Appl. Statist. (1999)
48, *Part* 2, *pp.* 253–268

Multilevel modelling of the geographical distributions of diseases

Ian H. Langford,

University of East Anglia, Norwich, University College London, and Institute of Education, London, UK

Alastair H. Leyland

University of Glasgow, and Institute of Education, London, UK

and Jon Rasbash and Harvey Goldstein

Institute of Education, London, UK

[Received February 1997. Final revision June 1998]

Summary. Multilevel modelling is used on problems arising from the analysis of spatially distributed health data. We use three applications to demonstrate the use of multilevel modelling in this area. The first concerns small area all-cause mortality rates from Glasgow where spatial autocorrelation between residuals is examined. The second analysis is of prostate cancer cases in Scottish counties where we use a range of models to examine whether the incidence is higher in more rural areas. The third develops a multiple-cause model in which deaths from cancer and cardiovascular disease in Glasgow are examined simultaneously in a spatial model. We discuss some of the issues surrounding the use of complex spatial models and the potential for future developments.

Keywords: Cancer epidemiology; Geographical epidemiology; Multilevel modelling; Random coefficient models; Spatial analysis

1. Introduction

Geographical epidemiologists are increasingly using complex methods of statistical analysis to investigate the distribution of diseases, such as cancers, by using data which are aggregated into small areas such as postcode sectors (Elliott *et al.*, 1992, 1995). The analysis of such geographically distributed disease data tends to fall into one of two broad categories which reflect different motivations and goals.

The first category, exploratory analysis, produces maps of the distribution of disease to provide health researchers with a visual display which can suggest, via patterns and spatial trends, useful avenues of research into causal processes. Atlases of such maps attempt to reflect the distribution of a range of diseases over a large geographical area (e.g. Kemp *et al.* (1985) and Statistics Canada (1991)). However, the use of the relative risk of a disease, i.e. the number of observed cases divided by the number of expected cases for each area, for this purpose may lead to problems for areas with small populations. Such areas, usually in rural locations, tend to have extreme relative risks as the number of expected cases in the

Address for correspondence: Ian H. Langford, Centre for Social and Economic Research on the Global Environment, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK.
E-mail: i.langford@uea.ac.uk

0035–9254/99/48253

denominator is low. Conversely, if we map probability values for excesses or deficits of cases rather than relative risks, these tend to occur in areas with large populations, usually in urban areas, as the probability value is related to the sample size (see Clayton and Kaldor (1987) and Langford (1994)). Here, we try to achieve a compromise by relating the relative risk in each area to the global distribution of relative risks for all the areas in our sample, or the local distribution of relative risks in areas that are geographically close to each other. The example of Section 3.1 shows how this approach can be implemented as part of a multilevel modelling analysis using data on all-cause mortality in Greater Glasgow for 1993. The data are aggregated into postcode sectors which are relatively small areas with differing sizes of population at risk, so the use of smoothed relative risks is important to account for the issues discussed above. Spatial smoothing is also applied, where areas which are adjacent to each other are assumed to have more similar relative risks. Section 3.3 uses the flexibility of the multilevel model to develop a multivariate spatial analysis, where deaths from two different causes, cancer and circulatory diseases, are modelled together. We discuss in Sections 2 and 4 how residuals may be extracted to provide information for mapping the distributions of these diseases.

The second type of analysis is inferential analysis, in which a number of explanatory variables, some of which may have a spatial component, are used to explain variation in a particular disease of interest. The emphasis here is on the testing of specific hypotheses, or prior beliefs, about the distribution of the disease and associated, potentially causal, factors (Langford, 1995; Langford and Bentham, 1996). Accounting for spatial correlation between areas enables more reliable inferences to be made. In Section 3.2, using data collected on prostate cancer incidence in the 56 old local government districts of Scotland, we demonstrate that choosing between different spatial models is not always straightforward. The aim of this analysis is to investigate whether more rural districts, defined as having higher proportions of the male workforce employed in agriculture, forestry and fishing, have a higher incidence of prostate cancer.

In this paper, data are investigated which consist of observed and expected counts of cases of disease in discrete spatial units. Hence, for a population of geographical areas, a number of cases occur within a specified population at risk in each area. Whether we are embarking on an exploratory or inferential analysis, it is useful to break down the likely influences on the distribution of a disease into three separate categories:

(a) within-area effects, such as population at risk, or individual characteristics;
(b) hierarchical effects, arising from small areas being grouped into larger areas, for administrative, cultural or geographical reasons, e.g. a number of local authority districts may be grouped into health boards which have different methods of treatment or classification of a disease;
(c) neighbourhood effects — areas which are close to each other in geographical space may share common environmental, social or demographic factors influencing the incidence or outcome of disease. Also, as areas are usually formed by using geopolitical boundaries which are unrelated to the disease of interest, we may wish to use spatial smoothing of the distribution of relative risks to remove any artefactual variation brought into the data by the method of data aggregation.

