Bayesian Thinking in Spatial Statistics

Lance A. Waller
Department of Biostatistics
Rollins School of Public Health
Emory University
1518 Clifton Road NE
Atlanta, GA 30322
E-mail: lwaller@sph.emory.edu
Abstract

In the sections below we review basic motivations for spatial statistical analysis, review three general categories of data structure and associated inferential questions, and describe Bayesian methods for achieving inference. Our goal is to highlight similarities across spatial analytic methods, particularly with regards to how hierarchical probability structures often link approaches developed in one area of spatial analysis to components within other areas. By choosing to highlight similarities, we focus on general concepts in spatial inference, and often defer details of several interesting and current threads of development to the relevant literature. We conclude with a listing of some of these developing areas of interest and references for further reading.
1 Why spatial statistics?

Data are always collected at particular locations whether these be in a forest, at a particular street address, in a laboratory, or in a particular position on a gene expression array. In many cases, the location may provide additional insight into circumstances associated with the data item, in short “where” we collect a measurement may inform on “what” or “how much” we measure. The field of spatial statistics involves statistical methods utilizing location and distance in inference. Some methods are extensions of familiar techniques such as regression, generalized linear models, and time series, while others derive from particular models of stochastic processes in space.

Historically, spatial issues lurk in the background of many classical statistical settings, including the design and analysis of agricultural field trials. The recognition of spatial gradients in soil properties impacting yield led to various randomization schemes to effectively remove the spatial effect from consideration. Explicit models of spatial pattern often drew motivation from similar methods in time series, with the recognition that, in space, one no longer has a simple ordering of observations. This loss of ordering complicates the extension of popular temporal models to the spatial domain, for example, Markov processes and autoregressive schemes, while important, take on additional complexity in higher dimensions, as we will see below. Key early work in spatial analysis appears in the work of Moran (1948, 1950), Whittle (1954), and Bartlett (1964, 1975). These are followed by foundational work in spatial prediction (Matheron 1963, Gandin 1963), spatial autoregressive models (Besag 1974), and spatial point processes (Ripley 1977, Diggle 1983). One measure of the increasing interest in and development of the field is a comparison of Ripley’s (1981) 252-page and Cressie’s (1993) 900-page texts addressing spatial statistics as a whole, in addition to the recent and rapid increase in texts addressing particular areas of application and/or theory (Stein 1999, Chiles and Delfiner 1999, Lawson 2001, Lawson and Denison 2002, Webster and Oliver 2001, Waller and Gotway 2004).

Much of the recent explosion in the use of spatial statistics can be tied to the relatively simultaneous increases in both spatial data availability, including accurate location
measurements via global positioning system (GPS) technology, and rapid improvements in computational power allowing advances in both spatial data structures (particularly within geographic information systems) and in the fitting of extraordinarily complex models via sophisticated algorithms. Similar advances in computational speed and algorithm development drive the increases in the use of Bayesian methods in this setting, in particular the recognition and further development of Markov chain Monte Carlo (MCMC) algorithms allowing great flexibility in fitting spatially structured statistical models. Hierarchical models provide not only a means to fit spatial models, but the hierarchical structure itself allows the statistical modeler new ways of conceptualizing spatial patterns and processes.

In the sections below, we set the stage with a review of three basic types of spatial data, their associated inferential questions, and features of spatial data which drive many of the models described below. For each data type, we outline how Bayesian hierarchical structures address the associated inferential questions. While lengthy, this review addresses only basic questions and models and we end with a brief list of current research areas in spatial statistics where Bayesian ideas aid development.

2 Features of spatial data and building blocks for inference

A primary feature driving many methods of spatial analysis is described by Tobler’s “First Law of Geography”: “Everything is related to everything else, but near things are more related than far things” (Tobler 1970). Statistically, Tobler’s “law” refers to positive spatial autocorrelation in which pairs of observations taken nearby are more alike than those taken farther apart. Allowing correlated observations weakens the usual statistical assumption of independent observations and complicates analysis in several ways. First, if we assume independent observations, any observed spatial patterns can be modeled as a spatial trend in the expected values of the observations. If we allow correlation, observed spatial patterns may be due to a trend in expected values, correlation among observations with the same expectation, or some combination of the two. Second, we lose statistical precision in correlated
data resulting in a reduced “effective sample size” (the number of independent observations containing the same amount of information as our set of correlated observations), see Cressie (1993, pp. 14-15) for an example. The presence of spatial correlation drives two major areas of spatial analysis outlined below: fitting models with residual spatial correlation and using spatial correlation to predict observations at locations where no measurement was taken.

In addition, spatial data complicate usual notions of statistical asymptotics. At least two sorts of asymptotics exist: increasing domain wherein we observe data in a study area with increasing area, or infill asymptotics wherein we observe more and more data within the same, fixed study area. Increasing domain asymptotics reflect traditional ideas more closely than infill asymptotics, but are not realistic if the study area under question already covers a large portion of the earth. As a result, there is considerably less reliance on asymptotic results in spatial statistics than in other branches of statistics.

Spatial statistical methods may be classified by the inferential questions of interest. These questions are motivated by application and often fall into categories based on the type of data available. Cressie (1993) provides three useful categories of spatial data that also serve to categorize both inferential questions and inferential approaches. We present these in order of data complexity.

First, consider spatial point process data consisting of a set of observed locations in a defined study area. We consider the locations themselves as the realization of some random process and seek inference regarding the properties of this process. Examples include the locations of trees in a forest, neurons in the brain, and galaxies in the universe. Questions of interest include:

- Are observations equally likely at all locations? If not, where are observations more or less likely?
- Are there (spatially-referenced) covariates that drive the probability of observations occurring at particular locations?
- Does the presence of an observation at a particular location either encourage or inhibit
further observations nearby?

- If observations do impact the probability of nearby observations, what is the range of influence of observations?

Much of the literature on spatial point processes involves modelling of stochastic processes, and comparison of competing models describing observed spatial patterns. Statistical questions address estimation of model parameters and assessment of fit of various models.

Next, suppose we have \textit{geostatistical} data consisting of a set of measurements taken at a fixed set of locations, e.g., ozone levels measured at each of a set of monitoring stations. In this case, the locations are set by design and not random. An inferential question of interest is prediction of the same outcome at locations where no measurement was taken. Examples of such predictions appear each day in weather maps of current temperatures interpolated from a set of official monitoring stations. The literature on spatial prediction builds from a fairly simple concept: spatial correlation suggests that one should give more weight to observations near the prediction location than to those far away. Spatial prediction theory explores how to optimally set these weights based on estimates of the underlying spatial autocorrelation structure.

Finally, we may observe data from a set of regions partitioning the study area. Such data are referred to as \textit{lattice data} by Cressie (1993) and \textit{regional data} by Waller and Gotway (2004). Lattices may be regularly or irregularly spaced, for example pixels in an image or counties within a state, respectively, so we use the term “regional” to avoid confusion with literature that assumes the term “lattice” implies a regular lattice. Regional data generally involve summary measures for each region, e.g., number of residents in an enumeration district, average income for residents of the region, or number of items delivered within a postal delivery zone. Inferential questions often involve accurate estimation of summaries from regions with small sample sizes (“small area estimation”), or regression or generalized linear modelling linking outcomes and covariates measured on the same set of regions. In the first case, statistical methods involve how best to “borrow strength” from other regions in order to improve estimates within each region. In the second, methods involve accurate estimation
of model parameters with adjustment for spatial correlation between nearby regions.

Since the inferential questions vary with data type, we describe models for each data type separately. We address methods for the data types in the reverse of the order of data complexity above since typically the analytic methods become more complex and less familiar as we move from regional regression/generalized linear model approaches, to geostatistical prediction, to point process models. While the approaches outlined derive from the different data types, we present associated models in a framework illustrating similarities in structure illustrating methodological commonalities in the statistical approaches underlying the three data structures.

3 Small area estimation and parameter estimation in regional data

We begin by addressing two primary analytic questions commonly encountered in the analysis of regional (lattice) data, namely, small area estimation and parameter estimation in generalized linear models with (residual) spatial correlation. We will illustrate how hierarchical models address both issues.

There is a large literature on small area estimation including the comprehensive review by Ghosh and Rao (1994) and the recent text by Rao (2003). We do not attempt a thorough review here, but borrow from several key approaches and link them to Bayesian hierarchical models for spatially correlated regional data. However, most of the attention in the small area estimation literature involves accurate estimation for subsets of larger samples, rather than on spatial patterns per se, although, as we will see below, adding spatial structure links the small area estimation literature and that addressing models for “disease mapping”, i.e., small area estimation of local disease incidence or prevalence rates.

In a nutshell, the structure we describe involves a hierarchical model with random effects (intercepts) inducing “borrowed” strength. The widely-cited approach of Fay and Herriot (1979) involves a Gaussian linear model with a Gaussian random intercept and provides a base model from which many small area estimation approaches derive. Briefly, consider the
setting where we observe $Y_i$ for each region $i$, $i = 1, \ldots, I$, and we assume

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma_Y^2).$$

If $\mu_1 = \mu_2 = \ldots = \mu_I = \mu$, estimation proceeds either via maximum likelihood or some design-based estimate of the shared, global population mean.

However, suppose we aren’t willing to assume equal means and instead wish to estimate each $\mu_i$? Without further assumptions, we can only base our local estimates on the observation (or perhaps observations) from region $i$, resulting in considerable reduction in sample size and accompanying loss in statistical precision. This a limitation of design-based estimation procedures and is the basis of the small area estimation problem (where “small” refers to the local sample size, not the geographic area).

If we are willing to add additional modelling assumptions, we can provide model-based or perhaps composite estimators of the local means that “borrow strength” from data in other areas to supplement each local sample size. The precise balance of input from the local and other areas depends on the particular model used. Suppose we consider the following random effects model:

$$Y_i | v_i \overset{ind}{\sim} N(\mu + v_i, \sigma_Y^2)$$

$$v_i \overset{ind}{\sim} N(0, \sigma_v^2)$$

where the zero-mean Gaussian random effects $v_i$ provide random variation among the local means, centered around the global mean $\mu$. Standard derivation of Bayesian estimators minimizing squared error loss result in a compromise “shrinkage” estimator based on a weighted average of the global sample mean and the local data where the weights depend on the relative sizes of the two variance components $\sigma_Y^2$ and $\sigma_v^2$.

