The Bayesian Modeling of Disease Risk in Relation to a Point Source

Jon C. WAKEFIELD and Sara E. MORRIS

Recently there has been increased interest, from both the media and the public, in the question, "Is there an excess of disease risk close to a prespecified point source?" To address this question, routinely available public health data may be analyzed. In the United Kingdom, as in many countries, health data and the associated population data that are required for comparison, are available as aggregated counts. In this article we propose to analyze such data using a Bayesian disease mapping framework. This framework allows the extra-Poisson variability that is frequently encountered to be accommodated through random effects that may be unstructured or display spatial dependence. The disease risk-spatial location relationship is modeled using a simple but realistic parametric form. The random effects may be used for diagnostic purposes, in particular to assess the appropriateness of the distance-risk model. The choice of prior distribution is extremely important in this context and we develop an informative prior distribution that is based on epidemiological considerations and on additional analyses of data that are obtained from a larger "reference" region within which the study region is embedded. We argue that a particularly useful inferential summary for public health purposes is the predictive distribution. For example, we may obtain the distribution of the number of cases that would be expected to occur within a specified distance of the putative source (given a population size, by age and sex, and a time period). The approach is illustrated using data from an investigation into the incidence of stomach cancer close to a municipal solid waste incinerator. The sensitivity to the prior distribution and the presence or absence of spatial random effects is examined. To determine whether the increase in risk detected in the study is persistent, we analyze incidence data from the four-year interval following the study period. We finally describe a number of extensions including the modeling of data from a number of sites using a four-stage hierarchical model. This model is statistically realistic and, more importantly, allows the epidemiological question to be answered with greater reliability.

KEY WORDS: Disease mapping; Environmental epidemiology; Overdispersion; Predictive distributions; Prior distributions; Public health; Spatial epidemiology.

1. INTRODUCTION

There is now great interest, from the public and the media alike, in environmental health issues and, in particular, in the possible effects of industrial pollution on health. If environmental factors are playing a part in the etiology of a disease, then we would expect "clustering" of disease cases (relative to the population at risk) in areas where and at times when (possibly after a suitable lag) the environmental factor is present. The following two specific questions are of great interest.

1. Does a disease exhibit spatiotemporal clustering?
2. Is there evidence that a particular spatiotemporal configuration of cases is a cluster?

These two questions are related. For example, knowledge that a disease, in general, does or does not exhibit clustering aids in the analysis and informed interpretation of an alleged cluster. The first question is mathematically well defined (Alexander and Cuzick 1992) and is inherently easier to answer than the second. In this article, we are concerned with the situation in which the second question is of the form, "Is there evidence of increased risk close to a putative source?" Besag and Newell (1991) refered to this as a "focused" question while evidence for clustering is a "general" question.

Boyle, Walker, and Alexander (1996) provided an interesting account of investigations of leukemia clusters and clearly illustrated the difficulties of interpretation associated with a single cluster. In particular, the manner by which the investigation came to light is of vital importance for an interpretation to be placed on a statistical significance level. For example, the analysis may have arisen as a result of an a priori epidemiological hypothesis concerning the etiology of the disease, in which case interpretation follows in the usual manner. If, however, the analysis arises in response to the public–media highlighting of a particular source, then the p values associated with conventional tests are not appropriate because the same data are being used to both generate and test the hypothesis.

In the situation examined in this article, a specific point source is suspected to be responsible for the cluster. In this case the hypothesis may be tested more satisfactorily if data are available from time periods before and after the point source became operational (allowing for a suitable lag period). Clearly the existence of data from more than a single site also provides far greater evidence for the existence of a relationship between a specific environmental hazard and disease risk. Beyond the interpretative problems that lead from the selection mechanism, there are further difficulties related to data quality and confounding (Elliott, Martuzzi, and Shaddick 1995; Wakefield and Elliott 1999). We consider these issues in more detail in later sections.

The generic situation we consider is that in which we have aggregated health and population data, that is, we do not have the exact locations of individuals with and without the disease. Case-control data in which exact locations are available are preferable, but are more expensive to collect (aggregated data are often routinely available), and suffer from potential difficulties of selection bias. The regions that define the level

Jon Wakefield is Associate Professor, Departments of Statistics and Biostatistics, University of Washington, Seattle, WA 98195 (E-mail: *jon@stat.washington.edu*). Sara Morris is Research Associate, Small Area Health Statistics Unit (SAHSU), Department of Epidemiology and Public Health, Imperial College School of Science, Technology and Medicine, St. Mary's Hospital, London, W2 1PG, United Kingdom (E-mail: *sem30@ic.ac.uk*). SAHSU is sponsored by the Departments of Health and the Environment and the Health and Safety Executive. This work was supported, in part, by an equipment grant from the Wellcome Trust (0455051/Z/95/Z). The authors thank Professor Paul Elliott, Director of the Small Area Health Statistics Unit, for helpful suggestions, and the editor, associate editor, and three referees for detailed comments that greatly improved the article.

of aggregation are generally defined for administrative purposes and so, in terms of the exposure, are arbitrary. In general the aggregation level of cases and populations at risk will not be the same; Best, Ickstadt, and Wolpert (1999) and Mugglin, Carlin, and Gelfand (1999) considered methods for dealing with this problem.

Additional information may include area-specific confounders (covariates). For example, in aggregate data studies in which individual-level data are unavailable, socioeconomic status (SES) has been shown to be a powerful predictor of disease risk (Jolley, Jarman, and Elliott 1992; Kleinschmidt, Hills, and Elliott 1995). An important and difficult issue is the interpretation of such an observation. Clearly for many diseases, an area-level measure of SES may act as a surrogate for known risk factors such as diet, alcohol consumption, and smoking status of the individuals of the area, but the association may have a component that is truly area level (for example, access to health services) or is related to the experiences of those in the area, beyond the SES of the individual (to give a contextual variable).

We note that, in general, when group-level data are considered, there is always the possibility of "ecological bias" because individual-level relationships coincide only with those at the group level under strict circumstances (the groups are areas in our context). Relevant issues include the mathematical form of the risk–exposure model, the presence of within- and between-group confounders, mutual standardization, and effect modification (e.g., Greenland and Morgenstern 1989; Greenland 1992; Richardson 1992; Greenland and Robins 1994).

In this article, we propose a Bayesian hierarchical model for analysis, utilizing a simple but realistic distance–risk function (as advocated by Diggle, Elliott, Morris, and Shaddick 1997). We argue that an important aid in the assessment of the public health implications of a point source is the predictive distribution for the number of cases, a natural summary in a Bayesian approach. The hierarchy allows the explicit modeling of overdispersion (extra-Poisson variability) in terms of random effects that may display spatial dependence. The existence of overdispersion frequently has been reported in disease mapping studies (e.g., Mollié and Richardson 1991) and may be due to data anomalies or unmeasured risk factors that may or may not display spatial structure. Examples of such anomalies in the numerator include double counting and underregistration of cases; in the denominator, examples include underenumeration and migration.

To illustrate our framework, we present a case study in which the association between stomach cancer incidence and the distance from a municipal incinerator in the northeast of England is investigated. Prior distributions are derived from previous related studies and from an analysis of data from a reference region within which the study region is contained. When spatially dependent random effects are included in the model, there is the possibility of confounding between exposure and risk factors accounted for by these spatial random effects. The use of informative prior distributions for the hyperparameters of the random effects distribution allows this possibility to be investigated by "fixing" the level of spatial dependence.

The structure of this article is as follows. In the next section, the study that we consider is introduced. Section 3 contains a discussion of previous approaches and Section 4 contains our Bayesian approach, with subsections that consider the specific model utilized, predictive distributions, and computation. The choice of prior distribution is extremely important and we devote Section 5 to this topic. In Section 6, we present our analysis of the stomach cancer data. The final section contains a concluding discussion, including possibilities for future work.

2. CASE STUDY—STOMACH CANCER AROUND MUNICIPAL SOLID WASTE INCINERATORS

Elliott et al. (1996) investigated whether proximity to municipal incinerators was associated with an increased risk of cancer by analyzing incidence data in the vicinity of all 72 such incinerators in Great Britain for the period 1974–1986. Pollutants emitted from municipal waste incinerators include heavy metals (especially lead, cadmium, and mercury), acidic gases, organic compounds, and partially combusted organic materials. Some of these substances have been classified as likely or possible human carcinogens. We illustrate how the relationship between the incidence of stomach cancer and a putative source of pollution can be investigated at one particular point source. The incinerator that we select was operable for the period 1940–1976. Confidentiality does not allow disclosure of the exact location, but the site is a coastal town in the northeast of England. We chose the site because there was evidence of increased risk in the vicinity of this incinerator. As outlined in Section 1, interpretation of the analysis is not straightforward. However, evidence of an increase in risk does provide a more interesting context within which issues including the choice of prior distribution and the inclusion of random effects, may be examined. In Section 7 we obtain data for the period 1987–1991 so that we can address the substantive question of the possibility of increased risk in the vicinity of the incinerator under study.

The study area was chosen to be a circular region of radius 7.5 km, centered on the incinerator. The size of this region was chosen to reflect the extent of the potential effect of emissions. Within this region, the number of cases of stomach cancer was determined; each case has an associated postcode. A postcode contains, on average, 14 households. In the full study reported in Elliott et al. (1996), a 10-year lag from the time each site became operable was used to allow for the development of disease (see Rothman and Greenland 1998 for a discussion of latency periods). Estimates of the population at risk were obtained from the 1981 decennial census. This provides, at the time of the census, the number of individuals living in particular census-defined enumeration districts (EDs) by sex and 18 five-year age bands. An ED contains, on average, 400 individuals. The study region contains 44 such EDs. Expected numbers, adjusting for the known risk factors age and sex (Nomura 1997) were then calculated based on national stomach cancer rates over the period of study. Note that EDs are larger than postcodes and so we aggregate the case data to EDs, and this is the ecological level of the analysis. Figure 1(a) shows the study region with the ED centroids indicated; the radii of the circles are proportional to the expected numbers of cases.

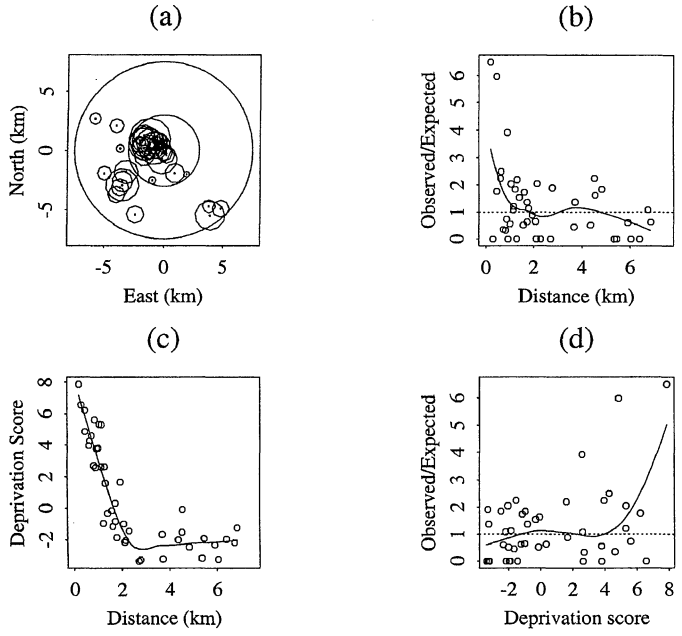

Figure 1. Study Region Characteristics. (a) Positions of enumeration district (ED) centroids in relation to the incinerator (represented by the origin), The large concentric circle represents the extent of the study region and the smaller circle has radius 3 km. The additional circles are centered on the ED centroids and have radii proportional to the expected number of cases. (b) Standardized incidence ratios plotted versus distance. (c) An index of socioeconomic status plotted versus distance. (d) Standardized incidence ratios plotted versus the index of socioeconomic status. The solid lines on (b)–(d) denote lowess-smoothers.