The use of empirical Bayes and fully Bayesian techniques allows the distribution of a disease to be affected by various models of spatial and environmental processes which arise from different underlying beliefs about aetiology (Bernardinelli *et al.*, 1995; Bernardinelli and Montomoli, 1992; Cisaghli *et al.*, 1995; Clayton and Bernardinelli, 1992; Clayton and Kaldor, 1987; Langford *et al.*, 1998; Langford, 1994, 1995; Lawson, 1994; Lawson and Williams, 1994;

Mollié and Richardson, 1991; Schlattmann and Böhning, 1993). Two main statistical methodologies have been used to model geographically distributed health data in this way. The first is Markov chain Monte Carlo (MCMC) methods using Gibbs sampling (Gilks *et al.*, 1993) often implemented via the BUGS software (Spiegelhalter *et al.*, 1995). The second is multilevel modelling techniques based on iterative generalized least squares (IGLS) procedures (Goldstein, 1995) which are the focus of this paper. These two kinds of method can be described as using the Bayesian and empirical Bayesian models respectively, and we discuss the differences between the two approaches. In the following section, we detail the methodology and computational algorithms that are necessary to model the three types of effect described above within the IGLS framework. Section 3 presents the three examples of analyses of geographical health data mentioned previously. The models were all fitted using the multilevel modelling software MLn (Rasbash and Woodhouse, 1995). The discussion focuses on issues surrounding both the theory and the methodology of building complex spatial models and how these models should be interpreted, and provides pointers for future research.

2. Methods

2.1. The linear random coefficients model

The basic model of fixed and random effects described by Goldstein (1995) and Breslow and Clayton (1993) is

$$Y = X\beta + Z\theta \tag{1}$$

with a vector of observations Y being modelled by explanatory variables X and associated fixed parameters β, and explanatory variables Z with random coefficients θ. (θ represents all the random coefficients for Y and hence contains residual terms in the model.) The fixed and random design matrices X and Z will not, in general, be the same. Goldstein (1995), pages 38–41, described a two-stage process for estimating the fixed and random parameters (the variances and covariances of the random coefficients) in successive iterations using IGLS. A summary of this process follows. We estimate the fixed parameters in an initial ordinary least squares regression. From the vector of residuals from this model we can construct initial values for V, the dispersion matrix for Y. Then, we iterate the following procedure, first estimating fixed parameters in a generalized least squares regression as

$$\hat{\beta} = (X^{\mathrm{T}} V^{-1} X)^{-1} X^{\mathrm{T}} V^{-1} Y \tag{2}$$

and again calculating residuals $\tilde{Y} = Y - X\hat{\beta}$. We now form the matrix product of these residuals and stack them into a vector $Y^* = \mathrm{vec}(\tilde{Y}\tilde{Y}^{\mathrm{T}})$. By doing so we can estimate the variance of the random coefficients θ, $\gamma = \mathrm{cov}(\theta)$, as

$$\hat{\gamma} = (Z^{*\mathrm{T}} V^{*-1} Z^*)^{-1} Z^{*\mathrm{T}} V^{*-1} Y^*, \tag{3}$$

where V^* is the Kronecker product of V, namely $V^* = V \otimes V$, noting that $V = E(\tilde{Y}\tilde{Y}^{\mathrm{T}})$, and Z^* is the appropriate design matrix for the random coefficients. Assuming multivariate normality, the estimated covariance matrix for the fixed parameters is

$$\mathrm{cov}(\hat{\beta}) = (X^{\mathrm{T}} V^{-1} X)^{-1} \tag{4}$$

and, for the random parameters, Goldstein and Rasbash (1992) showed that

$$\mathrm{cov}(\hat{\gamma}) = 2(Z^{*\mathrm{T}} V^{*-1} Z^*)^{-1}. \tag{5}$$

2.2. The Poisson model

Consider a population of areas with the ith area having O_i observed cases and E_i expected cases, where E_i may be calculated from the incidence in the population N_i for each area as

$$E_i = N_i \sum O_i / \sum N_i, \tag{6}$$

and E_i may be divided also into different age and sex bands. We can write a basic Poisson model with heterogeneity effects as

$$O_i \sim \text{Poisson}(\mu_i) \qquad \text{with } \log(\mu_i) = \log(E_i) + \alpha + x_i\beta + u_i \tag{7}$$

where $\log(E_i)$ is treated as an offset, α is a constant and x_i is an explanatory variable with coefficient β. This model may be generalized to include any number of explanatory variables. We take account of the distribution of cases *within* each area by assuming that the number of cases has a Poisson distribution. In contrast, the u_i represent heterogeneity effects between areas (Clayton and Kaldor, 1987; Langford, 1994), which may be viewed as constituting extra-Poisson variation caused by the variation among underlying populations at risk in the areas considered. However, we also want to take account of the fact that the relative risks may be spatially autocorrelated. One way of doing this is to treat the model as a 'multiple-membership' model (Goldstein, 1995; Goldstein *et al.*, 1998), where each area is a member of a higher level unit which contains its nearest neighbours, e.g. those areas with which it shares a common boundary. We can write this model as

$$\begin{aligned} &O_i \sim \text{Poisson}(\mu_i), \\ \log(\mu_i) = \log(E_i) &+ \alpha + x_i\beta + \sum_j u_i \surd w_{ij}, \end{aligned} \tag{8}$$

where w_{ij} are weights, and $w_{ij} = 0$ if district j is not adjacent to district i. If a district had three neighbours, then we could construct weights such that $w_{ii} = 0.5$, and for the adjacent districts $w_{i1} = w_{i2} = w_{i3} = 0.167$, so that $\Sigma_j\, w_{ij} = 1$. However, this formulation does not allow us to examine heterogeneity and spatial effects independently, so we have used a model developed for the distribution of relative risks of a disease by Besag *et al.* (1991). This can be written as

$$O_i \sim \text{Poisson}(\mu_i) \qquad \text{with } \log(\mu_i) = \log(E_i) + \alpha + x_i\beta + u_i + v_i. \tag{9}$$