The basic idea extends from the fairly straightforward Gaussian-Gaussian linear model to hierarchical generalized linear mixed models (GLMMs). As an illustration, we consider a family of hierarchical Poisson regression models first proposed by Clayton and Kaldor (1987) to provide local (small area) estimates of disease risk, based on observed disease counts.
To begin, consider the standard development of Poisson regression as a generalized linear model (GLM) with a log link and an offset. Suppose we observe a number of events, $Y_i$, and a vector of covariates, $X_i$, for each region $i, i = 1, \ldots, I$. In addition, let $\beta$ denote the vector of parameters corresponding to covariate effects and let $E_i$ denote the number of events “expected” in the absence of any covariate effects. The $E_i$ are assumed to be fixed and known and serve as offsets in the GLM effectively removing known differences in expected counts due to, e.g., varying numbers at risk or varying age structures in each region. The model is then

$$
Y_i \overset{ind}{\sim} \text{Poisson}(E_i \exp(\zeta_i)),
$$

$$
\zeta_i = X_i \beta, \text{ for } i = 1, \ldots, I.
$$

The expected number of events $E_i$ is typically defined as the product of the number of units under study in region $i$, say $n_i$, and a constant “baseline” risk, and covariates provide multiplicative increases or decreases in this risk defined by the parameters in $\beta$. The resulting Poisson GLM defines the natural logarithm of the mean count as

$$
\log[E(Y_i)] = \log(E_i) + X_i \beta,
$$

yielding a model with an offset $E_i$ and a multiplicative impact of each covariate on the model-based expected observations.

The small area estimation question of interest is whether the baseline risk of disease for individuals within a region is the same between all regions. In addition, we seek estimates of the model parameters in $\beta$.

Consider the small area estimation problem in the absence of covariates. In this case we are simply fitting an intercept-only GLM (with offset $E_i$), but with a separate intercept in each region so only the data from region $i$ can contribute to the estimate of $\zeta_i$. The addition of a mean-zero Gaussian random intercept $v_i$ to each region results in the following model

$$
Y_i|v_i \overset{ind}{\sim} \text{Poisson}(E_i \exp[\beta_0 + v_i]),
$$

$$
v_i \overset{ind}{\sim} N(0, \sigma_v^2), \text{ for } i = 1, \ldots, I.
$$
Note that the observations are conditionally independent given the values of the random effects \( v_i \), but will be marginally correlated. Similar models appear in the analysis of longitudinal data and the basic two-stage hierarchy allows one to consider the probability in two stages. A Bayesian interpretation of the model views the first stage as the likelihood (built from the product of the conditionally independent components) and views the distributions of the \( v_i \)s as a set of exchangeable prior distributions. To complete the second stage, we assign vague (even flat) priors to the “fixed” effect \( \beta_0 \), which is well identified through the likelihood.

The addition of the random effect helps our estimation problem because, in effect, we induce some commonality between the local \( \zeta_i \)s and change our estimation problem from estimating \( I \) separate \( \zeta_i \)s to estimation of two parameters: \( \beta_0 \) and \( \sigma_v^2 \). The end result is a local estimate defined by a weighted average of the observed data in location \( i \) and the global overall mean. Clayton and Kaldor (1987), Marshall (1991), and Waller and Gotway (2004, Section 4.4.3) provide details of an empirical Bayes approach using data estimates of \( \beta_0 \) and \( \sigma_v^2 \). In a fully Bayesian approach we assign a hyperprior distribution to \( \sigma_v^2 \) (e.g., a conjugate inverse gamma distribution) to complete the model.

We note the approach outlined above does not incorporate any spatial similarity between observations or error terms. All local estimates are a compromise between the local data and a global estimate based on all of the data. Clayton and Kaldor (1987) modified the hierarchical structure defining a spatially structured prior distribution for the random effects to create local empirical Bayes estimates based on a compromise between the local data and the average of regions nearby, essentially borrowing strength locally rather than globally. Besag at el. (1991) extend the approach to a fully Bayesian setting using MCMC and clarify some technical points with the spatially structured prior. While the model remains the same, to avoid confusion with the exchangeable \( v_i \)s above, we follow Besag et al. (1991) and denote the spatially correlated random effects by \( u_i, i = 1, \ldots, I \).

We describe two classes of spatially structured prior formulations. First, consider a joint
multivariate Gaussian prior distribution for \( \mathbf{u} \) with spatial covariance matrix \( \Sigma_u \), i.e.,

\[
\mathbf{u} \sim \text{MVN}(\mathbf{0}, \Sigma_u).
\]  

(1)

Diggle et al. (1998) illustrate this approach defining the elements of \( \Sigma_u \) via an exponential correlation function with one parameter governing the rate of decay with distance and another defining overall variation (impacting the overall level of smoothness in the resulting estimates). The rate of decay and smoothness parameters are then assigned independent uniform priors. While conceptually more straightforward, the jointly-specified spatial prior is only beginning to see broad application for reasons that become apparent below.

As an alternative to the joint prior distribution defined in equation (1), Clayton and Kaldor (1987) propose a set of conditionally autoregressive priors (CAR priors) for the \( u_i, i = 1, \ldots, I \). In addition, Breslow and Clayton (1993) apply CAR priors as random effects distributions within likelihood approximations for GLMMs. The CAR priors see wide use in the area of disease mapping, or regional smoothing of rates (proportions) of rare events. Clayton and Bernardinelli (1992), Mollié (1996) and Wakefield et al. (2000) review the application of such models to disease mapping issues.

The widespread application of CAR models merits a closer look at their structure and implementation, paraphrased from the development in Waller and Gotway (2004, pp. 409-423). First, consider a subset of Gaussian CAR priors defining the prior mean of each \( u_i \) as a weighted average of the other \( u_j, j \neq i \),

\[
u_i|u_{j\neq i} \sim \text{N} \left( \frac{\sum_{j\neq i} c_{ij} u_j}{\sum_{j\neq i} c_{ij}}, \frac{1}{\tau_{\text{CAR}} \sum_{j\neq i} c_{ij}} \right), \quad i = 1, \ldots, N.
\]

(2)

Here, the \( c_{ij} \)s denote spatial dependence parameters defining which regions \( j \) are “neighbors” to region \( i \), and \( \tau_{\text{CAR}} \) denotes a hyperparameter related to the conditional variance of \( u_i \) given the other elements of \( \mathbf{u} \). We set \( c_{ii} = 0 \) for all \( i \), so regions are not neighbors to themselves. Many applications utilize adjacency-based weights where \( c_{ij} = 1 \) if region \( j \) is adjacent to region \( i \), and \( c_{ij} = 0 \), otherwise. The \( c_{ij} \) are typically considered fixed and defined by the physical or political geography of the study area. Other weighting options appear in the literature (Best et al. 1999) but are much less widely applied.
While the hierarchical spatial GLMM defined above contains sensible pieces and MCMC algorithms offer a means to fit them, the end result is a rather complicated hierarchical model and several features merit further attention.

First, consider the relationship between the joint and the conditional specifications of spatially structured priors given in equations (1) and (2), respectively. While both the joint and conditional specifications are possible, the connection between the user-defined $c_{ij}$s and the resulting (prior or posterior) spatial covariance structure of random effects in $\mathbf{u}$ rarely, if ever, provides a closed-form, functional relationship between the spatial dependence parameters $\{c_{ij}\}$, and the elements of the covariance matrix $\Sigma_u$, as we illustrate below.

Besag and Kooperberg (1995) clarify the connection between the autoregressive spatial dependence parameters and the spatial covariance matrix, and we highlight several key points from their development here. First, if $\mathbf{u}$ follows a multivariate Gaussian distribution with covariance $\Sigma_u$, then the density, $f(\mathbf{u})$, follows

$$f(\mathbf{u}) \propto \exp \left( -\frac{1}{2} \mathbf{u}' \Sigma_u^{-1} \mathbf{u} \right).$$

(3)

Multivariate Gaussian theory defines the associated conditional distributions via

$$u_i | u_{j \neq i} \sim N \left( \sum_{j \neq i} \left( \frac{-\Sigma^{-1}_{u,ij}}{\Sigma_{u,ii}^{-1}} \right) u_j, \frac{1}{\Sigma_{u,ii}^{-1}} \right),$$

(4)

where $\Sigma^{-1}_{u,ij}$ denotes the $(i, j)$th element of the precision matrix $\Sigma_u^{-1}$. Notable features include: the conditional mean for $u_i$ is a weighted sum of $u_j, j \neq i$, and the conditional variance is inversely proportional to the diagonal of the inverse of $\Sigma_u$. Going from a set of user-specified conditional Gaussian distributions to the associated joint distribution is less straightforward, requiring constraints on the weights defining the conditional mean and variance to ensure both a Gaussian joint distribution and a valid covariance matrix $\Sigma_u$ (cf., Besag 1974, Besag and Kooperberg 1995, and Arnold et al. 1999).

Both the conditional mean and the conditional variance in equation (4) depend on elements of the \textit{inverse} of the covariance matrix $\Sigma_u$. Computationally, this implies an MCMC implementation of the joint specification based on straightforward updates from full conditional distributions will involve matrix inversion at each update of the covariance param-
eters. This suggests some computational advantage to the CAR prior formulation, since
the CAR priors (effectively) limit modelling to the elements of $\Sigma_u^{-1}$, rendering the inver-
sion unnecessary. In fact, the elegant theory in Besag (1974) provides the basis for both
conditionally-specified spatial modelling and conditional updates in MCMC algorithms, al-
though the latter component was not widely appreciated for some time. In essence, one
could say that CAR priors are almost “custom-made” for MCMC implementation.

If the conditional formulation is more convenient for MCMC implementation, but the
relationship between the $c_{ij}$s and the elements of $\Sigma_u$ complicated by the matrix inversion,
how do we define sets of $c_{ij}$s that correspond to valid and reasonable spatial covariance
structures?

To address this issue, we equate the spatial dependence parameters in equation (2) to
their counterparts in equation (4) yielding

$$\frac{c_{ij}}{\sum_j c_{ij}} = \frac{-\Sigma_u^{-1}_{u,ij}}{-\Sigma_u^{-1}_{u,ii}}, \quad (5)$$

and

$$\tau_{CAR} \sum_j c_{ij} = \Sigma_u^{-1}_{u,ii}. \quad (6)$$

Therefore, $c_{ij} = \Sigma_u^{-1}_{u,ij}/\tau_{CAR}$ and symmetry of $\Sigma_u$ requires symmetry of the spatial dependence
parameters, $c_{ij}$, in the collection of CAR priors defined by equation (2).