Table 1 summarizes the population, health, and exposure data for the site. The standardized incidence ratio (SIR) is given by the ratio of the observed to expected counts and is the maximum likelihood estimate (MLE) of the relative risk under a Poisson model. Figure 1(b) shows the SIRs plotted versus distance with a local smoother imposed. From this plot, there appears to be an association: EDs closer to the incinerator are subject to increased risk.

As noted in the previous section, SES can be a powerful predictor of disease. It is well documented that social class is a powerful risk factor for stomach cancer (see Nomura 1997 and references therein). We utilize an area-level measure of SES derived from three census variables, namely the proportion unemployed, the proportion of overcrowded households, and the average social class of household, in each ED. Each of these scores was then standardized to have zero mean and unit standard deviation across Great Britain and the sum was taken to give the SES score. The variable is, therefore, continuous, with large values indicating high deprivation and low values indicating affluence. Figure 1(c) plots the SES index versus distance and illustrates that areas surrounding the incinerator are relatively deprived. Figure 1(d) illustrates that, as expected, high deprivation is associated with high risk.

The link between SES, health, and exposure to pollutants is complex. In our analysis we are using the obvious causal interpretation that ill health may be caused by both SES and exposure. We acknowledge, however, that each of ill health and environmental pollution may contribute toward higher levels of deprivation. Here we assume that deprivation is not on the causal pathway between pollution and health, that is, we treat deprivation as a confounder. To calculate expected numbers that adjust for deprivation, we determine disease rates across levels of deprivation within a reference area. Whereas deprived individuals are more likely to live in areas of high pollution, it is possible that these rates reflect pollution levels also and so there exists the possibility that some of the effect of pollution may be lost in a particular study. For specific disease-exposure relationships and rare exposures, this becomes less of a problem, however (Dolk et al. 1995).

3. MODELING DISEASE RISK

We now consider how disease risk may be modeled as a function of spatial location in relation to a point source, and we review and highlight the shortcomings of a number of previously proposed approaches. A more detailed description of the methods, and a more comprehensive analysis of the data considered here using non-Bayesian approaches, can be found in Morris and Wakefield (2000).

We first note that the ideal exposure measure is the cumulative dose (or perhaps the dose weighted by time since exposure) due to the pollution source received by each of the occupants in the study region. This is clearly an unrealistic goal. We also note that a person does not spend 24 hours a day at their residence. In some instances, pollution monitors may be located close to the source and the measurements obtained from these may be combined with modeled emissions to produce concentrations of emissions by location. Such emission modeling may be based on the characteristics of the incinerator (e.g., height of chimney, speed and volume of emissions),

Table 1. Summary Statistics Across Enumeration Districts for 1974–1986

	Study region				Reference region			
Variable	*Total*	*Min*	*Max*	*Mean*	*Total*	*Min*	*Max*	*Mean*
Population	36,824.4	314.9	1674.0	836.9	82,318.0	63.1	824.3	393.9
Expected counts	71.8	0.3	3.9	1.6	353.8	0.2	3.1	1.7
SIRs	—	0	6.3	1.3	—	0	4.4	1.0
Disease counts	85	0	10	1.9	335	0	7	1.6
Socioeconomic status	—	−3.39	7.9	0.7	—	−4.2	7.1	−0.8
Distance	—	0.2	6.9	2.5	—	—	—	—

NOTE: The reference region is used to inform the prior distribution. Populations have been adjusted for underenumeration and so are not necessarily integers. Expected counts for the study region have been adjusted for age and sex. Socioeconomic status is an area-level measure derived from three census variables concerning unemployment, overcrowding, and social class. Distances are between the location of the incinerator and the population centroid of enumeration districts.

the geography (e.g., proximity to hills and valleys), and meteorology (e.g., wind direction, temperature). In the future, due to increased numbers of regulations on emissions and compulsory monitoring, such information may be routinely available, but this is, in general, not true now. It is also worth mentioning that for cancers in particular, study periods are typically greater than 10 years and it is rare to obtain retrospective pollution information over such a long period.

Partly due to these considerations, and shortcomings in the data (i.e., numerator and denominator inaccuracies), many of the statistical developments over the past 10 years have concentrated on semiparametric modeling techniques, which make few assumptions on the form of the risk function in relation to a point source. In the next section we describe a number of these techniques.

3.1 Semiparametric Approaches

A simple method for investigating these kinds of data is the so-called near versus far analysis, where a region within a particular distance from source is chosen to define near. The disease rates in the two regions are then compared. For our data we defined the near region as those EDs that have their centroids within 3 km of the incinerator; this region is indicated on Figure 1(a). Adjusting for age and sex in the expected numbers, and using a one-sided Poisson test for equality of Poisson means, a p value of .055 was obtained. Adjusting additionally for SES within the expected numbers, using quintiles, yielded a p value of .074. Hence the significance is reduced after adjustment for SES, as expected from Figure 1, (c) and (d).

There are a number of obvious drawbacks to this method, in particular, the somewhat arbitrary choice of near and far and the inability to include covariate information except within the expected numbers. The choice of near and far is also likely to be crucial in determining the power of the test.

A popular semiparametric method for investigating excess risk close to a point source is Stone's test (Stone 1988; Bithell and Stone 1990). Let $\lambda_1, \ldots, \lambda_n$ denote the true risks in each of n areas ranked via increasing distance from the point source. The null hypothesis of constant risk may then be compared with the alternative of nonincreasing risk as a function of increasing distance, that is, H_1: $\lambda_1 \geq \cdots \geq \lambda_n$. The method provides a p value but not (directly) a modeled distance–risk function. Bithell (1992) provided an interesting discussion of Stone's test and related methods; see Morris and Wakefield (2000) for recent extensions to Stone's test. For our data, a p value of .001 (obtained from 999 Monte Carlo simulations under H_0) was obtained both without and with adjustment for deprivation. Hence this test strongly suggests that the relationship between risk and distance is nonincreasing.

The foregoing methods suffer from a number of additional difficulties. First, there is a limited ability to incorporate known information concerning the form of the risk–distance, or more generally the risk–location, relationship. A related problem is that it is not possible to obtain the predicted number of cases in areas at different distances from the source and this, as we shall argue in Section 4.2, is a highly informative summary of the analysis from a public health perspective. Finally, no account is taken of the extra-Poisson variation that is commonly encountered in spatial epidemiology.

Kernel methods for modeling disease risk have been considered by Bithell (1990), Kelsall and Diggle (1995a,b), Lawson and Williams (1993), and Anderson and Titterington (1997). These techniques are useful for exploratory analyses, but their more formal use is hindered by the difficulties in choosing a value for the smoothing parameter.

3.2 Parametric Approaches

A Poisson process framework for the analysis of health data in the vicinity of a putative source was described by Diggle (1990) in the context of case-control data, and the extension of these models to aggregated data was provided by Diggle, Elliott, Morris, and Shaddick (1997). Suppose that Y_i, $i = 1, \ldots, n$, denotes the observed number of cases in each of n areas. Further let E_i denote the expected number of cases based on the age–sex profile of the population at risk in area i and let $Z_i = (Z_{i1}, \ldots, Z_{im})^T$ be a vector of m area-specific confounders. Then the Y_i may be assumed to be independent and identically distributed (iid) Poisson random variables with mean $\rho E_i \exp(Z_i^T \phi) f(d_i; \theta)$, where $\phi = (\phi_1, \ldots, \phi_m)^T$, d_i is the distance between the population-weighted centroid of area i and the point source, and ρ is a parameter that relates the overall risk in the study region to that in the reference region from which the expected numbers were calculated. Cook-Mozaffari et al. (1989) analyzed data of this kind using a log-linear function of distance. This model may be fitted with standard software, but unfortunately it produces modeled risks that decline to zero (and not baseline) as distance tends to infinity. Regressing against the reciprocal of distance removes this problem, but results in the risk at the source (which is a useful summary) being undefined. For the latter model, inference may also be very sensitive to the distance between the point source and the nearest area centroid.

The particular function used by Diggle (1990) was

$$f(d_i; \theta) = 1 + \alpha \exp(-\beta d_i^2), \tag{1}$$

where $\theta = (\alpha, \beta)$. The model depends on just two parameters and provides a simple yet plausible form of relationship between risk and distance. It was argued by Diggle et al. (1997) that the quality of the data often will not allow the consideration of more complex forms. There are many advantages of such a parametric approach, an obvious one being that modeled risk functions are produced, along with associated interval estimates. Note that this model ignores directional effects; a two-dimensional location has been summarized in terms of a distance. A more complex anisotropic model that allows for differential risk at different orientations and nonmonotonicity in risk with distance was proposed by Lawson (1993). Such models also may be embedded within the Bayesian framework that we describe in the next section. The disadvantage of a parametric approach is that inference becomes less reliable the further the true location–risk relationship moves from the form assumed. Consequently, it is vital to produce diagnostics that allow the appropriateness of the model to be examined.

Two (nested) models were considered by Diggle et al. (1997). Their "step" model is given by $f(d; \theta) = 1 + \alpha$ for $d \leq \delta$ with $f(d; \theta) = 1$ for $d > \delta$, whereas in

the "full" model, $f(d;\theta) = 1+\alpha$ for $d \le \delta$, and $f(d;\theta) = 1+\alpha\exp[-((d-\delta)/\beta)^2]$ for $d > \delta$. The step model reflects the simple near versus far model of excess disease risk, but estimates the extent of the near–far region. The full model allows a smooth transition between the step model and (1). For the step and full models, $\theta = (\alpha, \delta)$ and $\theta = (\alpha, \beta, \delta)$, respectively. Diggle et al. (1997) allowed for overdispersion via $\text{var}(Y_i) = \kappa \times \text{E}[Y_i]$, for $i = 1, \ldots, n$, with estimation proceeding via quasilikelihood. Hence unstructured extra-Poisson variability is acknowledged, but not spatial dependence between areal counts. To incorporate the latter via quasilikelihood is not straightforward (see, for example, Breslow and Clayton 1993).

