Gamerman, D. (2000). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. New York: CRC Press.

Gelfand, A. and Ghosh, S. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, **85**: 1–11.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**: 1–19.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D. (2004). *Bayesian Data Analysis*. London: Chapman and Hall/CRC Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. PAMI*, **6**: 721–741.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. and Kirby, A.J.

(1993). Modelling complexity: Applications of Gibbs sampling in medicine. *Journal of the Royal Statistical Society B*, **55**: 39–52.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**: 97–109. 44

Hossain, M. and Lawson, A.B. (2006). Cluster detection diagnostics for small area health data: with reference to evaluation of local likelihood models. *Statistics in Medicine*, **25**: 771–786.

Kauth, R.J. and Thomas, G.S. (1976). The tasseled cap – a graphic description of the spectral-temporal development of agricultural crops as seen by landsat. In: *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data*, pp. 41–51. West Lafayett: Purdue University.

Lawson, A.B. (2006). *Statistical Methods in Spatial Epidemiology*, 2nd edn. New York: Wiley.

Lawson, A. B. (2008) *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. London: Chapman and Hall/CRC Press.

Lawson, A.B., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P. and Divino, F. (2000). Disease mapping models: an empirical evaluation. *Statistics in Medicine*, **19**: 2217–2242. Special issue: Disease mapping with emphasis on evaluation of methods.

Lawson, A.B., Browne, W.J. and Vidal-Rodiero, C.L. (2003). *Disease Mapping with WinBUGS and MLwiN*. New York: Wiley.

Lawson, A.B. and Cressie, N. (2000). Spatial statistical methods for environmental epidemiology. In: Rao, C.R. and Sen, P.K. (eds), *Handbook of Statistics: Bio-Environmental and Public Health Statistics*, volume 18, pp. 357–396. Amsterdam: Elsevier.

Lawson, A.B. and Kleinman, K. (eds) (2005). *Spatial and Syndromic Surveillance for Public Health*, p. 45. New York: Wiley.

Lee, P. (2005). *Bayesian Statistics*, 4th edn. London: Arnold.

Móller, J. and Waagpetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. New York: CRC/Chapman and Hall.

Neal, R.M. (2003). Slice sampling. *Annals of Statistics*, **31**: 1–34.

Ripley, B.D. (1987). *Stochastic Simulation*. New York: Wiley.

Robert, C. (2001). *The Bayesian Choice: A Decision-theoretic Motivation*. New York: Springer Verlag.

Robert, C. and Casella, G. (2005). *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.

Schabenberger, O. and Gotway, C. (2004). *Statistical Methods For Spatial Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press.

Scheiner, S.M. and Gurevitch, J. (2001). *Design and Analysis of Ecological Experiments*, 2nd edn. London: Oxford University Press.

Sosin, D. (2003). Draft framework for evaluating syndromic surveillance systems. *Journal of Urban Health*, **80**: i8–i13. supplement.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. *Journal of the Royal Statistical Society*, **64**: 583–640.

Stein, M. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging*, p. 46. New York: Springer Verlag.

Wakefield, J. (2004). A critique of statistical aspects of ecological studies in spatial epidemiology. *Environmental and Ecological Statistics*, **11**: 31–54.

Waller, L. and Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. New York: Wiley.

Webster, R. and Oliver, M. (2001). *Geostatistics for Environmental Scientists*. New York: Wiley.