More generally, the family of CAR priors is broader than those defined by equation (2).
The weights defining the conditional mean need not be symmetric themselves so long as the
diagonal elements of $\Sigma_u^{-1}$ compensate appropriately via

$$\left(\frac{-\Sigma_u^{-1}_{u,ij}}{-\Sigma_u^{-1}_{u,ii}}\right) \Sigma_u^{-1}_{u,ii} = \left(\frac{-\Sigma_u^{-1}_{u,ji}}{-\Sigma_u^{-1}_{u,jj}}\right) \Sigma_u^{-1}_{u,jj}. \quad (6')$$

(While equation (6) may seem trivial, recall that in practice we define the additive weights
$\left(-\Sigma_u^{-1}_{u,ij}/\Sigma_u^{-1}_{u,ii}\right)$, and conditional variances $1/\left(\Sigma_u^{-1}_{u,ii}\right)$ without regard to the specific non-diagonal
elements of $\Sigma_u^{-1}$, so verification is important for any proposed conditional structure.)

The class of CAR priors defined by equation (2) (with associated symmetric $c_{ij}$s) includes
the widely-applied adjacency weights ($c_{ij} = 1$ when regions $i$ and $j$ share a boundary, $c_{ij} = 0$
otherwise), and (symmetric) distance-decay weights such as those considered by Best et al. (1999), and we continue to limit attention to this subset of CAR priors here, for simplicity and to raise some practical issues associated with these particular prior distributions.

Results in Besag (1974) indicate the set of CAR priors defined in equation (2) uniquely defines a corresponding multivariate normal joint distribution

$$\mathbf{u} \sim \text{MVN}(\mathbf{0}, \Sigma_u), \quad (7)$$

with $\Sigma_{u,ii}^{-1} = \sum_j c_{ij}$ and $\Sigma_{u,ij}^{-1} = -c_{ij}$. However, for symmetric $c_{ij}$s, the sum of any row of the matrix $\Sigma_u^{-1}$ is zero, indicating $\Sigma_u^{-1}$ is singular (non-invertible), and the corresponding covariance matrix $\Sigma_u$ in equation (7) is not well-defined. This holds for any symmetric set of spatial dependence parameters $c_{ij}$ (including the adjacency-based $c_{ij}$s appearing in many applications). Remarkably, the singular covariance does not preclude application of the model with such weight matrices (cf., Besag and Kooperberg 1995), but it does prevent easy transition from the spatial dependence parameters to closed-form spatial covariances.

CAR priors with singular associated covariance matrices can still be applied since pairwise contrasts $u_i - u_j$ are well-identified. More explicitly, note the full class of CAR priors (including those defined by equation (2)) falls into the class of pairwise difference prior distributions defined by Besag et al. (1995). These distributions are improper priors since they only define contrasts (differences) between pairs of values $u_i - u_j, j \neq i$, but they do not identify an overall mean value for the elements of $\mathbf{u}$ (since such distributions define the value of each $u_i$ relative to the values of the others). In this case, any likelihood function based on data allowing estimation of an overall mean also allows the class of improper pairwise difference priors to generate proper posterior distributions. In practice, we assure this by adding the constraint

$$\sum_{i=1}^{N} u_i = 0.$$  

The term “intrinsic” is used when defining statistics on contrasts and autoregressions with singular covariance matrices are often referred to as “intrinsic autoregressions” (Künsch 1987, Besag and Kooperberg 1995).
The descriptions above reveal a set of conditional prior distributions defined by a sensible set of spatial dependence parameters that do not translate directly into a closed-form model of spatial correlation or covariance. In addition, we also defined a joint specification that allows direct parametric modelling of the covariance function but is computationally intensive to fit (due to the matrix inversion component). Both approaches induce spatial similarity between observations by “borrowing strength” from neighboring values, both provide posterior estimates of model parameters, and both allow for small area estimation via posterior predictive distributions. However, the conditionally specified model is much more attractive for practical MCMC implementation. At present, the literature contains many more examples of CAR priors than the multivariate normal formulation, but both are actively applied in many settings.

To this point, we have considered borrowing strength either globally or locally, but Besag et al. (1991) point out that we could include both effects in the same model via a convolution prior yielding the following model:

\[
Y_i | u_i \sim^d \text{Poisson}(E_i \exp[\beta_0 + u_i + v_i]), \\
u_i | u_j \neq i \sim N\left(\frac{\sum_{j \neq i} c_{ij} u_j}{\sum_{j \neq i} c_{ij}}, \frac{1}{\tau_{\text{CAR}} \sum_{j \neq i} c_{ij}}\right), i = 1, \ldots, N, \\
v_i \sim^d N(0, \sigma_v^2), \text{ for } i = 1, \ldots, I.
\]

To complete the hierarchy, we assign hyperpriors to parameters \(\tau_{\text{CAR}}\) and \(\sigma_v^2\). As mentioned above, typical applications define conjugate inverse gamma hyperpriors, and Ghosh et al. (1999) and Sun et al. (1999) define conditions on the inverse gamma distributions necessary to ensure proper posterior distributions. When we include both \(u\) and \(v\) in the model some care is required since \(\tau_{\text{CAR}}\) is related to the conditional variance of \(u_i | u_j \neq i\) and \(\sigma_v^2\) is related to the marginal variance of each \(v_i\), and we want to take care not to assign “unfair” prior weight to either global or local smoothing. This issue is explored by Bernardinelli et al. (1995) who note that the prior marginal standard deviation of \(v_i\) is roughly equal to the conditional standard deviation of \(u_i | u_j \neq i\) divided by 0.7, although this is only a rule of thumb and merits closer scrutiny (Banerjee et al. 2004).
While hierarchical models with CAR priors see broad application for parameter estimation and associated small area estimation for regional data, they certainly are not the only models for such data, nor necessarily optimal in any particular way. In addition to the features noted above regarding prior/posterior propriety and computational complexity, CAR-based hierarchical models are defined for the given set of regions and do not aggregate or disaggregate sensibly into larger or smaller regions, respectively, and should not be viewed as “discretized” versions of some latent, smooth random process.

In summary, the hierarchical modelling approach described above allows us to incorporate spatial correlation into generalized linear models, but to conveniently defer it to the second level of the model, thereby maintaining a relatively straightforward likelihood based on conditional independence given the random effects. In some cases it is possible to incorporate spatial correlation directly in the likelihood, but when we move away from Gaussian likelihoods, this exercise becomes much more complex (and, in fact, motivates some of the development and interesting examples/counterexamples in Besag 1974).

We now turn from Bayesian spatial extensions of familiar regression and generalized linear models to the problem of statistical prediction based on spatially located data.

4 Geostatistical prediction

Statistical prediction has long had a Bayesian flavor since under most standard criteria the optimal predictor is the conditional expectation of the predictand given the observed data, i.e., the posterior predictive mean of the unknown predictand based on a fully specified model for observations (cf. Bernardo and Smith 2000, pp. 482-483). In this section, we review the typical geostatistical data structure and briefly outline “classical” geostatistical prediction methods (discussed at an introductory level in Cressie 1989, Webster and Oliver 2001, Chapter 8 of Waller and Gotway 2004, and considered more fully in Cressie 1993, Chilès and Delfiner 1999), then discuss aspects of this approach that may be addressed by Bayesian thinking. Our outline follows closely development in Le and Zidek (1992), Handcock and Stein (1993), and the recent text by Banerjee et al. (2004). One advantage of the Bayesian
approach, as expanded by Banerjee et al. (2004), is a very general model-based view of prediction blurring distinctions between the regional models of the last section and models for geostatistical prediction, a point we return to at the conclusion of this section.

The basic structure of geostatistical data is as follows: we observe a vector of \( n \) measurements \( Z = (Z(s_1), \ldots, Z(s_n)) \) at locations \( s_1, \ldots, s_n \) in some domain \( D \). We consider the locations to be fixed by design (with the term “design” including a convenience sample of locations), and for the remainder of this section assume the vector of observations follows a multivariate Gaussian distribution with mean \( \mu \) and covariance matrix \( \Sigma_Z \), i.e.,

\[
Z \sim MVN(\mu, \Sigma_Z),
\]

It is typical to assume \( \mu = (\mu_1, \ldots, \mu_n) \) and \( \mu_i = x(s_i)'\beta \), where \( x(s_i) \) represents a vector of \( p \) covariates observed at location \( s_i \), and \( \beta \) the vector of associated parameters. The elements of \( x(s_i) \) may represent polynomial functions of location coordinates to accommodate polynomial trends in the data (a trend surface analysis), but can be viewed as any location-specific covariates linearly related to the expectation of \( Z(s) \). In addition, following the descriptions Handcock and Stein (1993) and Banerjee et al. (2004), suppose that the elements of \( \Sigma_Z \) can be described through a parametric covariance function

\[
\text{Cov}(Z(s_i), Z(s_j)) = \alpha K_\theta(s_i, s_j),
\]

where \( \alpha \) denotes a (positive) scale parameter and \( \theta \) a vector of parameters (noting that elements of \( \theta \) may also influence scale). As a result, we assume the observations \( Z \) represent a realization of a Gaussian random field and seek to use the structure of this stochastic process to provide predictive inference on elements of the field, say \( Z(s_0) \), at location \( s_0 \) where we observe no measurement.

### 4.1 Covariance functions and variograms

Before moving to prediction, a few comments on the covariance function are in order. First, if we consider a spatially-referenced data set as a partial realization of a spatial stochastic process, e.g., a Gaussian random field, and if we have a single set of observations, we essentially
have a single observation of the process, even though we have multiple observations within the process. In this case, spatial autocorrelation can provide some “repeatability” within the process allowing estimation and prediction. More formally, the process is said to be second-order stationary if the covariance function $K_\theta(s_i, s_j)$ depends only on the vector $s_i - s_j$ separating locations $s_i$ and $s_j$, and not the locations themselves. Furthermore, if $K_\theta(s_i, s_j)$ depends only on the distance $||s_i - s_j||$, we say the covariance function is isotropic. Neither stationarity nor isotropy are required for spatial prediction, but results are streamlined in these cases.