In each of the foregoing models, the parameters θ may be estimated by maximum likelihood estimation; this was the approach taken by Diggle (1990) and Diggle et al. (1997). The usual asymptotic arguments do not apply. In particular, the discontinuous nature of the step and full models gives a nonregular likelihood. Consequently, Diggle et al. (1997) based inference on a Monte Carlo approach in which sampling distributions of estimators and test statistics were investigated via simulation of replicate datasets. We prefer a reparameterized version of (1), namely

$$f(d;\theta) = 1+\alpha\exp\left[-\left(\frac{d}{\beta}\right)^2\right], \tag{2}$$

because in our experience the nonregularity of the preceding models makes reliable inference difficult (see also Diggle, Morris, and Wakefield 2000). Note that $\alpha = 0$ corresponds to no relationship between distance and risk.

4. BAYESIAN FORMULATION

4.1 The Model

In the context of aggregated data, our approach to the detection of excess risk close to a point source is to embed a simple location–risk model within a disease mapping framework. For a background to disease mapping, see Mollié (1996). In particular, we utilize the following three-stage hierarchical model.

First Stage: Data Model. For the observed count in area i we have

$$Y_i|\lambda_i \sim \text{Poisson}(E_i\lambda_i)$$

with

$$\log\lambda_i = \log\rho + Z_i^T\phi + \log f(d_i;\theta) + V_i + U_i, \tag{3}$$

for $i = 1, \ldots, n$. Following Besag, York, and Mollié (1991), we incorporate both nonspatial (V_i) and spatial (U_i) random effects.

Second Stage: Overdispersion Model. For the nonspatial (unstructured) random effects V_i, we assume

$$V_i|\sigma_v^2 \sim_{\text{iid}} N(0, \sigma_v^2),$$

whereas for the spatial random effects U_i there are a number of possibilities. In a Markov random field (MRF) model the spatial dependence is modeled through the *conditional* distributions $U_i|U_j, j \in \partial i$, where ∂i denotes the set of *neighbors* of area i, $i = 1, \ldots, n$. In a *joint* model, the collection $U = (U_1, \ldots, U_n)^T$ is modeled via a multivariate specification. We base our analysis on 1981 EDs. Unfortunately, the boundaries of these regions are unavailable to us; we only have the population-weighted centroids of each area. This restricts the range of models for spatial dependence that we may use. In particular, the frequently employed MRF intrinsic Gaussian autoregression (IGAR) model (e.g., Besag et al. 1991), with neighbors taken as areas with a common boundary, cannot be employed. We choose to use a joint (stationary) multivariate normal spatial model. In the following discussion, let $N_n(\mu, \Sigma)$ denote the n-dimensional normal distribution with mean vector μ and variance–covariance matrix Σ, and let d_{ij} denote the distance between the centroids of areas i and j. We then assume

$$U|\sigma_u^2, \psi \sim N_n(0, \sigma_u^2\Sigma_u(\psi)),$$

with the (i, j)th element of the correlation matrix $\Sigma_u(\psi)$ taken to be $\exp(-d_{ij}\psi)$. The parameter $\psi > 0$ reflects the strength of the spatial dependence. To interpret ψ, we note that the distance at which correlations fall to δ, $0 < \delta < 1$, is given by $\log(\delta^{-1})/\psi$. So as ψ decreases, spatial correlations increase. With this model, σ_u and σ_v are comparable because they are both marginal standard deviations. We note that we could utilize the IGAR model with a distance-based definition of neighborhoods; see Best, Arnold, Thomas, Waller, and Conlon (1999). Cressie and Chan (1989) considered distance-based weighting schemes in which the extent of spatial dependence is assessed via a variogram.

Third Stage: Prior Distributions. At this stage, we specify prior distributions for θ, ρ, ϕ, and for the parameters of the second stage distribution; for model (3) these parameters consist of σ_v, σ_u, ψ.

The preceding hierarchy, with distance–risk modeled via (2), is a nonlinear ecological regression model, Richardson (1992) provided a general discussion of ecological studies. Note that in the null model [i.e., the model in which $f(d;\theta) = 1$ for all d], $\exp(V_i)$ and $\exp(U_i)$ represent, respectively, nonspatial and spatial contributions to the residual relative risk of area i [relative to the overall risk $\rho E_i\exp(Z_i^T\phi)$].

Our aim is to provide both a mechanism for accommodating overdispersion and also, via examination of the posterior distributions of the V_i, U_i, a diagnostic to aid in modeling risk as a function of distance (or more generally, spatial location). We note that the use of spatial random effects in this context is contentious because the true risk–distance relationship may be smoothed away due to confounding between exposure (distance here) and unmeasured risk factors that are being accommodated by the U_i. This issue is controversial. Lawson (1996) advocated the modeling of spatial dependence in contexts such as those considered here to account for both "unobserved heterogeneity in the environment" and the natural clusters that occur with some diseases "due to possible genetic or even viral aetiology." Bithell (1996), in response, strongly opposed this view on the grounds that "interpretation becomes more

difficult and it is likely that estimates of the parameters of primary interest become less precise and stable as we attempt to gain more information from modest data sets." Due to these issues, we address the sensitivity of conclusions by considering analyses with and without spatial random effects, and with a range of prior distributions.

4.2 Predictive Distributions

An important public health question is, "What is the distribution of the number of health events we expect in a particular region and time period?" A predictive distribution for the number of cases of disease that will occur is an important aid to answering this question. Consider an area with expected number E^*, area-level covariates Z^*, and area centroid a distance d^* from the putative source. Here E^* depends on both the population at risk in the area and the time period under consideration. We denote by Y^* the random variable that represents the number of cases over this population and time period. We are then interested in

$$\Pr(Y^*|\text{data}) = \int \Pr(Y^*|\lambda) \times p(\lambda|\text{data})\, d\lambda, \tag{4}$$

where $Y^*|\lambda \sim \text{Poisson}(\lambda E^*)$.

Various choices are available for the relative risk function λ. An obvious choice is

$$\lambda = \rho f(d^*; \theta) \exp(Z^{*T}\phi), \tag{5}$$

although alternatives are available. If the area under consideration is one of the original study areas, with index i^* say, then we may consider

$$\lambda = \rho f(d_{i^*}; \theta) \exp(Z_{i^*}^T\phi + V_{i^*} + U_{i^*}), \tag{6}$$

where V_{i^*}, U_{i^*} are realizations from the posterior distribution $p(V_{i^*}, U_{i^*}|\text{data})$. A third possibility is

$$\lambda = \rho f(d^*; \theta) \exp(Z^{*T}\phi + V^* + U^*), \tag{7}$$

where V^*, U^* are considered to be exchangeable random effects and are drawn from the predictive distribution $p(V, U|\text{data})$. This predictive distribution is given by

$$p(V, U|\text{data}) = \int p(V|\sigma_v^2)p(U|\sigma_u^2, \psi)p(\sigma_v^2, \sigma_u^2, \psi|\text{data})\, d\sigma_v^2\, d\sigma_u^2\, d\psi.$$

The choice of relative risk function (5)–(7) is not straightforward and depends on the use to which the predictive function is to be put. For model checking, (6) may be preferable because area-specific random effects for the period of data collection are required. As discussed in Section 1, the random effects V and U may be accounting for unmeasured risk factors and/or data anomalies. To predict the future number of cases, if we believe that U and V are mainly accounting for data anomalies, then we may exclude the random effects and use (5). *Persistent* unmeasured risk factors again lead to the use of (6). Finally, if we wish to acknowledge the overdispersion in the data, (7) may be used; this is consistent with accounting for unmeasured, nonpersistent risk factors, and the possibility that data anomalies are reflected in future observations. It is straightforward to obtain a predictive distribution for the number of cases within all of the EDs within a particular radius of the point source.

We now discuss how the excess number of cases associated with the point source may be determined. We note that it is not possible to determine an estimate of the excess number of cases that *result* from exposure to the point source, that is, to place any causal interpretation on the risk–distance association. Suppose we fit the monotonic distance–risk model. We may obtain the predictive distribution of cases using the foregoing procedure, where the number of cases follows a Poisson distribution with mean $\sum_i E_i\lambda_i$ and i indexes the areas of the region of interest. In this situation, a careful choice of λ_i must be made, with (5) being the obvious candidate. We want to compare this with the hypothetical situation in which the point source is removed (after a suitable lag period). It is not appropriate to use the *null* model for this comparison because if there is an increase of risk with proximity to the point source, then this will be reflected in an increased value for ρ that will produce an increased number of cases. Instead the relevant Poisson mean is obtained by using the posterior distribution for ρ from the monotonic, and not the null, model. We may then compare the number of cases under the two predictive distributions. We illustrate such predictive distributions in Section 6. We reiterate that an obvious difficulty here is that ρ and θ from the monotonic model may reflect risk factors other than exposure to the emissions of the incinerator.

4.3 Computation

The model that we have specified is not analytically tractable and so we utilize Markov chain Monte Carlo (MCMC); see Gilks, Richardson, and Spiegelhalter (1996). Specifically we use the BUGS software (Spiegelhalter, Thomas, and Best 1998).

With such a sampling-based approach, it is straightforward to evaluate the predictive distribution given by (4), because

$$\Pr(Y^*|\text{data}) \approx \frac{1}{S}\sum_{s=1}^{S} \Pr(Y^*|\lambda^{(s)}),$$

where $\lambda^{(s)}$ is obtained from the relevant posterior distribution. For example, with λ given by (6), we would use samples from $p(\rho, \phi, V_{i^*}, U_{i^*}, \theta|\text{data})$. Samples from the posterior distribution of any function of interest may also be obtained, easily. Examples that we examine in Section 6 include residuals and the value of the risk–distance function at specified distances.

5. PRIOR DISTRIBUTIONS

In this section we describe how we specify the prior distributions for ϕ, ρ, θ and for the (hyper) parameters of the distributions of the random effects U and V. In environmental epidemiology, in general, it is not straightforward to specify prior distributions, because the simple models that are utilized are far removed from the exposure–confounder–disease mechanism. We first describe a preliminary study that we carried out to inform the specification of the prior distribution.

5.1 Preliminary Mapping Study

In Section 4.1 we outlined the potential problems of confounding between the effects of the point source and the unmeasured risk factors being accounted for by the random effects with spatial structure. In particular, we may dilute the pollution effect with the inclusion of random effects. To attempt to minimize this possibility, we carried out a mapping study in an area that did not contain the putative source to obtain a baseline measure of spatial variability in residual log relative risk in the absence of the pollution source. We note that in other types of investigation it may be more difficult to find a region in which the exposure is not present (e.g., an ecological study in which the exposure is the concentration of magnesium in the water supply). This mapping study also will give us information on the relationship between stomach cancer and deprivation. There is the possibility of over- or under-adjustment for deprivation and so it is of interest to examine the size of the association. The reference region that we use is the census district that contains (but excluding) the study region. This region contains 209 EDs. We obtain data on the incidence of stomach cancer in the period 1974–1986. The right-hand panel of Table 1 summarizes the population and health data for this region. The expected numbers were calculated via internal standardization using age–sex rates determined marginally (as opposed to a joint approach in which these parameters are estimated simultaneously with the other parameters of the model). When reference rates were calculated, we included the cases from the study region (which is why the sum of the observed counts does not equal the sum of the expected counts in Table 1). We note that this could distort the conclusions if, for example, the study region was a large fraction of the reference region and if the age–sex distribution close to the incinerator was atypical compared to the region as a whole (i.e., if the age-sex distribution is not independent of distance). For example, if there were raised incidence close to the incinerator and older people lived closer to the incinerator, then this would produce artificially inflated rates for the elderly.

Figure 2(a) displays the SIRs, Y_i/E_i, of the reference region; we see that the SIRs are widely spread. The gap between the two sets of areas corresponds to the circle of 7.5 km centered on the incinerator. The displayed region falls into two pieces because the central portion corresponds to the study region that has been excluded. From Table 1 we observe that the maximum relative risk estimate is 4.4, and the 5% and 95% points of the empirical distribution are 0 and 2.95. In a small-area study such as this it is well known that relative risk estimates may be dominated by sampling variability (Clayton and Kaldor 1987) and a hierarchical modeling approach has been advocated to smooth the ensemble of estimates.

The disease mapping model is given by $Y_i|\lambda_i \sim \text{Poisson}(E_i\lambda_i)$, where

$$\log \lambda_i = \log \rho + Z_i^T \phi + V_i + U_i \tag{8}$$

for $i = 1, \ldots, 209$. Again, V_i and U_i represent unstructured and spatially dependent random effects, respectively. The measure of SES that we utilize in the main study is an index based on three census variables (Section 2). Unfortunately this index

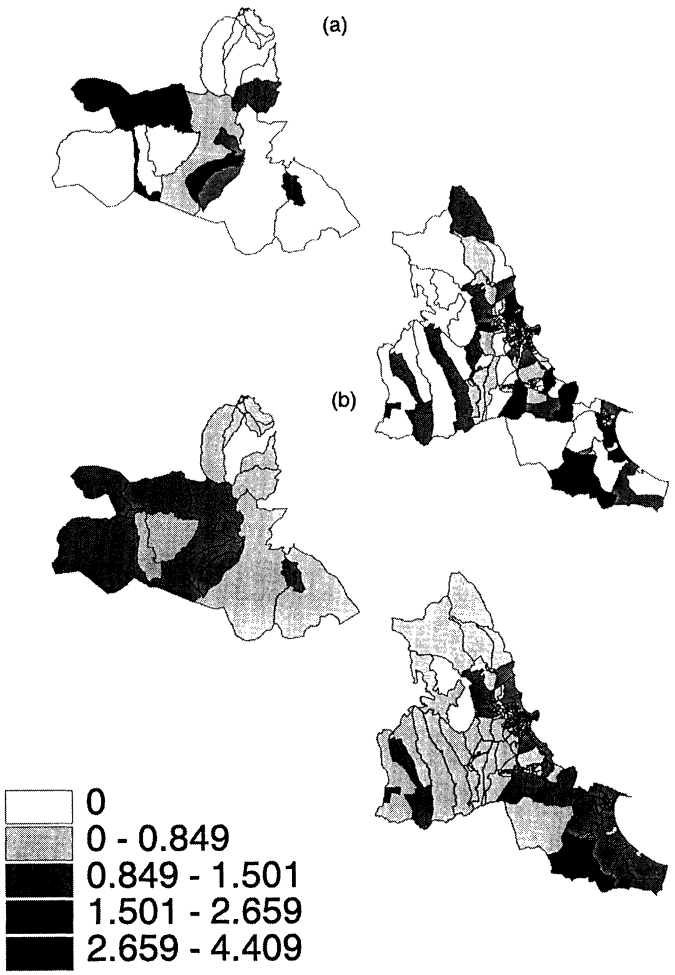

Figure 2. Maps of (a) Raw Relative Risk Estimates (SIRs), and (b) Smoothed Relative Risks From the Random Effects Model, in the Reference Region.

was not readily available for the reference region, so instead we use the so-called Carstairs index (Carstairs and Morris 1991). This index is census-based and is closely related to that used in the study region, except it adds a fourth variable, access to a car. We compared the two index for areas in which both were available and found them to be highly correlated. Hence it was thought that the regression coefficient for the reference region could be compared with that for the study region.

We now describe the prior distributions that were used in the analysis of data in the reference region. For the precision parameters σ_u^{-2} and σ_v^{-2}, we use the prior suggested by Kelsall and Wakefield (1999), namely Ga(.5, .0005). This prior often provides a realistic range of residual relative risks. By contrast, the commonly suggested choice of Ga(.001, .001) may lead to artificial inflation of the estimate of the random effects standard deviation because very small values are very unlikely under this prior. For the spatial dependency parameter ψ, we specify a uniform prior. We note that very small and very large values of ψ (corresponding to very strong and very weak spatial dependence, respectively) will not be identifiable from the data because of the finite size of the region and the lack of areas very close together. We assume that the two extreme dependencies that we can detect have correlations of

r $(0 < r < 1)$ at distances of d_1 and d_2, to give the endpoints of the distribution as $-\log r/d_k$, $k = 1, 2$. The size of the reference region is 58.9 km and we take $r = .01$, $d_1 = 29.5$, and $d_2 = 2.95$ to give the prior $U(.16, 1.56)$.

We are particularly interested in the random effects standard deviations σ_u and σ_v. Table 2 gives summaries of the posterior distribution for these parameters, ρ, ψ, and ϕ, under a range of analyses. The first two lines are comparable and give identical inference for ϕ. In the quasilikelihood model in which $\text{var}(Y_i) = \kappa \times E[Y_i]$, we obtained $\hat{\kappa} = 1.42$, indicating 42% overdispersion. In the joint model, we see that the variance of the spatial random effects is much larger than the variance of the unstructured random effects. Examination of the spatial dependence parameter reveals that the correlations fall to .5 and .01 at distances of approximately 0.8 and 5.2 km, respectively, indicating short range dependence.

The estimate of the ecological regression coefficient ϕ that describes the relationship between incidence and deprivation is sensitive to the choice of model for the random effects. In particular, when spatial random effects are included, the coefficient moves closer to being significantly different from zero, but with the opposite sign to that expected. In all cases, the 95% interval contains zero, however.

Figure 2(b) displays the posterior means of the relative risks after global and local smoothing (corresponding to the inclusion of unstructured and spatial random effects, respectively) have been carried out using the hierarchical model. Comparison with Figure 2(a) reveals that the spread in the relative risk estimates is far narrower. For example, there is 2.4-fold variability between the 5% and 95% percentiles of the empirical distributions of the posterior means of the smoothed relative risks (the actual values are .6 and 1.43).

5.2 Prior for ρ

The parameter ρ is a nuisance parameter that reflects the overall incidence in the area relative to the reference region that provides the rates for the expected numbers. For the null model, the MLE of ρ is given by Y_+/E_+, where $E_+ = \sum_i E_i$ and $E_i = \sum_j N_{ij} p_j$, and so we may obtain an approximate range for ρ by considering how large or small we might expect Y_+ to be. For example, if the probabilities of disease were twice as great across the study region when compared to the reference region, then we would have $\rho = 2$. We specify the normal prior $N(M_\rho, V_\rho)$ for $\log \rho$ with $M_\rho = 0$ and $V_\rho = .25$. This prior gives, for example, $\Pr(.5 < \rho < 3) = .903$. We note that, with respect to this parameter, the likelihood is well behaved and so we often will be able to specify a relatively flat prior.

5.3 Prior for ϕ

We now consider the regression coefficients that describe the relationship between confounders and risk. We assume that the priors for each regressor are independent and as a generic confounder suppose that a measure of SES is the covariate under consideration. We assume that the prior for ϕ is a normal distribution with mean M_ϕ and variance V_ϕ. One method for selecting M_ϕ and V_ϕ is the following. We first set $M_\phi = 0$ and then suppose that ΔR is the maximum ratio of risk that is thought to be possible between the most deprived $(Z_{\max})$ and the least deprived $(Z_{\min})$ areas. This ratio should be informed by background epidemiological knowledge and, in particular, previous studies. Then, whereas the risk in an area with an expected number of cases E is given by $\rho E f(d, \theta) \exp(Z^T \phi)$, we have

$$\Delta R > \exp[\phi(Z_{\max} - Z_{\min})], \tag{9}$$

which gives

$$\phi < \frac{\log(\Delta R)}{Z_{\max} - Z_{\min}}.$$

Let $\phi_{\max} = |\log(\Delta R)/(Z_{\max} - Z_{\min})|$. To place this largest possible value that may occur in the tail of the prior, we set $3\sqrt{V_\phi} = \phi_{\max}$. Alternatively, we may use a reference analysis to provide prior information.

The variance was determined using figure 1 of Elliott (1996) in which the ratio of stomach cancer incidence displays an approximate twofold increase from the least to the most deprived quintiles (which is consistent with Nomura 1997). For our data, we have the same score, but on a continuous rather than a discrete scale. We therefore use $\Delta R = 3$ in Equation (9), from which we obtain $V_\phi = .045^2$. This gives a 95% prior interval of $(-.088, .088)$. We note that this interval contains the posterior medians from the reference analysis in Table 2.

5.4 Prior for σ_v^2, σ_u^2, and ψ

The priors for the variance components σ_u^2 and σ_v^2 are chosen to be inverse gamma distributions. We utilize two sets of priors: an uninformative set that consists of Ga(.5, .0005) priors (discussed in Section 5.1) for both precisions; and an informative set in which the priors for σ_u^2 and σ_v^2 are derived from the preliminary study. Specifically, we match up the 5% and 95% points of the prior distributions with the posterior distributions obtained from the preliminary study. Following this procedure, we obtain prior distributions $\sigma_u^{-2} \sim \text{Ga}(1.0, .1)$ and $\sigma_v^{-2} \sim \text{Ga}(.5, .0005)$. Hence for the latter we see that the posterior and prior were identical in the preliminary study.

For ψ our default choice for the study region is the uniform distribution $U(1.2, 12.3)$. These endpoints are based on the procedure outlined in Section 5.1 with $r = .01$, $d_1 = 3.75$, and $d_2 = .38$.

Note that the priors for the variances are independent of the distance–risk model that we are using because the overdispersion is that which is inherent in the incidence of stomach cancer, even though we would expect the values of the variance components to decrease as we increase the complexity of the model.

5.5 Prior for α

For both of the parameters of interest, α and β, we concentrate on the likelihood function by using uniform prior distributions. For α we take a uniform prior distribution on the range $(-1, \alpha_{\max})$ with $\alpha_{\max}$ taken to be the maximum plausible increase in risk at source based on current epidemiological knowledge. In studies of environmental pollution from point sources, the increases in risk are often modest (unless there is an accident that results in a large increase of pollutants). Occupational studies tend to produce much larger increases.