Second, to link results to many classical geostatistical techniques, we also consider the variogram defined as

$$2\gamma(h) = \text{Var}[Z(s) - Z(s + h)],$$

where $h$ denotes a spatial lag or relocation vector. Also important is the semivariogram, $\gamma(h)$, defined as one-half of the variogram. Note that the variogram is defined as the variance of contrasts of observations, similar to the the pairwise difference priors defined in the preceding section, and the term intrinsic stationarity implies the existence of a valid variogram for all locations, depending on the spatial lag $h$ but not on particular locations $s$ and $s + h$ themselves, as suggested by the notation $\gamma(h)$. Intrinsic stationarity and second-order stationarity are similar but not identical concepts where intrinsic stationarity defines a slightly broader class of spatial processes (see Cressie 1993, pp. 68, and Banerjee et al. 2004, pp. 22-24). The notion of isotropy (rotation invariance) can apply to variograms in the same manner as it does for covariance functions. The semivariogram is related to the covariance function via the relationship

$$\gamma(h) = \text{Cov}(Z(s), Z(s)) - \text{Cov}(Z(s), Z(s + h)),$$

where $\text{Cov}(Z(s), Z(s))$ naturally refers to $\text{Var}(Z(s))$ and represents

$$\lim_{||h|| \to \infty} \gamma(h),$$

the limit of the semivariogram as the lag goes to infinity. This relationship holds for the set of second-order stationary processes, but does not automatically hold for intrinsically
stationary processes unless they are also second-order stationary. We restrict attention to second-order stationary processes for the remainder of this review.

Finally, the covariance function $\text{Cov}(s, s + h)$ (which we assume is defined by parametric family $\alpha K_\theta(s, s + h)$) must define a valid covariance function, i.e., must be a positive-definite function, i.e.,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(s, s + h) \geq 0,$$

for any $a_1, \ldots, a_n$ and strict inequality if not all $a_i$ are zero. Similarly, a valid semivariogram must be negative definite, i.e.,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \gamma(s, s + h) \leq 0,$$

for $a_1, \ldots, a_n$ such that $\sum_{i=1}^{n} a_i = 0$.

Determining the validity of a proposed covariance function or semivariogram often is not a straightforward task, but there is a long list of (often isotropic) parametric families that generate valid forms for particular spatial dimensions. These are reviewed in Cressie (1993), Banerjee et al. (2004), and Waller and Gotway (2004) among other places. For illustration, we define three such families here and defer details on others to the references listed.

First, letting $h = ||h||$, consider the exponential covariance function defined by:

$$K_\theta(h) = \exp(-\theta h)$$

for $h > 0$. Here the covariance declines as a smooth exponential function of lag distance. Next, consider the Gaussian covariance function defined by:

$$K_\theta(h) = \exp(-\theta^2 h^2)$$

for $h > 0$. This covariance function levels off as $h \to 0$, implying an area of fairly strong and consistent spatial correlation for a small range of distances. In fact, the Gaussian covariance function guarantees a predicted surface that is infinitely differentiable, yielding an exceedingly smooth surface. As such, many question its applicability as a parametric covariance family (Wackernagel 2003, Banerjee et al. 2004, pp. 36-37). Finally, Stein (1999)
and Banerjee et al. (2004, p. 36) among others recommend the use of the Matérn covariance function which includes both the exponential and Gaussian covariance functions as special cases, moving between them via a parameter governing the smoothness of the underlying process. The Matérn covariance function is defined as

\[
K_{\theta}(h) = \frac{1}{2^{\nu-1} \Gamma(\nu)} (2h\phi\sqrt{\nu})^\nu B_\nu(2h\phi\sqrt{\nu})
\]

for \( h > 0 \) where \( \theta \) denotes the vector \((\nu, \phi)\), \( \Gamma(\cdot) \) the gamma function, and \( B_\nu(\cdot) \) the modified Bessel function of order \( \nu \) (for details see Abramowitz and Stegun, 1965, Chapter 9). The parameter \( \nu > 0 \) represents a smoothness parameter where the Matérn model reduces to the exponential model for \( \nu = 1/2 \) and the Gaussian model as \( \nu \to \infty \). Note that the functions \( K_{\theta}(h) \) are multiplied by the scale parameter \( \alpha \) to complete definition of \( \text{Cov}(h) \).

For each definition, the case of \( h = 0 \) merits special consideration. For a spatially continuous process, \( \gamma(0) = 0 \) since the value \([Z(s) - Z(s)] \) is identically zero (and its variance zero) for any given realization of the process. However, in practice, measurements are often made with error, or there may be very small, local discontinuities such as a small “nugget” in a distribution of gold ore in a mining application. As a result \( \lim_{h \to 0} \gamma(h) = \eta \) may not be zero, but rather some positive value, referred to in the geostatistical literature as a nugget effect. Practically speaking, a non-zero nugget effect allows local, non-spatially correlated noise in the process (somewhat in the spirit of the exchangeable noise of the regional data modelling section above) and we will note its role in prediction below.

Next, consider the semivariograms corresponding the parametric families of covariance functions defined above. For second-order stationary processes, the corresponding semivariograms are

\[
\gamma(h; \eta, \alpha, \theta) = \eta + \alpha [1 - \exp(-\theta h)] \text{ (exponential)},
\]

\[
\gamma(h; \eta, \alpha, \theta) = \eta + \alpha [1 - \exp(-\theta^2 h^2)] \text{ (Gaussian)},
\]

and

\[
\gamma(h; \eta, \alpha, \theta) = \eta + \alpha \left[ 1 - \frac{(2h\theta\sqrt{\nu})^\nu}{2^{\nu-1} \Gamma(\nu)} B_\nu(2h\theta\sqrt{\nu}) \right] \text{ (Matérn)},
\]

all for \( h > 0 \), and all equal \( \eta \), the nugget effect, as \( h \to 0 \).
4.2 Kriging: classical spatial prediction

Building from these ideas of covariance and semivariograms, we next review classical geostatistics. To begin, suppose all covariance/variogram parameters are known, the family $K_{\theta}(s_i, s_j)$ is correctly specified, all $Z(s)$ share the same mean $\mu$ (an assumption we later weaken), and we seek the best linear unbiased predictor (BLUP) of $Z(s_0)$ for location $s_0$,

$$
\tilde{Z}(s_0) = \sum_{i=1}^{n} \ell_i Z(s_i),
$$

$$
= \ell'Z.
$$

The goal is to find the vector of weights $\ell_i, i = 1, \ldots, n$, minimizing the mean square prediction error

$$
MSE = E[(\tilde{Z}(s_0) - Z(s_0))^2].
$$

Unbiasedness requires $\sum_{i=1}^{n} \ell_i = 1$ so we have a constrained optimization problem, traditionally solved via addition of a Lagrange multiplier $m$. Such predictions are referred to as kriging predictions based on the early work of South African mining engineer D. G. Krige (Krige 1951), and given a formal mathematical basis by Matheron (1963). Traditionally, we find the optimal predictions by solving the kriging equations:

$$
\ell = \Gamma^{-1} \gamma
$$

(8)

where

$$
\ell = (\ell_1, \ldots, \ell_n, m)'
$$

$$
\gamma = (\gamma(s_0 - s_1), \ldots, \gamma(s_0 - s_n), 1)'
$$

and the elements of $\Gamma$ are

$$
\Gamma_{ij} = \begin{cases} 
\gamma(s_i - s_j) & i = 1, \ldots, n; \\
1 & i = n + 1; j = 1, \ldots, n; \\
0 & i = j = n + 1.
\end{cases}
$$
So (8) becomes

\[
\begin{bmatrix}
\ell_1 \\
\ell_2 \\
\vdots \\
\ell_n \\
m
\end{bmatrix} = 
\begin{bmatrix}
\gamma(s_1 - s_1) & \cdots & \gamma(s_1 - s_n) & 1 \\
\gamma(s_2 - s_1) & \cdots & \gamma(s_2 - s_n) & 1 \\
\vdots & & \ddots & \vdots \\
\gamma(s_n - s_1) & \cdots & \gamma(s_n - s_n) & 1 \\
1 & \cdots & 1 & 0
\end{bmatrix}^{-1}
\begin{bmatrix}
\gamma(s_0 - s_1) \\
\gamma(s_0 - s_2) \\
\vdots \\
\gamma(s_0 - s_n)
\end{bmatrix}.
\]

Note that, in this formulation, the semivariogram defines the optimal weights. The matrix \(\Gamma\) utilizes the semivariogram to summarize spatial covariances between data values while the vector \(\gamma\) summarizes covariances between each data value and the prediction location \(s_0\). Also note that we require a separate set of weights \(\ell\) for each prediction location, \(s_0\). However, only the right hand side of equation (8) changes with the prediction locations (through \(\gamma\)). Since \(\Gamma\) depends only on the data locations and not on the prediction locations, we need only invert \(\Gamma\) once and then multiply by the associated \(\gamma\) vector to obtain a prediction for any \(s_0\) in \(D\).

The minimized MSPE, also known as the kriging variance, also derives from the kriging weights \(\ell\) and the vector \(\gamma\), specifically,

\[
\sigma^2_{\text{krige}}(s_0) = \ell'\gamma = \sum_{i=1}^{n} \ell_i \gamma(s_0 - s_i) + m = 2 \sum_{i=1}^{n} \ell_i \gamma(s_0 - s_i) - \sum_{i=1}^{n} \sum_{j=1}^{n} \ell_i \ell_j \gamma(s_i - s_j).
\]

Taken together, equations (8) and (9) define ordinary kriging prediction where we have a fixed and spatially constant (but unknown) mean. Next, let us consider the more general case considered above where the mean is a linear function of location-specific covariates, i.e., \(\mu = (\mu_1, \ldots, \mu_n), \mu_i = x(s_i)' \beta\). In this case, suppose our observations may be modelled as

\[
Z(s) = x(s)' \beta + \delta(s) + \varepsilon(s),
\]

the sum of a fixed (local) mean defined by observed covariates and unobserved parameters, a random intercept, \(\delta(s)\), itself modelled by a mean-zero Gaussian random field, and spatially-independent noise, \(\varepsilon(s)\). Note that the spatial correlation structure is captured through
the covariance structure of the random intercept $\delta(s)$, and additional measurement error or small-scale variation (i.e., a nugget effect) is reflected in the variance of $\varepsilon(s)$. The basic structure mirrors that discussed for longitudinal models by Diggle, Liang, and Zeger (1994, Section 5.2), and for spatial models by Diggle et al. (1998), Stein (1999), and Banerjee et al. (2004).