Table 2. Posterior Quantities, 50% (2.5%, 97.5%) for the Disease Mapping Analysis of the Reference Region

Random effects model		ρ	σ_v	σ_u	ψ	ϕ
Quasilikelihood		0.93 (0.81, 1.1)	—	—	—	−0.025 (−0.083, 0.033)
Unstructured	—	0.85 (0.72, 0.99)	0.43 (0.024, 0.64)	—	—	−0.025 (−0.085, 0.032)
—	Spatial	0.77 (0.55, 1.0)	—	0.38 (0.11, 0.65)	0.96 (0.25, 1.5)	−0.059 (−0.13, 0.0049)
Unstructured	Spatial	0.75 (0.55, 0.96)	0.035 (0.015, 0.37)	0.39 (0.21, 0.63)	0.88 (0.23, 1.5)	−0.062 (−0.13, 0.0029)

A number of point source studies have been carried out by the Small Area Health Statistics Unit in the United Kingdom. Examples include all incinerators of waste solvents and oils in Great Britain (Elliott et al. 1992a), a single petrochemical works at Baglan Bay, Wales (Sans et al. 1995), radio and TV transmitters (Dolk et al. 1997a, b), cokeworks (Dolk et al. 1999), a pesticides factory (Wilkinson et al. 1997), and industrial complexes that include major oil refineries (Wilkinson et al. 1999). These have reported excesses in risk at source in the range 0.1–1.0, which gives us a lower bound on the size of $\alpha_{\max}$. In Elliott et al. (1992b), a point source study was carried out to investigate increased risk of mesothelioma in the vicinity of Plymouth docks. This analysis revealed an estimated excess of 11 at source, but further analysis revealed that this excess was due to occupational, rather than environmental, risk factors.

5.6 Prior for β

As a prior for β we take a uniform distribution on the range $(0, \beta_{\max})$. To determine $\beta_{\max}$, we think in terms of the more intuitive distance–risk function. In particular, we specify two values, r and q, such that at a distance $r \times d_{\max}$ $(0 < r < d_{\max})$ we believe that the risk will have fallen to below $1 + q\alpha (0 < q < 1)$. The size that was chosen for the study region aids in this choice. This formulation implies that

$$f(rd_{\max}) = 1 + q\alpha \geq 1 + \alpha \exp\left[-\frac{(rd_{\max})^2}{\beta^2}\right],$$

which gives

$$\beta \leq \beta_{\max} = \frac{rd_{\max}}{(-\log q)^{1/2}}.$$

Figure 3 shows 20 simulations from the prior distribution with $d_{\max} = 7.5$ km, $\alpha_{\max} = 10$, and $\beta_{\max} = 3.15$. The latter is derived from the choices $r = .9$, $q = .01$ and corresponds to the belief that at a distance of 6.75 km from source, the excess risk will be less than 1% of the excess at source. Note from the figure that, as desired, all of the simulations from our model produce negligible risk by 7.5 km. If this were not true, then a larger study region would need to be chosen.

In Section 6 we address the sensitivity to the prior by considering values of $\alpha_{\max}$ in the range 2–20 and values of $\beta_{\max}$ in the range 1–7.

6. FULL ANALYSIS

In this section we use the prior distributions of the previous section and carry out analyses of the stomach cancer incidence data in the proximity of the incinerator. The analyses were carried out using the MCMC strategy outlined in Section 4.3. Convergence was assessed via informal assessment of posterior summary distributions across two chains started from different points in the parameter space. On this basis, a burn-in of 1,000 iterations was used and, depending on the model, a further 30–50,000 iterations produced samples that were used for inference.

Table 3 summarizes the posterior distribution under various prior specifications. When we pass from the null to the monotonic model, the posterior median of the nonspatial lack of fit σ_v is reduced from .052 to .044, whereas the corresponding reduction for the spatial standard deviation σ_u is .50 to .29. This shows how the unexplained variability is explained by the distance–risk relationship.

We note that again the estimate of the regression coefficient ϕ is sensitive to the choice of random effects distribution and on the prior. Again all 95% intervals contain zero, however. Given that, if anything, the relationship was reversed in the reference study, we should investigate whether we are overadjusting for deprivation. The danger is that deprivation has little effect here and, in fact, the effect of the incinerator is being reduced by allowing some of the excess risk to be absorbed into ϕ (see Fig. 1). We carried out an analysis in which ϕ was fixed at the posterior mean from the reference analysis −.062, but again there was little sensitivity in the parameters of interest (Table 3).

We also carried out a number of analyses to investigate the effect of the prior on ψ. Although ϕ was sensitive to the choice, there was little change in inference on α and β when various uniform priors on ψ were taken (Table 3 displays the

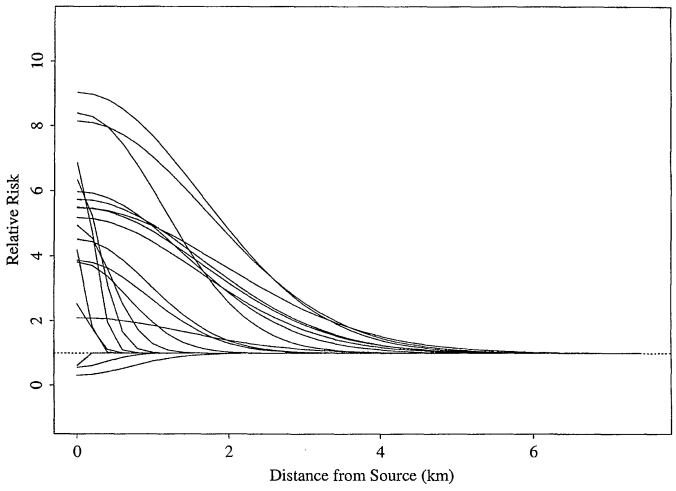

Figure 3. Twenty Simulations From the Prior Distribution of the Risk–Distance Relationship. Prior specifications (see text) were d_{max} = 7.5 km, $\alpha_{max} = 10$, $\beta_{max} = 3.15$, $r = 0.9$, and $q = 0.01$.

Table 3. Posterior Quantiles, 50% (2.5%, 97.5%), for the Full Analysis of the Study Region. The Fixed Values Correspond to the Posterior Medians From the Reference Analysis

Model	Random effects	Prior dist.	α	β	ρ	σ_v	σ_u	ψ	ϕ
Null	$V_i + U_i$	1	—	—	0.98 (0.57, 1.5)	0.052 (0.014, 0.58)	0.50 (0.21, 1.0)	8.9 (2.3, 12.0)	0.053 (−0.024, 0.12)
Monotonic	V_i	2	6.4 (2.1, 9.8)	0.43 (0.28, 0.71)	0.91 (0.67, 1.2)	0.048 (0.014, 0.42)	—	—	0.0010 (−0.061, 0.067)
Monotonic	U_i	2	6.4 (2.1, 9.8)	0.43 (0.28, 0.70)	0.90 (0.64, 1.2)	—	0.045 (0.014, 0.45)	6.8 (1.5, 12.0)	0.0012 (−0.061, 0.068)
Monotonic	$V_i + U_i$	2	6.3 (2.0, 9.8)	0.43 (0.28, 0.72)	0.91 (0.67, 1.2)	0.045 (0.014, 0.37)	0.041 (0.014, 0.30)	6.7 (1.5, 12.0)	0.0015 (−0.061, 0.067)
Monotonic	$V_i + U_i$	1	5.9 (1.2, 9.8)	0.45 (0.25, 1.80)	0.84 (0.46, 1.3)	0.044 (0.014, 0.35)	0.32 (0.16, 0.78)	7.5 (1.5, 12.1)	0.0031 (−0.065, 0.075)
Monotonic	$V_i + U_i$	1	7.6 (2.9, 9.9)	0.48 (0.30, 1.10)	0.85 (0.50, 1.3)	0.047 (0.014, 0.41)	0.30 (0.15, 0.72)	7.04 (1.5, 12.1)	−0.062 (*fixed*)
Monotonic	$V_i + U_i$	1	6.4 (2.0, 9.8)	0.43 (0.28, 0.76)	0.87 (0.55, 1.6)	0.046 (0.013, 0.47)	0.29 (0.15, 0.80)	0.88 (*fixed*)	0.00050 (−0.062, 0.071)

*Prior distributions are 1, informative or 2, uninformative.

case in which ψ was fixed to the posterior median from the preliminary study).

The posterior marginal distributions for ρ, α, β, ρ, σ_v, and σ_u for the analysis with the informative prior are displayed in Figure 6. The prior distributions are also presented on this plot. These histograms are useful and easily constructed using the MCMC sampling-based approach. Profile likelihoods, which are the classical equivalent, are far more difficult to determine for models such as those considered here. It is clear from Figure 6(c) that there is very little information in the data concerning an upper value for α. With the MCMC approach, it is straightforward to make statements such as $\Pr(\alpha > 0|\text{data}) = .999$. We also note that, as was our aim, the posterior distributions of the random effects variances [panels (e) and (f)] are close to their prior distributions.

Figure 5 shows the 5%, 50%, and 95% posterior intervals for the fitted curve $f(d, \theta)$ versus d, along with the MLE from the quasilikelihood approach outlined in Section 3.2.

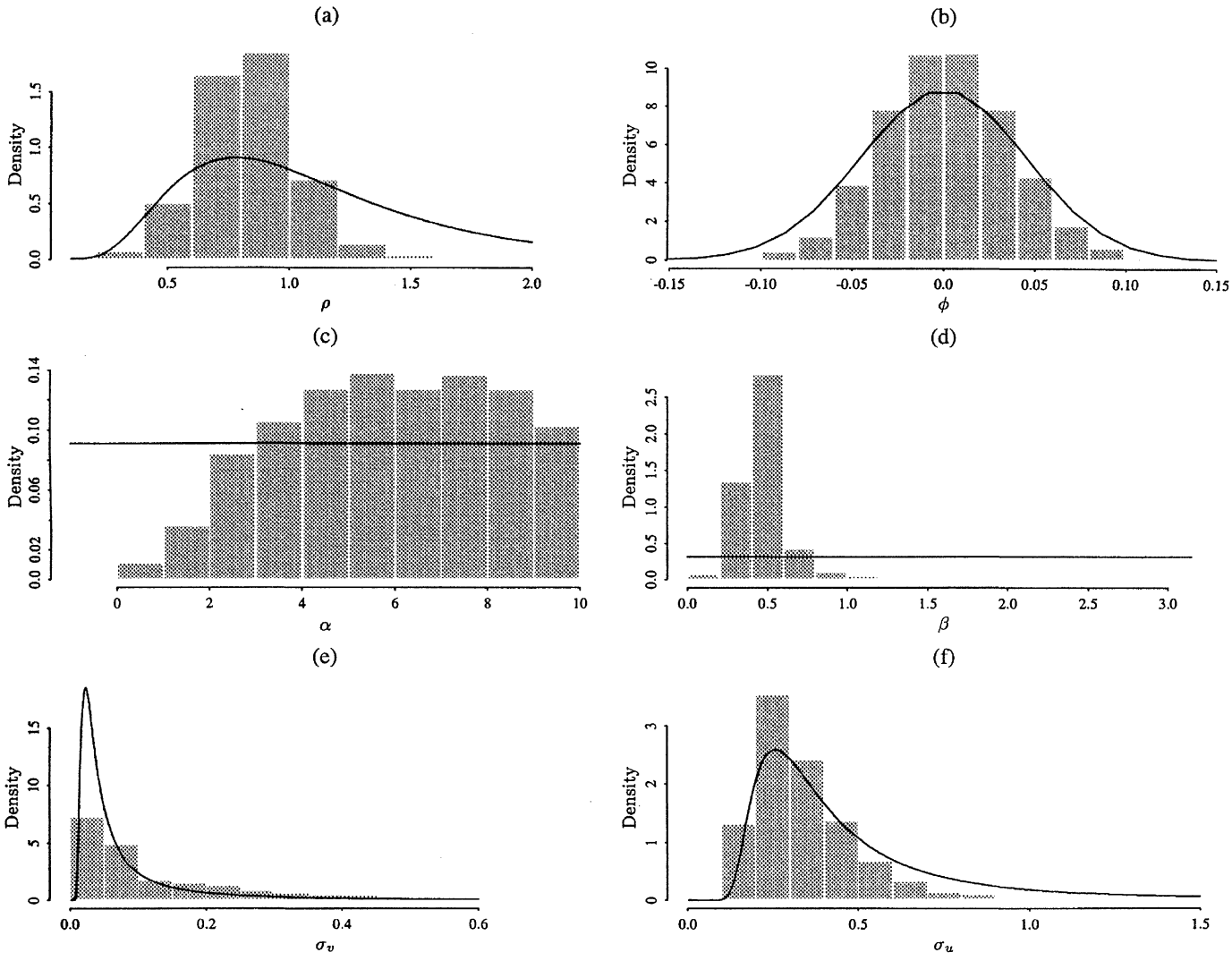

Figure 4. Posterior Distributions for (a) ρ, (b) ϕ, (c) α, (d) β, (e) σ_v, and (f) σ_u. The solid line represents the prior distribution.

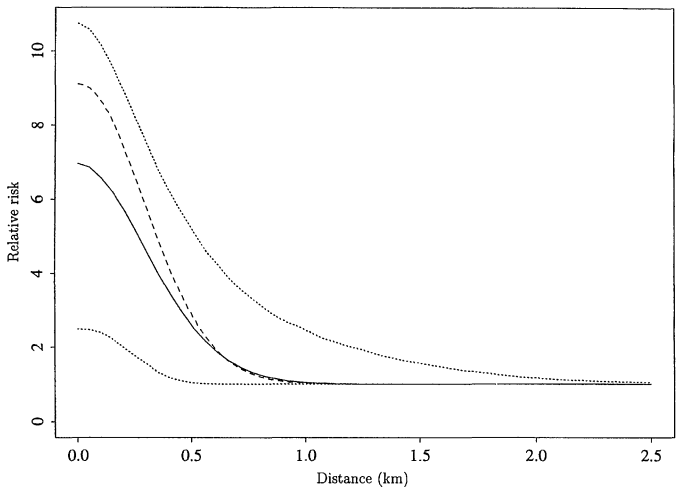

Figure 5. Bayesian Median Point Estimate (solid line) and MLE (dashed line) of Modeled Risk Function. Also shown is a Bayesian 95% credible region (dotted lines). For clarity, the distance axis has been truncated at 2.5 km.

The posterior summaries were constructed as follows. A parameter of interest $f(d;\theta)$ was defined for a fine discretization of distances d. The posterior distribution of the risk function was then summarized by substituting samples $\theta^{(s)}, s = 1, \ldots, S$, from $p(\theta|\text{data})$ into f. The upper endpoint of the interval estimate for small values of d again indicates the lack of information in the data with regard to an upper value for α. Notice that the 90% interval is above unity to a distance of approximately 300 m and that the increased risk at source decreases rapidly with distance. This observation indicates that the study region was chosen to be large enough and any effect is very localized. Almost identical plots were obtained from analyses with and without spatial effects. We note that the results are highly influenced by SIRs of 6.49 at a distance of 0.2 km and 5.98 at 0.45 km. It would be advisable to check the cases in these areas because data anomalies will have large influence.

Figure 6 shows the posterior medians of the error terms $V_i, U_i, i = 1, \ldots, 44$, plotted against distance for the null and monotonic models. A plot of these random effects against

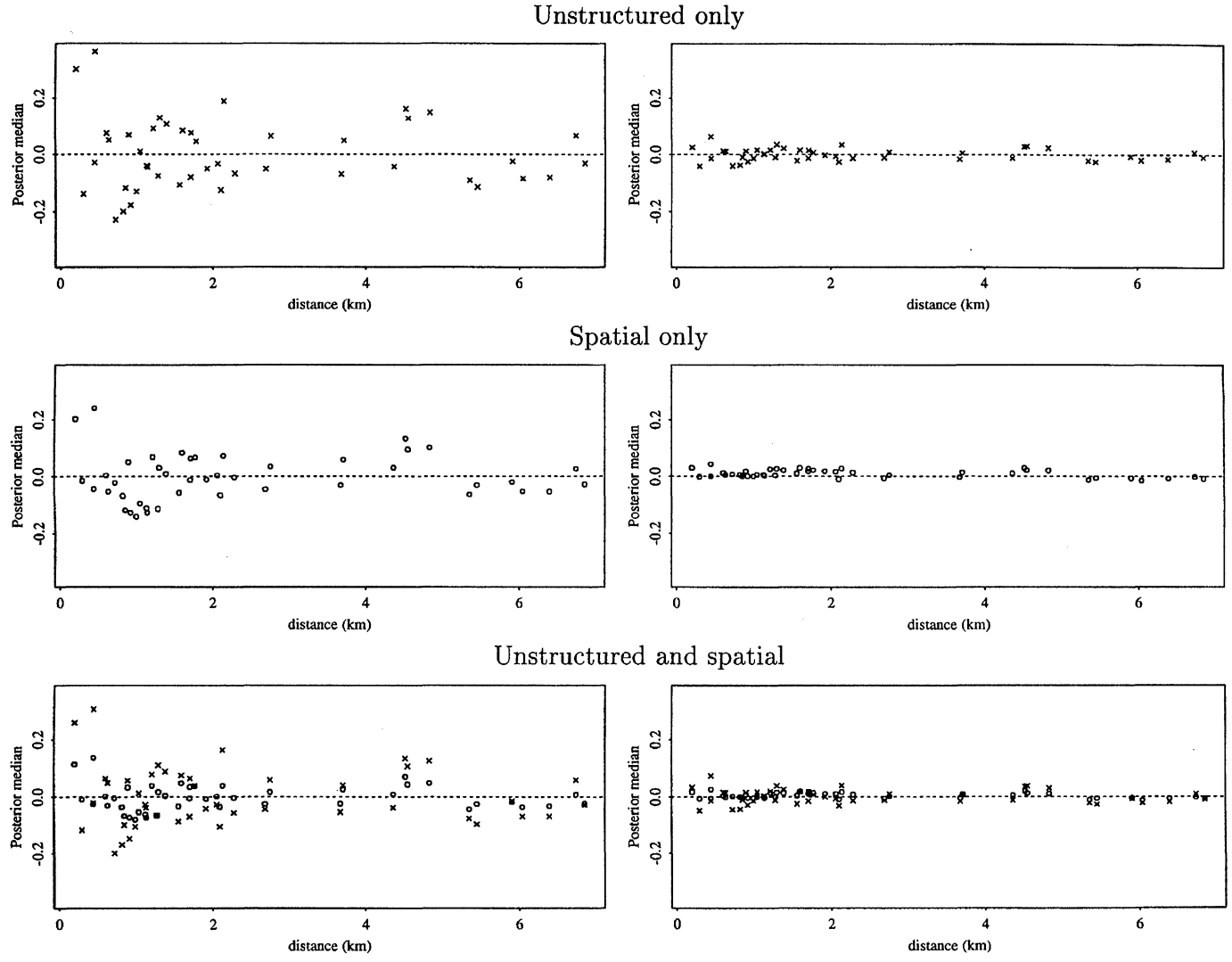

Figure 6. Posterior Medians of the Random Effects V_i (crosses) and U_i (circles), $i = 1, \ldots, 44$. The left panel displays the random effects for the null model, and the right panel displays the random effects for the monotonic model.

orientation showed no systematic pattern (hence providing some justification for our simple isotropic model). The differences in the ranges between the left and right panels shows the improvement in fit from incorporating the distance–risk model. We also calculated the posterior distributions of the Bayesian first-stage residuals given by $(Y_i - E_i\lambda_i)/(E_i\lambda_i)^{1/2}$. These showed the same pattern as the second-stage residuals V_i, U_i, although the latter were more informative here because the first-stage residuals are difficult to interpret for small counts because they may take only a small set of values.

Figure 6 displays the sensitivity of the posterior distributions for α and β to the upper limits on the uniform priors, $\alpha_{\max}$ and $\beta_{\max}$. We see that the prior does have an influence, although there is reasonable stability for β in the range (1, 6). For α, the upper quantiles are particularly sensitive to the prior and essentially remain close to $\alpha_{\max}$. As previously commented, this reflects the fact that there is little information in the likelihood due to the sparsity of data close to source (both in terms of the number of EDs that are close and the size of the expected numbers in these EDs).

From an epidemiological perspective, the modeled risk function is of obvious interest, but from a public health perspective, the predicted number of cases of stomach cancer can be highly informative. Figure 8 shows the predicted survivor functions $\Pr(Y_k^* \geq y|\text{data})$ $(y = 0, 1, 2, \ldots)$ for four EDs at various distances from the incinerator and under the null and monotonic models. From this plot we may, for example, state that the probability of 5 or more deaths in an ED whose centroid is .2 km from the source and over the same time as the data collection period is virtually zero under the null model, but approximately .7 under the monotonic model.

For illustration, we compare predictive distributions for the number of cases over the whole study region with and without random effects. The expected numbers were taken as those in the study that correspond to the predictions being over the same population and number of person years as the study. When we include random effects, we sample U, V from the predictive distribution of random effects, rather than the posterior for $p(U, V|\text{data})$. This approach is consistent with the random effects representing risk factors that are randomly distributed across areas and not specific to the areas under study. Table 4 gives predictive distribution summaries under a variety of models. As expected from the discussion in Section 4.2, under the null model the predicted number of cases is almost exactly equal to the number observed. In the null model, the posterior median of ρ is .98, whereas in the analysis in which the monotonic model was used, the estimate was .87. When random effects are included, the predictive distribution is far wider. Comparisons between the last four lines of the table indicate a difference of approximately 18 cases between the predictions with and without the monotonic risk model.

7. DISCUSSION

In this article, we have taken a Bayesian approach to the modeling of disease risk in relation to a point source. We have embedded the approach of Diggle et al. (1997) within a hierarchical framework. This hierarchy includes random effects that allow for spatial and nonspatial overdispersion and may be used as diagnostic tools to aid in model refinement. With this framework, the full range of Bayesian techniques can be utilized to provide informative analyses. In particular, we have stressed the incorporation of prior information and predictive distributions for inference, and considered the assessment of model adequacy and sensitivity analyses. We now describe a number of extensions and issues.