We now assume the parametric form of $K_\theta$ is known, but the parameter values in $\theta$ are not. In other words we seek the vector of prediction weights $\ell(\theta)$ defining the best linear unbiased predictor

$$\tilde{Z}(s_0) = \ell(\theta)'Z,$$

where the optimal weights depend on the unknown covariance parameters in $\theta$. Stein (1999, pp. 7-9) reviews classical linear prediction theory noting that the best linear prediction of $Z(s_0)$, assuming $\beta$ known is

$$\tilde{Z}(s_0) = x(s_0)'\beta + kK^{-1}(Z - X\beta),$$

where $X$ denotes the $n \times p$ design matrix with $i$th row $x(s_i)$, $k$ the $n \times 1$ vector with $i$th element $K_\theta(s_0, s_i)$, and $K$ the $n \times n$ matrix of covariances $\alpha K_\theta(s_i, s_j), i, j = 1, \ldots, n$. For $\beta$ unknown, we can replace $\beta$ by its generalized least squares estimate (the best linear unbiased estimate of $\beta$), namely,

$$\hat{\beta} = \left(X'K^{-1}X\right)^{-1}X'K^{-1}Z,$$

yielding

$$\tilde{Z}(s_0) = x(s_0)'\left(X'K^{-1}X\right)^{-1}X'K^{-1}Z + kK^{-1}\left(Z - X\left(X'K^{-1}X\right)^{-1}X'K^{-1}Z\right).$$

(11)

Rearranging terms yields the prediction weights

$$\ell(\theta) = \left[K^{-1} - K^{-1}X\left(X'K^{-1}X\right)^{-1}X'K^{-1}\right]k + K^{-1}X\left(X'K^{-1}X\right)^{-1}x(s_0),$$

(12)

which Stein (1999, pp. 7-8) shows are identical to those minimizing the MSPE under the unbiasedness constraint

$$E(\ell(\theta)'Z) = E(Z(s_0)).$$
\[ \ell(\theta)'X\beta = x(s_0)'\beta. \]

The associated (minimized) mean square error is given by
\[ K_0(s_0, s_0) - k'K^{-1}k + (x(s_0) - X'K^{-1}X)'(X'K^{-1}X)^{-1}(x(s_0) - X'K^{-1}X). \] (13)

Equations (12) and (13) define the universal kriging equations. In the case of the constant mean used to present the kriging equations above, we have \( x(s) = 1 \) for any location \( s \).

In this setting, referred to in the literature as ordinary kriging, the unbiasedness constraint implies \( k'K^{-1}X = 1 \), and equation (11) reduces to \( \ell(\theta) = K^{-1}k \), the covariance function analogue to the semivariogram-based ordinary kriging equations above.

We note that the optimal predictors are only optimal if we consider the covariance function (equivalently the semivariogram for intrinsically stationary processes) fixed and known. In practice, this is not the case and most applications of classical geostatistical prediction “plug in” estimates of semivariogram parameters as the unknown values. As a result, such approaches often omit the variation associated with semivariogram estimation in the prediction process and the prediction errors do not reflect this uncertainty. Classical solutions involve exploration of “empirical” best linear unbiased prediction or EBLUPs, considered in some detail by Zimmerman and Cressie (1992) and Harville and Jeske (1992). The desire to account for uncertainty in covariance parameter uncertainty motivated both Le and Zidek (1992) and Handcock and Stein (1993) in their proposals of Bayesian kriging which we outline in the next section.

However, before leaving our (admittedly brief) review of classical kriging, we note that there is a very wide literature on variography, the modelling and estimation of (semi)variograms and their associated parameters. Such issues are of modelling interest in its own right as well as for their role in prediction and we refer interested readers to Cressie (1993) and Chilès and Delfiner (1999) for details and relevant references.

### 4.3 Bayesian kriging

A Bayesian approach brings additional flexibility to the classical prediction framework outlined above. First of all, the issue of incorporating uncertainty in covariance parameters
follows directly from posterior inference. More specifically, Bayesian prediction derives from the posterior predictive distribution which integrates over the posterior distribution of all model parameters, i.e.,

\[
p(Z(s_0)|Z, X, x(s_0)) = \int p(Z(s_0), \beta, \alpha, \eta, \theta|Z, X, x(s_0))d\beta d\alpha d\eta d\theta,
\]

\[
= \int p(Z(s_0)|Z, X, x(s_0), \beta, \alpha, \eta, \theta)p(\beta, \alpha, \eta, \theta|Z, X)d\beta d\alpha d\eta d\theta.
\]

In many cases, this integration may be numerically or theoretically intractable leading to the use of MCMC methods for evaluation.

In addition to providing a mechanism for accounting for uncertainty in variance-covariance modelling, Bayesian thinking enters the spatial prediction process in additional ways as well. One simple advantage is the conceptual ease of accommodating functions of model parameters. For example, many researchers model environmental exposure measurements via a log-normal distribution where one treats the natural logarithm of observed values as observations from a Gaussian distribution. Classical geostatistical prediction as outlined above provides BLUP of the log-exposure values, but such are no longer linear or even unbiased predictions on the original (untransformed) scale. Adjustments exist (Cressie 1993, pp. 135-137), but must be derived for any reparameterization. In the Bayesian context, transformations are less problematical, as predictions derive from the posterior predictive distribution rather than a necessarily linear combination of observations. This is particularly evident in MCMC implementation where a sample from the posterior distribution of model parameters may be transformed to a sample from the posterior distribution of the transformed parameters with little effort.

Bayesian thinking also aids in constructing spatial models, providing a means of inference for complex, multi-level, hierarchical models. Recalling the preceding section’s discussion of how a hierarchical structure provides convenient spatial modelling opportunities for regional data (by deferring spatial structure to a second stage of the model), we see similar opportunities in Bayesian geostatistical prediction. For example, if we consider the model formulation in equation (10), the prior distribution of the spatial random effects \(\delta(s)\) defines the spatial structure of the data (or, more accurately, the spatial structure of the expected value of the
data). That is, thoughtful development of prior distributions for the random effects allows a variety of modelling options in addition to a means of “bookkeeping” components of uncertainty. In addition, hierarchical structures also allow options for efficient computation of MCMC schemes.

To see the impact of the choice of hierarchical structure consider the following formulations of the model in equation (10) outlined by Banerjee et al. (2004, pp. 131-132). First, define the initial stage of the model as

$$Z | \beta, \alpha, \theta, \eta \sim N(X'\beta, \alpha K(\theta) + \eta I)$$

where $I$ is the identity matrix and we specifically denote the dependence of the matrix $K$ on the covariance parameters in $\theta$. To complete the model one assigns prior distributions to model parameters $\beta$, $\alpha$, $\theta$, and $\eta$. Next, consider a structure with first-stage conditional on model parameters and on the conditional random effects $\delta(s)$, i.e.,

$$Z | \beta, \alpha, \theta, \eta, \delta \sim N(X'\beta + \delta, \eta I)$$

where $\delta$ denotes the vector with elements $\delta(s_i), i = 1, \ldots, n$. Now, we define a second stage of the model giving prior distributions to parameters $\beta$ and $\eta$, and

$$\delta | \alpha, \theta \sim N(0, \alpha K(\theta)),$n

then complete the model with prior distributions for $\alpha$ and $\theta$. Briefly deferring a discussion of the prior distributions for model parameters for a moment, consider implications of these two model formulations. While both formulations ultimately base inference on the joint posterior of model parameters, $p(\beta, \alpha, \theta, \eta | Z)$, this distribution is (assigning independent prior distributions for notational convenience)

$$p(Z | \beta, \alpha, \theta, \eta)p(\beta)p(\alpha)p(\theta)p(\eta)$$

in the first case and

$$p(Z | \beta, \eta, \delta)p(\delta | \alpha, \theta)p(\beta)p(\eta)p(\alpha)p(\theta)$$
in the second case, where the former is marginalized over the random effects vector $\delta$. The difference between the two formulations appears when considering MCMC implementation where Banerjee et al. (2004, p. 132) note that the marginalized formulation is often better behaved since the matrix $\alpha K(\theta)$ is often nearly singular while $\alpha K(\theta) + \eta I$ often is not.

While a Bayesian setting provides some conceptual and computational niceties, in application we must complete the hierarchical model by defining reasonable prior distributions that (a) do not overly influence predictions and (b) ensure proper posterior distributions. Models typically support vague or even uniform priors for the linear mean parameters $\beta$ but prior specification for the nugget effect and other covariance parameters ($\eta$, $\alpha$, and the elements of $\theta$) requires more thought, and a variety of approaches appear in the literature. We next briefly review a few general strategies for prior specification for covariances.

Starting broadly, one could ignore parameteric covariance functions and instead define a prior distribution on the entire dispersion matrix $\Sigma_Z$, perhaps using a conjugate inverse (or inverted) Wishart distribution (cf., Box and Tiao 1973, pp. 459-470 for general results). While conjugacy offers some theoretical attraction, note that the inverse Wishart prior on $\Sigma_Z$ no longer limits attention to specifically spatial covariance structures, and Daniels and Kass (1999) outline questionable performance of the inverse Wishart prior in some hierarchical modelling settings. There is increasing attention paid to nonconjugate Bayesian analysis of covariance/correlation matrices in general (Daniels and Kass 1999, Daniels and Kass 2001, Wong et al. 2003, and Liechty et al. 2004) and with respect to particular applications in time series (Daniels and Cressie 2001) and longitudinal analysis (Daniels and Zhao 2003). Such approaches often provide “shrinkage” between an independence structure and specified correlation structures.

Moving more specifically to spatial hierarchical models of the sort defined in equation (10) with an assumed parametric family for pairwise covariances, Banerjee et al. (2004, p. 131) recommend the general pragmatic solution of selecting proper, informative priors for each variance-covariance parameter, often starting with conjugate inverse Gamma distributions for variance/scale parameters such as $\alpha$ and $\eta$. The authors also caution that very vague
priors on such parameters may lead to essentially improper posterior distributions hence MCMC convergence failure, so some care is still required to maintain posterior propriety without unduly limiting prior ranges of parameters.

Expanding on this issue, Berger et al. (2001, 2003) move toward objective Bayesian spatial analysis by considering reference and Jeffreys’ priors for variance-covariance parameters in a Gaussian random field with no nugget effect. The authors illustrate that the Gaussian random field structure raises several interesting aspects not previously encountered in the reference prior literature. In particular, the model provides an example where Jeffreys’ prior applied independently to each component yields an improper posterior. In addition, a popular approximation in the derivation of reference priors does not hold for the multivariate Gaussian data arising in a Gaussian random field. The results of Berger et al. (2001, 2003) provide tantalizing information regarding assignment of prior distributions for spatial covariance parameters, but also point to the need for additional development.

As a final note, the model formulation outlined above extends to non-Gaussian models by adding spatial random effects inside the link function of a generalized linear model, creating a generalized linear mixed model. More specifically, consider a generalized linear model where observations $Z(s)$ follow a distribution in the exponential family and

$$g[E(Z(s_i))] = z(s_i)'\beta + \delta(s_i)$$

for link function $g(\cdot)$ where $\delta(s)$ still denotes a set of random effects modelled by a a zero-mean Gaussian random field (noting the observations are not necessarily Gaussian). In this case, the functional form of the likelihood changes to reflect the assumed distribution within the exponential family, but remains in an attractive product over the observations, conditional on the spatial random effects in $\delta(s)$ which are modelled at the second stage via a familiar spatial Gaussian form, very similar to the appealing structure of the Bayesian models for regional data presented in the preceding section. Such models fall into the class of model-based geostatistics introduced by Diggle et al. (1998) with relevant issues raised in the accompanying discussion of this paper, and further considered by Zhang (2004) and Banerjee et al. (2004).
The link in model structure between geostatistical and regional inference is appealing, but does not precisely equate the two categories of models. First, key inferential differences remain: in geostatistical data we assume that we could predict \( Z(s) \) at any (point) location \( s \in D \), while regional data represent summaries over regions, and predicting a county-specific outcome for some hypothetical county falling between existing counties makes little sense. Furthermore, geostatistical models assume a field comprised of an infinite number of random variables, some of which are observed, while regional models often consist of a finite number of random variables, most or all of which are observed. Approaches layering latent geostatistical fields under observed regional data sets have been considered and offer areas of further development (Best et al. 2000). Finally, shifting between point and regional data support involves special consideration falling under the term “change of support”, see Cressie (1993, pp. 284-289) for a description of classical approaches, Gotway and Young (2002) for a recent review, and Mugglin and Carlin (1998), Mugglin et al. (2000), and Banerjee et al. (2004, pp. 168-169, and Chapter 6) for a hierarchical Bayesian perspective.

We next consider the third area of inference, namely the analysis of spatial point processes.

5 Bayesian Thinking in Spatial Point Processes

A spatial point process is a stochastic process whose realizations consist of a set of \( n \) events occurring at locations \( (x_1, \ldots, x_n) \) in the study area of interest \( D \), where the locations and total number of events \( (n) \) are random variables. We use the notation \( x \) for realized random locations to contrast with the fixed-by-design locations \( s \) in geostatistical data. Diggle (2003) and Waller and Gotway (2004, Chapter 5) provide general introductions to the statistical analysis of spatial point processes, and Stoyan et al. (1995), Lawson and Denison (2002), and Møller and Waagepeterson (2004) provide more detailed developments and applications. Here we briefly review definitions of key elements of a few particular classes of spatial point processes, namely spatial Poisson processes, Cox processes, and log Gaussian Cox processes, illustrating Bayesian concepts arising in their definition and application.
Statistical issues in spatial point processes often wander farther from traditional statistical inference and stray more into probabilistic modelling than do methods for regional modelling or geostatistical prediction, but we highlight some familiar themes from preceding sections to link the corresponding literatures.

5.1 Some spatial point processes of interest

5.1.1 Homogeneous Poisson processes

To begin, consider a homogeneous spatial Poisson process defined by the following criteria (Diggle 1983, p. 50, Stoyan et al. 1995, p. 33):

(i) The number of events occurring within a finite region \( D \) is a random variable following a Poisson distribution with mean \( \lambda |D| \) for some positive constant \( \lambda \) and \( |D| \) denoting the area of \( A \), and

(ii) Given the total number of events, \( n \), occurring within an area \( D \), the locations of the \( n \) events represent an independent random sample of \( n \) locations, where each point (a location where an event could occur) is equally likely to be chosen as an event.

Criterion (ii) represents the general concept of complete spatial randomness (CSR) (events uniformly distributed across the study area), and criterion (i) introduces the idea of an intensity \( \lambda \) representing the expected number of events per unit area. The Poisson distribution allows the total number of events observed to vary from realization to realization while maintaining a fixed (but unknown) expected number of events per unit area. Dividing the total number of events observed by the total area provides a straightforward estimate of \( \lambda \), i.e., \( \hat{\lambda} = n/|D| \). This estimate serves well for many applications, however we note that other estimators can provide better performance in certain situations, particularly those involving estimates based on sparse samples of the set of observed events (Diggle and Cox 1983).

For further insight into the properties of homogeneous Poisson processes, consider an equivalent definition listed in Cressie (1993, p. 634):

(a) the numbers of events in non-overlapping regions are statistically independent,
(b) for any region $A \subseteq D$,
\[
\lim_{|A| \to 0} \frac{\Pr[\text{exactly one event in } A]}{|A|} = \lambda > 0
\]
where $|A|$ = the area of region $A$, $D$ is the domain of interest (study area), and

(c)
\[
\lim_{|A| \to 0} \frac{\Pr[\text{two or more events in } A]}{|A|} = 0.
\]

Component (a) is particularly important to the analysis of regional counts, e.g., the number of events observed in a partition of the study area into small areas. Diggle (1983, p. 50) formally establishes the link between criteria (i) and (ii) and component (a) above. Component (b) implies the probability of a single event in an increasingly small area $A$ (adjusted for the area of $A$) is a constant ($\lambda$) independent of the location of region $A$ within the study area of interest. Component (c) implies the probability of two or more events occurring in precisely the same location is zero. As above, the quantity $\lambda$ is the Poisson parameter, or the intensity of the process, and is equal to the mean number of points per unit area.

Since the intensity of events is constant at all locations in the study area, we say the process is homogeneous. Mathematically, stationarity and homogeneity are related but separate concepts (e.g., a process defined within a finite study area cannot be stationary since translation outside of the study area is undefined).

While the homogeneous Poisson process provides a mathematical model for the lack of spatial pattern, various extensions to the Poisson process provide additional structure for modelling spatial pattern. We limit attention to a set of processes maintaining independence between event locations, i.e., the occurrence of an event at a particular location does not impact the probability of other events occurring nearby. Interested readers are directed to Stoyan et al. (1995), Diggle (2003), Geyer (XXXX), and Møller and Waagepetersen (2004) for models of processes with interevent dependence.

### 5.1.2 Heterogeneous Poisson processes

Our first generalization is to allow the intensity function to be a spatially varying function $\lambda(x)$, yielding a heterogeneous Poisson process defined by...
(i*) The number of events occurring within a finite region $A$ is a random variable following a Poisson distribution with mean $\int_A \lambda(x)dx$, and

(ii*) Given the total number of events, $n$, occurring within an area $A$, the $n$ events represent an independent random sample of $n$ locations, with the probability of sampling a particular point $s$ proportional to $\lambda(x)$.

Note that the number of events observed in disjoint regions still follow independent Poisson distributions but now with expectation $\int_A \lambda(x)dx$. In addition, events are distributed according to a spatial density function proportional to $\lambda(x)$. In a heterogeneous Poisson process observed events may “cluster” together in areas with higher values of $\lambda(x)$, but, as noted above, the events remain mutually independent of one another.

Linking this definition back to the ideas of stationarity and isotropy defined for geostatistical data, note that heterogeneity (as defined by a spatially varying intensity function) necessarily implies non-stationarity as the point process is no longer translation invariant (stationarity). Since we define isotropy with respect to rotation invariance around a single point (typically referred to as the origin), heterogeneity in the intensity function results in an anisotropic process only if $\lambda(x)$ itself is anisotropic (i.e., not symmetric around the origin).

5.1.3 Cox processes

To further generalize our model, consider a Cox process (Cox 1955) where the (heterogeneous) intensity function itself represents a realization of a random process, $\Lambda(x)$. That is, suppose we wish to generate a realization from a Cox process. First, we generate the number $n$ of events from a Poisson distribution, and generate a particular intensity surface $\lambda(x)$ from the process $\Lambda(x)$ (ignoring for the moment the precise algorithm for doing so). Next, conditional on $n$ and $\lambda(x)$, we generate $n$ locations from the spatial probability density proportional to $\lambda(x)$.

To further crystalize the various types of processes considered, generating repeated realizations with $n$ and $\lambda(x)$ fixed corresponds to a binomial process (Stoyan et al. 1995), generating repeated realizations with $\lambda(x)$ fixed and $n$ drawn from a Poisson distribution
corresponds to a heterogeneous Poisson process, and generating repeated realizations drawing
\( \lambda(x) \) from \( \Lambda(x) \) and \( n \) from a Poisson distribution each time corresponds to a Cox process.

In a sense, a Cox process represents a hierarchical probability model with a spatially
defined prior distribution on the intensity function. However, unless \( \Lambda(x) \) is defined para-
metrically, the problem is a nonparametric one.

5.1.4 Log Gaussian Cox processes

Next, suppose we limit attention to the set of log Gaussian Cox processes, i.e., Cox processes
where the intensity surface process is defined by \( \Lambda(X) = \exp(Z(x)) \) where \( Z(x) \) denotes
a Gaussian random field with mean \( E(Z(x)) = \mu(x) \) and covariance function \( K(x_1, x_2) \)
perhaps parameterized as in the geostatistical models above, and with some smoothness
restrictions defined in Møller et al. (1998). \( Z(x) \) need not be stationary (i.e., have a station-
ary covariance function), but stationarity is often convenient assumption. The family of log
Gaussian Cox processes places a by-now familiar Gaussian random field model within the
hierarchical structure of Cox processes, linking some ideas from preceding sections to our
point process models.

5.2 Inferential issues

The models outlined above build increasing layers of complexity and variation onto the
basic stochastic allocation of events to locations. Questions of interest include estimation
of the intensity surface \( \lambda(x) \) giving rise to the observed data in order to identify areas with
higher or lower probabilities of event locations for the observed realization, and description
of the intensity generating process \( \Lambda(x) \). Approaches may be non-parametric or parametric
depending on the sorts of models considered. We may be interested in inserting a parametric
structure in \( \lambda(x) \), for instance inserting a parametric distance decay component around a
source of environmental pollution. An estimate of the rate of decay aids in assessments
of putative increased incidence of adverse health events in residents living nearby (Diggle
1990). The (latent) Gaussian random field driving intensity estimation in log Gaussian Cox
processes suggests assessment of parameters in the mean function \( \mu(x) \) and the covariance
function \( K(\mathbf{x}_1, \mathbf{x}_2) \).