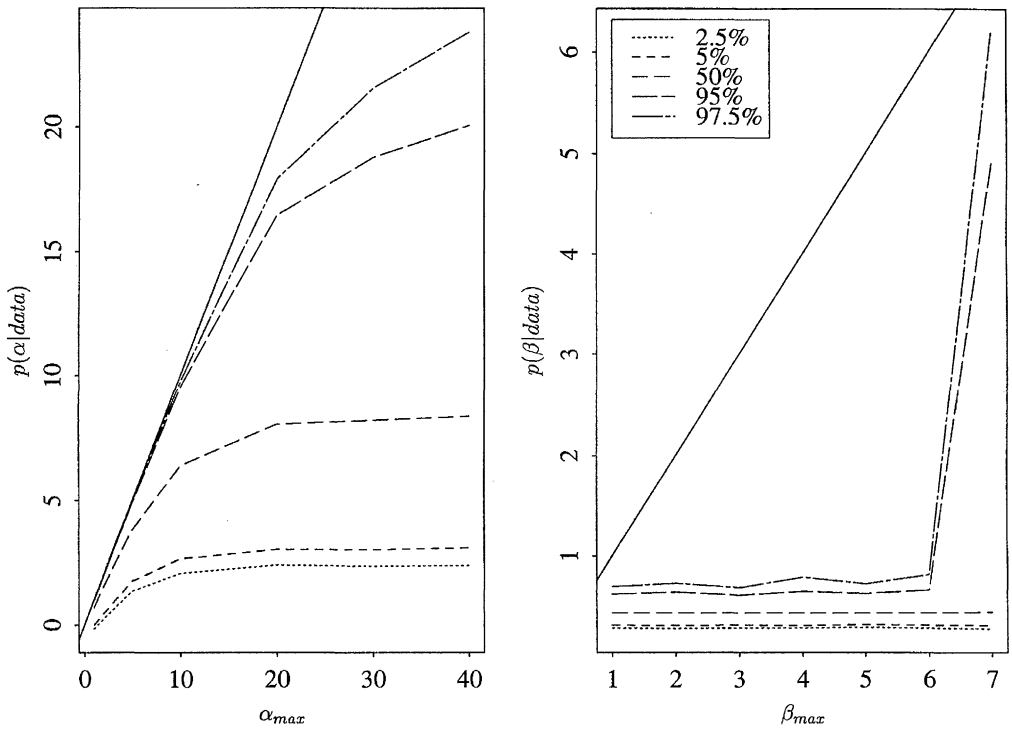

Figure 7. Sensitivity of Posterior Distributions for (a) α (as a function of α_{max}) and (b) β (as a function of β_{max}). Solid lines denote $y = x$.

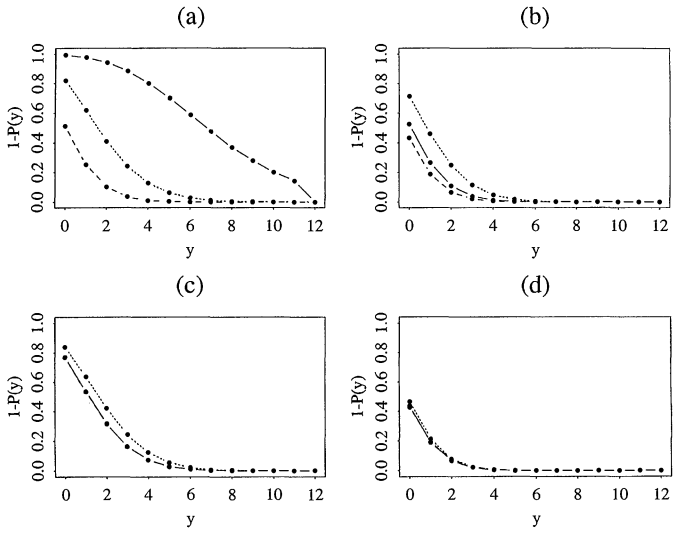

Figure 8. Predictive Survivor Functions, $S(y) = Pr(Y \geq y|data)$, $y = 0,1,2,\ldots$, of the Number of Cases for Enumeration Districts at Distances From the Point Source of Pollution of (a) 0.2 km, (b) 1.1 km, (c) 4.5 km, and (d) 6.9 km. The solid line represents these functions for the monotonic model, the dashed line for the null model, and the dotted line from the posterior under the monotonic model but with the distance–risk relationship removed.

Point Data. If point data are available, then a Bayesian version of the Bernoulli model of Diggle and Rowlingson (1994) can be implemented in a straightforward manner (see Diggle et al. 2000).

Socioeconomic Status. It is well documented that SES can be a strong predictor of health outcomes (e.g., Jolley et al. 1992). Here we have used a nonspecific measure (the Carstairs index), although it is strongly predictive of stomach cancer incidence (Elliott 1996). The explanation of the strong relationship between area-level measures of SES and health remains an important and challenging problem.

Errors in Variables. In general, in spatial epidemiology many of the important components of the model that are treated as fixed are in fact estimates. For example, the health data may be subject to double counting, underascertainment, and coding errors and the population data may be subject to migration and underenumeration. Best and Wakefield (1999) have examined various models for numerator and denominator errors in the context of a mapping study. Exposure measures may be modeled or based on extrapolation, and confounders such as SES may be based on census data determined only on specific occasions. Exposures and confounders may also be measured on individuals even though the health and population data are measured at the area level. Often exposure and confounder variables are incorporated directly into the model without acknowledging that these variables are actually estimates. Each of these misspecifications can be modeled using an errors-in-variables approach. MCMC is then a very natural way to carry out the computations, because the required conditional distributions are of simple form. Richardson and Gilks (1993) provided a discussion of errors in variables in epidemiology from a Bayesian perspective. Unfortunately, validation data are rare in the context considered here and the approach should be viewed as a sensitivity analysis.

More Realistic Disease Risk–Location Models. The relationship between disease risk and spatial location has been assumed to be of simple form. A more complex form, for example, taking into account directional effects, may also be incorporated into our framework, but extensive data are likely to be required to support such a model. In Section 6 we saw great sensitivity to the prior distribution for α, indicating that there is little information in the data for even this very simple model. In general, a modeled pollution surface is preferable to a simple distance measure of exposure.

Prior Choice. The choice of prior distributions in spatial epidemiology remains a challenging issue. In Section 5.1 our uninformative priors for σ_v and σ_u were taken to be identical, although it is not clear that this in any way reflects placing equal prior weight on unstructured and spatial random effects. Independent priors on the two variances may not always be appropriate also.

Data From Many Sources. In the original study (Elliott et al. 1996), data from all 72 municipal incinerators in Great Britain were analyzed. The analysis of the totality of this data clearly make the substantive conclusions far stronger. A natural method for analyzing these data is the following. Let θ_k, denote the parameters of the $k = 1, \ldots, 72$ incinerators, parameterized so that each of the elements lies on the whole of the real line [e.g., $\theta = (\log(\alpha - 1), \log\beta)^T$ for model (2)]. These parameters could then be assigned a distributional form, for example, $\theta_k \sim N(W_k\mu, \Sigma)$, where W_k denote incinerator–study area-specific covariates such as height of chimney–speed of emissions–prevalent wind direction (where such data exist). In other words, separate populations of incinerators are defined. The parameter μ summarizes the average values and the effect of covariates on θ. Elements of the variance–covariance matrix Σ summarize the variability of the elements of θ across all incinerators. This model alleviates the need to specify a strongly informative prior distribution for θ because the data from all sites would choose the appropriate normal distribution, with priors specified for μ and Σ. The effects of confounders could also be allowed to vary between different point sources. For example, deprivation may have a different interpretation in different health regions. These models could be fitted in a straightforward manner using MCMC methods.

Table 4. Median of Predictive Distribution (5% and 95% Quantiles) for the Number of Cases in the Study Area With the Same Populations and Over the Same Time Period (so that the expected numbers are identical to those in the study)

Model	*Random effects*	*Predictive summary*
Null	Not included	84 (61, 111)
Null	Included	73 (25, 183)
Monotonic	Not included	85 (61, 111)
Monotonic	Included	83 (43, 169)
Monotonic *f*	Not included	66 (45, 90)
Monotonic *f*	Included	64 (32, 125)

NOTE: The label "Montonic *f*" denotes that the predictions were obtained using ρ from the monotonic model but with flat risk [i.e., $f(d;\theta) = 1$; see text for details].

Such a modeling approach, combined with predictive distributions, provides an aid in the determination of the potential health effects of a new point source exchangeable with those already examined.

Dolk et al. (1998) use a simple form of this model in a study of congenital malformations in the vicinity of landfill sites across Europe. In their analysis, the excess near to each landfill was assumed random across sites.

Substantive Findings. In this article, we have suggested a hierarchical framework within which point source data may be analyzed. Our approach was illustrated using data from a single incinerator. In the full study, Elliott et al. (1996) examined a range of cancer endpoints across all 72 incinerators in Great Britain and found limited evidence of an excess. Examination of the available case notes and histopathology of liver cancer cases has produced refined estimates of the excess (Elliott, Eaton, Shaddick, and Carter, 2000).

We selected our study area purposefully to illustrate our methodology on an interesting dataset. As a postscript we obtained data from a period immediately following the study period (1987–1991). The incinerator in question closed down in 1976. There were limited data available (22 cases in the study region) and we used data from the county (which is larger than the previously used reference region) to obtain rates for the expected numbers. Population data were obtained from the 1991 census. We included unstructured random effects only and used the default prior distributions of Section 6. In particular the prior on α was uniform on the range $(-1, 10)$. The posterior median for α was 2.9 with 95% interval $(-0.4, 9.2)$. Hence, there does not appear to be strong evidence for a persistent increased risk. Interpretation is again difficult, however, because the period of study was greater than 10 years after closure of the incinerator and it is not clear how many new cases would result from previous exposure (in particular when we consider migration). Information on the residence history of the cases would be particularly useful. In general, evidence of a persistent increased risk is consistent with both a long latency period and the existence of unmeasured risk factors that are unconnected with the incinerator and are responsible for the excess. Choosing between these scenarios is difficult and must be done on epidemiological as well as statistical grounds. We finally note that the level of sophistication of the analysis will be strongly influenced by the quality of the data.

[Received September 1997. Revised July 2000.]

REFERENCES

Alexander, F., and Cuzick, J. (1992), "Methods for the Assessment of Disease Clusters," in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, eds. P. Elliott, J. Cuzick, D. English, and R. Stern, New York: Oxford University Press, pp. 238–250.

Anderson, N., and Titterington, D. (1997), "Some Methods for Investigating Spatial Clustering with Epidemiological Applications," *Journal of the Royal Statistical Society*, Ser. A, 160, 87–105.

Besag, J., and Newell, J. (1991), "The Detection of Clusters in Rare Diseases," *Journal of the Royal Statistical Society*, Ser. A, 154, 143–155.

Besag, J., York, J., and Mollié, A. (1991), "Bayesian Image Restoration With Two Applications in Spatial Statistics," *Annals of the Institute of Statistics and Mathematics*, 43, 1–59.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A., and Conlon, E. M. (1999). "Bayesian Models for Spatially Correlated Disease and Exposure Data," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 131–156.

Best, N. G., Ickstadt, K., and Wolpert, R. L. (in press), "Spatial Poisson Regression for Health and Exposure Data Measured at Disparate Spatial Scales," *Journal of American Statistical Association*.