We review three approaches for predicting \( \lambda(\mathbf{x}) \) based on the observed event locations \((\mathbf{x}_1, \ldots, \mathbf{x}_n)\), based on the structure of heterogeneous Poisson and Cox processes. In all cases considered here, we build inference (at least conceptually) from the hierarchical nature of a Cox process by treating the unknown intensity surface as a random quantity and defining associated prior distributions and likelihoods, yielding posterior inference for \( \lambda(\mathbf{x}) \).

To begin, we consider an approach due to Heikkinen and Arjas (1998) merits attention not only due to the approach in general, but also due to interesting conceptual issues involving posterior simulation and inference. Heikkinen and Arjas (1998) build on earlier work by Arjas and Gasbarra (1994) regarding Bayesian estimation of the intensity (hazard rate) of temporal Poisson processes. The basic idea is to build a prior probability model on the space of random step functions defining distributions for the number, locations, and sizes of steps. Coupled with the likelihood based on the observed locations, we aim to generate such step functions from the associated posterior distribution of the intensity function. Arjas and Gasbarra (1994) and Arjas and Andreev (1996) note that each sample from the posterior represents a piecewise constant function so that any assumed “true” smooth intensity function is not a member of the class of intensities generated by the approach, rather the posterior reflects a family of approximations. More specifically, Arjas and Andreev (1996) note the pointwise posterior mean intensity need not belong to the class of sampled posterior intensity functions, but pointwise averages across sampled values typically yield smooth surfaces. We return to this issue after outlining the Heikkinen and Arjas (1998) formulation for spatial intensities.

Suppose the data represent a realization from a heterogeneous Poisson process with unknown intensity function \( \lambda(\mathbf{x}) \) defined in a study window \( D \subset \mathbb{R}^2 \). Our goal is posterior inference on \( \lambda(\mathbf{x}) \). Based on the properties of a heterogeneous Poisson process, the likelihood of an observed point pattern \((\mathbf{x}_1, \ldots, \mathbf{x}_n)\) is

\[
p(\mathbf{x}|\lambda(\cdot)) \propto \exp \left\{ -\int_D \lambda(\mathbf{x})d\mathbf{x} \right\} \prod_{i=1}^n \lambda(\mathbf{x}_i)
\]

where integration is with respect to the Lebesgue measure on \( \mathbb{R}^2 \).

We need to define a prior distribution for \( \lambda(\mathbf{x}) \). Heikkinen and Arjas (1998) define a prior
distribution on the space of piecewise constant surfaces defined by creating a tessellation (tiling) of \( D \) and defining a constant intensity within each tile. This is accomplished in three steps. First, we define a set of \( K \) generating points \( (\xi_1, \ldots, \xi_K) \) in an area \( E \supseteq D \). We allow generating points outside of \( D \) to reduce the impact of edge effects in the estimation of \( \lambda(x) \), and to allow posterior prediction of \( \lambda(x) \) for points outside of \( D \) (noting that as one moves away from \( D \), the data will have relatively less and the prior relatively more influence on inference). Second, we define the Voronoi tessellation (also known as the set of Theissen polygons), representing a partitioning of \( D \) into a tile (region) \( E_k \) for each generating point \( \xi_k \) representing the locus of all locations nearer to \( \xi_k \) than to any other generating point. Third, we define the constant intensity \( \lambda_k \) associated with each Voronoi tile. Taken together, the three steps yield a probability model defining a class of piecewise-constant intensity functions

\[
\lambda(x) = \sum_{k=1}^{K} \lambda_k 1\{x \in E_k\}
\]

where \( 1\{\cdot\} \) represents the indicator function. The choice of the Voronoi tessellation provides a computational advantage, since updating the tessellation to reflect the addition or deletion of a single generating point can be accomplished very efficiently (Preparata and Shamos 1985, pp. 204-225, Edelsbrunner 1987, Chapter 13, Okabe et al. 2000, Chapter 4).

Defining the prior distribution for such piecewise-constant intensities requires defining the number and distribution of the set of generating points and the distribution of intensity values for each tile. This is accomplished in a hierarchical fashion by defining the set of generating points as the realization of a homogeneous Poisson process with constant intensity \( \lambda_\xi \in (0, \infty) \) and zero probability assigned to the empty pattern. The properties of a Poisson process imply that the number \( K \) of generating points follows a Poisson distribution, and, given \( K \), the generating points are uniformly distributed across \( E \) so \( p(\xi) \propto \lambda^K \) for \( K > 0 \), \( p(\xi) = 0 \) otherwise.

Heikkinen and Arjas (1998) define the prior distribution of \( \lambda = (\lambda_1, \ldots, \lambda_K) \) conditional on the set of generating points \( \xi \), by defining a spatial Gaussian process for \( \eta = \log(\lambda) \). The spatial Gaussian process (as illustrated above for both regional and geostatistical data) provides a mechanism for incorporating spatial correlations, in this case due to the assumed
smoothness of the underlying intensity function. More specifically, Heikkinen and Arjas (1998) utilize a conditional autoregressive (CAR) prior for $\eta$ with neighbors defined by the Delauney triangulation of generating points (the graph connecting any pair of generating points whose associated Voronoi tiles share a common point or boundary). As a Delauney triangulation, the set of neighbors represents the “dual graph” to the Voronoi tessellation and is again computationally straightforward to update following the addition/deletion of a single generating point. Heikkinen and Arjas (1998) define a more general set of weights than the adjacency weights defined in the discussion of regional models above, but nonetheless still within the class of weights meeting the symmetry requirements defined in equation (5).

This procedure yields a joint prior distribution

$$p(\xi, \eta) = p(\eta|\xi)p(\xi)$$

with two hyperpriors to be defined, namely the constant intensity $\lambda_\xi$ in $p(\xi)$ and an overall scale parameter in the conditional autoregressive prior $p(\eta|\xi)$. Both hyperpriors influence the overall smoothness of the resulting posterior samples, with $\lambda_\xi$ influencing the (expected) number of generating points (more points tending to give “lumpier” surfaces), and the scale parameter (analogous to $\tau_{CAR}$ in equation (2)) the conditional variance of $\eta_k$ relative to its neighboring values.

Assembling the components, the joint posterior becomes

$$p(\xi, \eta|x) \propto p(x|\xi, \eta)p(\eta|\xi)p(\xi)$$

where, for the piecewise-constant intensities under consideration, the likelihood becomes

$$p(x|\xi, \eta) = \exp \left[ \sum_{k=1}^{K} \{ N(E_k)\eta_k - \text{area}(E_k \cap D) \exp(\eta_k) \} \right]$$

where $N(E_k)$ represents the number of observed events in $x$ occurring within tile $E_k$ and $\text{area}(E_k \cap D)$ represents the Lebesgue measure of the area of the intersection of tile $E_k$ with the study area $D$.

To simulate samples from the posterior requires reversible jump Markov chain Monte Carlo iterations where each update requires either:
• a change in $\lambda_k$ for a tile $k$.

• adding a new generating point and associated tile, or

• deleting a generating point and adjusting the neighboring tiles appropriately to maintain a Voronoi tessellation of the remaining generating points.

The probabilities of which action to take depends on the current number of tiles, and details in Heikkinen and Arjas (1998) reveal each update only requires local calculations due to the use of the Voronoi tessellation.

As mentioned above with respect to Arjas and Gasbarra’s (1994) temporal intensity estimates, the set of posterior intensity functions is comprised (by definition) entirely of piecewise-constant surfaces, analogous to generating samples from a posterior distribution of histograms to approximate a density function (Arjas and Andreev 1995). While no single sampled value represents a smooth surface, the pointwise average of $M$ postconvergence samples, $\lambda^{(1)}, \ldots, \lambda^{(M)}$,

$$\hat{\lambda}(x) = \frac{1}{M} \sum_{m=1}^{M} \lambda^{(m)}(x)$$

does tend to yield a smooth surface approximating the posterior “mean” surface, even though such a surface is not a member of the class of surfaces generated by the posterior.

The end result of Heikkinen and Arjas’ (1998) approach is an algorithm for posterior inference for intensity functions based on a series of clever approximations, each of which has strong computational advantages (e.g., the Voronoi tessellation, Delauney neighborhoods, and the use of the CAR prior for $\eta$). Several examples involving both observed and simulated examples appear in Heikkinen and Arjas (1998, 1999) and illustrate the flexibility of the approach to accurately detect general modes, trends, and sharp transitions. More recently, Heikkinen (2003) provides updated suggestions for such models, discussing in particular alternatives to the CAR prior to streamline implementation.

Heikkinen and Arjas’s (1998) use of a CAR prior for $\eta = \log(\lambda)$ suggests a log-Gaussian Cox process structure. Independently, but almost simultaneously to Heikkinen and Arjas (1998), Møller et al. (1998) considered Bayesian prediction of $\lambda(x)$ in their definition of
log Gaussian Cox processes, using a joint formulation of the Gaussian process defined for a discrete grid of locations within the study area $D$. Møller et al.'s (1998) implementation utilizes a grid of locations spanning $D$ and provides prediction of $\lambda(x)$ at each grid point via Metropolis-adjusted Langevin algorithm (Besag 1994, Grenander and Miller 1994, Roberts and Tweedie 1997).

The probabilistic hierarchy mirrors that of Heikkinen and Arjas (1998), beginning with the likelihood (defined more explicitly here in terms of the latent Gaussian process)

$$p(x|Z) \propto \exp\left(-\int_D \exp(Z(x))dx\right) \prod_{i=1}^n \exp(Z(x_i)).$$

In general, the posterior distribution of $\lambda(x)$ is defined by transformation of the posterior distribution of $Z(s)$ given by

$$p(Z(x) \propto p(x|Z)p(Z)$$

where $p(Z)$ is defined as a prior over the family of Gaussian random fields. Heikkinen and Arjas (1998) operationalize this prior (hence posterior) via the definition of generating points and discretizing the study area by Voronoi tessellation.

In contrast, Møller et al. (1998) operationalize the prior (hence posterior) via the posterior distribution of $\lambda(s_{0,1}), \ldots, \lambda(s_{0,K})$ for a set of locations $s_{0,k}, k = 1, \ldots, K$. We note that locations $s_{0,k}, k = 1, \ldots, K$ are fixed, user-defined prediction locations (typically a set of grid points) not random realizations from a point process so we revert to geostatistical notation for prediction locations $s_0$. However, we also note the prediction locations also play a role similar to Heikkinen and Arjas’ (1998) generating points $\{\xi_k\}$ in that they support predictions of $\lambda(x)$, so we again index by $k = 1, \ldots, K$, although here the total number of prediction locations $K$ remains fixed throughout.