Best, N. G., and Wakefield, J. C. (1999), "Accounting for Inaccuracies in Population Counts and Case Registration in Cancer Mapping Studies," *Journal of the Royal Statistical Society*, Ser. A, 162, 363–382.

Bithell, J. (1990), "An Application of Density Estimation to Geographical Epidemiology," *Statistics in Medicine*, 9, 691–701.

——— (1992), "Statistical Methods for Analysing Point-source Exposures," in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, eds. P. Elliott, J. Cuzick, D. English, and R. Stern, New York: Oxford University Press, pp. 221–30.

——— (1996), Response to "Use of Deprivation Indices in Small Area Studies," by A. Lawson, *Journal of Epidemiology and Community Health*, 50, 690.

Bithell, J., and Stone, R. (1990), "On Statistical Methods for Analysing the Geographical Distribution of Cancer Cases Near Nuclear Installations," *Journal of Epidemiology and Community Health*, 43, 79–85.

Boyle, P., Walker, A., and Alexander, F. (1996), "Historical Aspects of Leukaemia Clusters," in *Methods for Investigating Localized Clustering of Disease*, eds. F. Alexander and P. Boyle, Lyon, France: International Agency for Research on Cancer, pp. 1–20.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Carstairs, V., and Morris, R. (1991), *Deprivation and Health in Scotland*, Aberdeen: Aberdeen University Press.

Clayton, D., and Kaldor, J. (1987), "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping," *Biometrics*, 43, 671–682.

Cook-Mozaffari, P., Darby, S., Doll, R., Forman, D., Hermon, C., Pike, M., and Vincent, T. (1989), "Geographical Variation in Mortality from Leukaemia and Other Cancers in England and Wales in Relation to Proximity to Nuclear Installations, 1969–78," *British Journal of Cancer*, 59, 476–485.

Cressie, N., and Chan, N. H. (1989), "Spatial Modelling of Regional Variables," *Journal of the American Statistical Association*, 84, 393–401.

Diggle, P. J. (1990), "A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point," *Journal of the Royal Statistical Society*, Ser. A, 153, 340–362.

Diggle, P. J., Elliott, P., Morris, S. E., and Shaddick, G. (1997), "Regression Modelling of Disease Risk in Relation to Point Sources," *Journal of the Royal Statistical Society,* Ser. A, 160, 491–505.

Diggle, P. J., Morris, S. E., and Wakefield, J. C. (2000), "Point Source Modeling Using Matched Case-Control Data," *Biostatistics*, 1, 89–105.

Diggle, P., and Rowlingson, B. (1994), "A Conditional Approach to Point Process Modeling of Raised Incidence," *Journal of the Royal Statistical Society,* Ser. A, 157, 433–440.

Dolk, H., Mertens, B., Kleinschmidt, I., Walls, P., Shaddick, G., and Elliott, P. (1995). "A Standardisation Approach to the Control of Socioeconomic Confounding in Small Area Studies of Environment and Health," *Journal of Epidemiology and Public Health*, Supplement, S9–S14.

Dolk, H., Elliott, P., Shaddick, G., Walls, P., and Thakrar, B. (1997a), "Cancer Incidence Near Radio and Television Transmitters in Great Britain: All High Power Transmitters," *American Journal of Epidemiology*, 145, 10–17.

Dolk, H., Shaddick, G., Walls, P., Grundy, C., Thakrar, B., Kleinschmidt, I., and Elliott, P. (1997b), "Cancer Incidence Near Radio and Television Transmitters in Great Britain: Sutton Coldfield Transmitter," *American Journal of Epidemiology*, 145, 1–9.

Dolk, H., Vrijheid, M., Armstrong, B., Abramsky, L., Bianche, F., Garne, E., Nelen, V., Robert, E., Scott, J. E. S., Stone, D., and Tenconi, R. (1998), "Risk of Congenital Anomalies Near Hazardous-Waste Landfill Sites in Europe: The EUROHAZCON Study," *Lancet*, 352, 423–427.

Dolk, H., Thakrar, B., Walls, P., Landon, M., Grundy, C., Suez-Lloret, I., Wilkinson, P., and Elliott, P. (1999), "Mortality Among Residents Near Cokeworks in Great Britain," *Occupational and Environmental Medicine*, 56, 34–40.

Elliott, P. (1996), "Small-Area Studies," in *Environmental Epidemiology: Exposure and Disease*, eds. R. Bertollini, M. Lebowitz, R. Saracci, and D. Savitz, Lewis Publishers, pp. 187–199.

Elliott, P., Hills, M., Beresford, J., Kleinschmidt, I., Jolley, D., Pattenden, S., Rodrigues, L., Westlake, A., and Rose, G. (1992a), "Incidence of Cancer of the Larynx and Lung Near Incinerators of Waste Solvents and Oils in Great Britain," *Lancet*, 339, 854–858.

Elliott, P., Westlake, A., Hills, M., Kleinschmidt, I., Rodrigues, L., McGale, P., Marshall, K., and Rose, G. (1992b), "The Small Area Health Statistics Unit: A National Facility for Investigating Health Around Point Sources of Environmental Pollution in the United Kingdom," *Journal of Epidemiology and Community Health*, 46, 345–349.

Elliott, P., Martuzzi, M., and Shaddick, G. (1995), "Spatial Statistical Methods in Environmental Epidemiology: A Critique," *Statistical Methods in Medical Research*, 4, 149–161.

Elliott, P., Shaddick, G., Kleinschmidt, I., Jolley, D., Walls, P., Beresford, J., and Grundy, C. (1996), "Cancer Incidence Near Municipal Solid Waste Incinerators in Great Britain," *British Journal of Cancer*, 73, 702–707.

Elliott, P., Eaton, N., Shaddick, G., and Carter, R. (2000), "Cancer Incidence Near Municipal Solid Waste Incinerators in Great Britain 2: Histopathological and Case-Notes Review of Primary Liver Cancer Cases," *British Journal of Cancer*, 82, 1103–1106.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Greenland, S. (1992), "Divergent Biases in Ecologic and Individual-Level Studies," *Statistics in Medicine*, 11, 1209–1223.

Greenland, S., and Morgenstern, H. (1989), "Ecological Bias, Confounding and Effect Modification," *International Journal of Epidemiology*, 18, 269–274.

Greenland, S., and Robins, J. (1994), "Ecological Studies: Biases, Misconceptions and Counterexamples," *American Journal Epidemiology*, 139, 747–760.

Jolley, D., Jarman, B., and Elliott, P. (1992), "Socio-economic Confounding," in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, eds. P. Elliott, J. Cuzick, D. English, and R. Stern, New York: Oxford University Press, pp. 115–124.

Kelsall, J. E., and Diggle, P. J. (1995a), "Kernel Estimation of Relative Risk," *Bernoulli*, 1, 3–16.

——— (1995b), "Nonparametric Estimation of Spatial Variation in Relative Risk," *Statistics in Medicine*, 14, 2335–2342.

Kelsall, J. E., and Wakefield, J. C. (1999), Discussion of "Bayesian Models for Spatially Correlated Disease and Exposure Data," by N. G. Best, R. A. Arnold, A. Thomas, L. A. Waller, and E. M. Conlon, in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York Oxford University Press, p. 151.

Kleinschmidt, I., Hills, M., and Elliott, P. (1995), "Smoking Behaviour Can Be Predicted by Neighbourhood Deprivation Measures," *Journal of Epidemiology and Community Health*, 49(Suppl. 2), S72–S77.

Lawson, A. (1993), "On the Analysis of Mortality Events Associated with a Prespecified Fixed Point," *Journal of the Royal Statistical Society,* Ser. A, 156, 363–77.

——— (1996), "Use of Deprivation Indices in Small Area Studies" (letter), *Journal of Epidemiology and Community Health*, 50, 689.

Lawson, A., and Williams, F. (1993), "Applications of Extraction Mapping in Environmental Epidemiology," *Statistics in Medicine*, 12, 1249–1258.

Mollié, A. (1996), "Bayesian Mapping of Disease," in *Markov Chain Monte Carlo in Practice*, eds. W. Gilks, S. Richardson, and D. Spiegelhalter, London: Chapman and Hall.

Mollié, A., and Richardson, S. (1991), "Empirical Bayes Estimates of Cancer Mortality Rates Using Spatial Models," *Statistics in Medicine*, 10, 95–112.

Morris, S. E., and Wakefield, J. C. (2000), "Assessment of Disease Risk in Relation to a Pre-specified Source," in *Spatial Epidemiology: Methods and Applications*, eds. P. Elliott, J. Wakefield, N. Best, and D. Briggs, New York: Oxford University Press, pp. 153–184.

Mugglin, A. S. Carlin, B. P., and Gelfand, A. E. (in press), "Fully Model-Based Approaches for Spatially Misaligned Data," *Journal of the American Statistical Association.*

Nomura, A. (1997), "Stomach Cancer," in *Cancer Epidemiology and Prevention (2nd ed.)*, eds. D. Schottenfeld and J. F. Fraumeni, New York: Oxford University Press, pp. 707–724.

Richardson, S. (1992), "Statistical Methods for Geographical Correlation Studies," in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, eds. P. Elliott, J. Cuzick, D. English, and R. Stern, New York: Oxford University Press, pp. 181–204.

Richardson, S., and Gilks, W. R. (1993), "Conditional Independence Models for Epidemiological Studies with Covariate Measurement Error," *Statistics in Medicine*, 12, 1703–1722.

Rothman, K., and Greenland, S. (1998), *Modern Epidemiology (2nd ed.)*, Lippincott-Raven, Philadelphia:

Sans, S., Elliott, P., Kleinschmidt, I., Shaddick, G., Pattenden, S., Walls, P., Grundy, C., and Dolk, H. (1995), "Cancer Incidence and Mortality Near the Baglan Bay Petrochemical Works, South Wales," *Occupational and Environmental Medicine*, 52, 217–224.

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1998), "WinBUGS User Manual," version 1.1.1, Cambridge, UK (available at http://www.mrc-bsu.cam.ac.uk/bugs+).

Stone, R. (1988), "Investigations of Excess Environmental Risks around Putative Source: Statistical Problems and a Proposed Test," *Statistics in Medicine*, 7, 649–660.

Wakefield, J. C., and Elliott, P. (1999), "Issues in the Statistical Analysis of Small Area Health Data," *Statistics in Medicine*, 18, 2377–2399.

Wilkinson, P., Thakrar, B., Shaddick, G., Stevenson, S., Pattenden, S., Landon, M., Grundy, C., and Elliott, P. (1997), "Cancer Incidence around the Pan Britannica Industries Pesticide Factory, Waltham Abbey," *Occupational and Environmental Medicine*, 54, 101–107.

Wilkinson, P., Thakrar, B., Walls, P., Landon, M., Falconer, S., Grundy, C., and Elliott, P. (1999), "Lymphohaematopoietic Malignancy Around All Industrial Complexes That Include Major Oil Refineries in Great Britain," *Occupational and Environmental Medicine*, 56, 577–580.