Waagepeterson (2004) summarizes the approach of Møller et al. (1998) by noting the posterior distribution of $(Z(s_{0,1}), \ldots, Z(s_{0,K}))$, given the observed locations $x_1, \ldots, x_n$ (and assuming $K \geq 1$) can be expressed as

$$p(Z(s_{0,1}), \ldots, Z(s_{0,K})|x) \propto E[p(x|Z(s_{0,1}), \ldots, Z(s_{0,K}))p(Z(s_{0,1}), \ldots, Z(s_{0,K})).$$

(14)
The posterior of the entire process $Z|\mathbf{x}$ is given by the set of finite-dimensional posterior distributions in equation (14) for $K \geq 1$ and all sets of grid points $s_{0,1}, \ldots, s_{0,K}$ in $D$. The details of the MCMC approach are beyond the scope of this article, but are discussed in Møller et al. (1998), Møller and Waagepetersen (2004), and Waagepetersen (2004).

As in the approach of Heikkenin and Arjas (1998), the posterior defined in (14) is analytically intractable and Monte Carlo approaches are not possible without discretizing $Z$. Møller et al. (1998) approximate $Z$ via a piecewise constant surface defined at each grid point. Waagepetersen (2004) shows that such an approximation converges to the true continuous process $Z$ as the number of grid points increases and grid spacing decreases.

A third, independent and simultaneously developed approach to modelling spatial intensities is given by Wolpert and Ickstadt (1998), whose motivation included a desire to develop a scalable hierarchical model for regional counts, i.e., a model whose spatial correlation structure is preserved under aggregation and disaggregation of regions, unlike the CAR models defined for regional data above. As with Heikkenin and Arjas (1998) and Møller et al. (1998), Wolpert and Ickstadt (1998) consider a hierarchical Cox process structure. To begin, the number of events observed in the study area follows a Poisson distribution with mean defined by the intensity function $\lambda(\mathbf{x}), \mathbf{x} \in D$. The intensity, in turn, represents the realization of a generating process $\Lambda(\mathbf{x})$.

Rather than a log Gaussian definition of $\Lambda(\mathbf{x})$, Wolpert and Ickstadt (1998) consider the intensity process to be defined by a weighted sum of kernels, i.e.,

$$\Lambda(\mathbf{x}) = \sum_c \delta_c \omega(\kappa_c, \mathbf{x}),$$

where $\delta_c$ denotes a set of weights, $\omega(\kappa_c, \mathbf{x})$ a kernel function centered at point $\kappa_c$ and evaluated at $\mathbf{x}$, and $\kappa_c, c = 1, \ldots, C$, denote points arising from a separate spatial point process. As a result, the observed points $\mathbf{x}$ represent the superposition of events generated by $C$ independent heterogeneous Poisson processes with intensities defined by the kernels $\omega(\kappa_c, \mathbf{x})$ centered at points $\kappa_c, c = 1, \ldots, C$.

Conceptually, it is helpful to think of such a process as a cluster process (Neyman 1939, Neyman and Scott 1958) wherein parent events $\kappa_c, c = 1, \ldots, C$, are scattered across $D$, 

37
and the \( c \)th parent generates a random number of \textit{offspring} events which are distributed according to an intensity \( \omega(\kappa_c, \mathbf{x}) \) around the location of the \( c \)th parent, \( \kappa_c \). The observed process consists only of the offspring points, \( \mathbf{x} \). More recently, Møller (2003) considered the broader class of \textit{shot noise Cox processes} (think of parent locations as the center of a shotgun blast and each offspring location as a point of impact of a particular pellet), which include a wide variety of cluster processes, including that proposed by Wolpert and Ickstadt (1998). While applications of such (cluster or shot noise) processes often involve assessment of spatial interactions between events from the same (unobserved) parent location, without additional information, the patterns of (correlated) locations from a cluster process are analytically indistinguishable from patterns of (independent) locations arising under a general Cox process with appropriately constructed intensity generating process, a relationship first formally derived by Bartlett (1964). In other words, we cannot determine from the observed pattern alone whether event locations are uniformly distributed in space but correlated with one another, are independent and heterogeneously distributed in space, or some combination of the two. However, even though the Wolpert and Ickstadt (1998) process can be viewed as a cluster process, a Bayesian analysis still allows posterior prediction of an underlying intensity surface, our underlying inferential goal.

Wolpert and Ickstadt’s (1998) proposed prior for \( \Lambda(\mathbf{x}) \) represents a gamma random field in contrast to the Gaussian random field utilized in the other approaches. Some advantages of the gamma field prior include: conjugacy of the random field, the desired “scalability”, and a convenient way to model uncertainty in location as a discrete sum of infinitely many point masses. For implementation, Wolpert and Ickstadt (1998) detail construction of a general MCMC approach with Metropolis steps based on an inverse Lévy measure.

To review, the primary difference between the three approaches outlined is in their respective definitions of a prior distribution for the unknown process \( \Lambda(\mathbf{x}) \) generating the intensity function \( \lambda(\mathbf{x}) \), and the resulting MCMC implementation. The family of log Gaussian Cox processes has many nice mathematical properties (Møller et al. 1998, and Møller and Waagepetersen 2004), and fits most seamlessly with the geostatistical models outlined above.
Møller and Waagepetersen (2002), Richardson (2003), and Møller (2003) provide additional discussion comparing approaches, noting the amount and type of prior knowledge coupled with the user’s analytic goals provide the primary motivations for choosing one approach over another. In particular, log Gaussian Cox processes offer an attractive structure for modelling mutually independent events whose locations result from a smoothly varying intensity function (e.g., due to smoothly varying environmental factors). The embedded Gaussian random field provides a direct link between log Gaussian Cox process models and the hierarchical prediction models described briefly above and outlined extensively in Banerjee et al. (2004). Shot noise Cox processes provide additional flexibility and seem most natural for modelling clustered phenomena. In addition, the weighted kernel mixture defining the intensity process for shot noise Cox processes bears similarity to Bayesian nonparametric models such as those incorporating Dirichlet process priors for density estimation (Escobar and West 1998). Even though inference for density and intensity functions are (proportionally) equivalent, there has been relatively little overlap to date between the Bayesian point process literature and the Bayesian density estimation literature.

The discussion above focused on inference for the intensity function as the most straightforward connection to models considered in preceding sections. For example, a log Gaussian Cox process differs from hierarchical geostatistical models primarily through the addition of an additional (lower) level to the hierarchy reflecting the Poisson process generating event locations. As in other arenas, the advantage of a Bayesian formulation is fully model-based inference allowing assessment of additional features of the model, whether that be counts of events in regions or estimates of parameters driving spatial heterogeneities or correlations.

6 Recent developments and Future directions

Our review of Bayesian thinking in spatial statistics only touches on a series of common themes across major inferential issues in spatial statistics. The central component to all approaches (indeed for all Bayesian analyses) is the development and application of a fully-defined probability model from which posterior inference flows. As noted by Box and Tiao
(1973, pp. 9-10), the appropriateness of the model may (should) always be challenged, but fully Bayesian (posterior-based) inference focuses criticism on the appropriateness of the probabilistic structure of the model, rather than the mode of inference, accuracy of approximation, or appropriateness of the particular choice of estimator.

While lengthy, the above review is by no means comprehensive. There are many interesting branches from the main path and we conclude our tour by highlighting related areas and current developments relating to Bayesian analysis of spatially-referenced data.


Another class of extensions to the family of CAR priors addresses multivariate CAR (MCAR) models allowing spatial similarity between and across multiple random quantities measured over the same study area (e.g., allowing correlation between levels of particulate matter composition with neighboring values of ozone levels). Relevant work includes Kim et al. (2001) and Carlin and Banerjee (2003).

A variety of additional applications utilize CAR priors for parameters governing local deviations in survival rates and covariate effects. The former typically assigns spatial prior distributions to frailty models (Banerjee and Carlin 2002, 2003). The latter seeks to provide model-based inference wherein covariate effects may vary from location to location. The growing popularity of *geographically weighted regression* (GWR, Fotheringham, Brunsdon, and Charlton 2002) in the social sciences illustrates the wide audience for such analyses but current implementations provide descriptive estimates and fall somewhat short of a comprehensive framework for inference. Bayesian approaches appear in Schootman and Sun (2004) who explore a specific regional data example, and in Banerjee et al. (2004, pp.
Advances in the Bayesian analysis of geostatistical data include a thorough discussion of model-based geostatistics in Diggle et al. (1998) and the methods and models explored in detail in Banerjee et al. (2004). Particular examples include Ecker and Gelfand’s (1997) Bayesian inference for semivariograms as a preface to their treatment of spatial prediction under anisotropy (Ecker and Gelfand 1999, 2003); and inference for spatial gradients in the predicted surface (Banerjee et al., 2003).

The preceding section illustrated the use of hierarchical models for inference regarding the intensity of heterogeneous Poisson processes in the Cox process setting. Other applications include inference regarding pairwise interaction parameters for correlated point point processes (Bognar and Cowles 2004, Bognar 2004). Unlike many MCMC-based Bayesian analyses, the models in Bognar and Cowles (2004) involve algorithms complicated by a variable proportionality constant, lessening one of the convenient advantages usually expected in MCMC approaches. Recent developments in the analysis of spatial point processes also include spatio-temporal formulations of both log Gaussian Cox processes (Brix and Diggle, 2001) and shot noise Cox processes (Brix and Chadoeuf 2002). On a more theoretical note, Schoenberg (1999) provides a means to transform general spatial point processes into inhomogeneous Poisson processes, a result that may allow application of the Cox process approaches above to a much wider class of spatial point process models.

Finally, there is also a growing literature of Bayesian approaches to address spatial classification problems. Knorr-Held and Rasser (2000), Knorr-Held (2003), and Denison and Holmes (2001) both consider such problems with regional data, classifying regions into spatially contiguous categories to identify similar summary values. Byers and Raftery (1998, 2002) address a similar problem with point data using Voronoi tiles to estimate the underlying intensity function (in a manner very similar to the approach of Heikkinen and Arjas 1998) in order to classify the study area into areas of high and low intensity.

In conclusion, we find great appeal for the application of Bayesian approaches to the analysis of spatial data, particularly through the use of hierarchical structures that play
important roles in the analysis of regional, geostatistical, and point process data. We hope
the discourse above provides a thorough (although by no means complete!) introduction to
Bayesian ideas in spatial statistical analysis and gives readers multiple resources for further
study.
References


43