The v_i are spatially dependent random effects and may have any one of a number of structures describing adjacency or nearness in space. However, before discussing the structure of these spatial effects, we must first account for the fact that we have a non-linear (logarithmic) relationship between the outcome variable and the predictor part of the model. There are two options.

(a) If the number of cases in each area is sufficiently large, say $O_i > 10$, then it may be reasonable to model the logarithm of the relative risks directly (Clayton and Hills, 1993), assuming that these follow a normal distribution. Heterogeneity effects can then be accommodated by weighting the random part of the model by some function of the population at risk in each area.

(b) When the normal distribution approximation is inappropriate, we can make a linearizing approximation to estimate the random parameters, i.e. the residuals $\hat{u}_i$ and $\hat{v}_i$ from the model, by using penalized quasi-likelihood (PQL) estimation with a second-order Taylor series approximation (Breslow and Clayton, 1993; Goldstein, 1995; Goldstein and Rasbash, 1996). We write model (7) as

$$\mu_i = E_i\, f(H) \qquad \text{with } H = \alpha + x_i\beta + u_i + v_i.$$

Then we linearize $f(H)$ by writing H_t for the value of the linear predictor H at iteration t. We now express $f(H_{t+1})$ as a function of $f(H_t)$ via a second-order Taylor expansion about current fixed and random part estimates, so that

$$\begin{aligned} f(H_{t+1}) &= f(H_t) + (\alpha_{t+1} - \hat{\alpha}_t) + x_i(\beta_{t+1} - \hat{\beta}_t) f'(H_t) + (u_{t+1,i} - \hat{u}_{t,i}) f'(H_t) \\ &\quad + (v_{t+1,i} - \hat{v}_{t,i}) f'(H_t) + (u_{t+1,i} - \hat{u}_{t,i})^2 f''(H_t)/2 + (v_{t+1,i} - \hat{v}_{t,i})^2 f''(H_t)/2, \end{aligned} \tag{10}$$

where the first three terms on the right-hand side of equation (10) provide the updating function for the fixed part of the model and the last four for the random part (see Goldstein (1995), section 5.1). For the Poisson distribution

$$f(H) = f'(H) = f''(H) = \exp(X_i\hat{\beta}_t + \hat{u}_i). \tag{11}$$

Hence, at each iteration we estimate about the fixed part of the model plus the residuals, u_i. A full description of this linearizing procedure can be found in Goldstein (1995) and Goldstein and Rasbash (1996). The procedure can lead to problems with convergence, or with the model failing due to arithmetic overflow when some of the residuals are particularly large. In these situations, the second-order term in equation (10) can be omitted, or, in extreme circumstances, estimates can be based on the fixed part of the model only. This latter estimation method is called marginal quasi-likelihood (MQL) (Breslow and Clayton, 1993; Goldstein, 1995) but may lead to biased parameter estimates. However, bootstrap procedures can potentially be used to correct for these biases (Goldstein, 1996; Kuk, 1995).

2.3. Estimating spatial effects in a multilevel model

Several possibilities for specifying the structure of the random effects in the model are available (see, for example, Besag *et al.* (1991) and Bailey and Gatrell (1995)). These models assume two components: a random effects or 'heterogeneity' term and a term representing the spatial contribution of neighbouring areas as in model (9) with intrinsic Gaussian distributions for each type of effect.

We adopt a different approach, which allows a more direct interpretation of the model parameters and can be fitted in a computationally efficient manner within a multilevel model. For the heterogeneity effects, this is not a problem, because we simply have a variance–covariance matrix with 1s or other specified values on the diagonal, and the model is analogous to fitting an iteratively weighted least squares model (McCullagh and Nelder, 1989). However, fitting spatial effects is more complex, as we are required to find off-diagonal terms in the variance–covariance matrix. This can be achieved through a careful consideration of the structure of the spatial part of the model. Our formulation of the spatial model is to consider each spatial effect v_i to be the weighted sum of a set of independent random effects v_j^* such that

$$v_i = \sum_{j \neq i} z_{ij} v_j^*. \tag{12}$$

The v_j^* can be considered to be the effect of area j on other areas, moderated by a measure of proximity z_{ij} of each pair of areas. The v_j^*, which are the residuals, can be estimated directly from the model because of their independence.

Returning to the matrix notation used in equation (1), we can rewrite equation (9) as

$$\log(\mu_i) = \pi_i = \{\log(E_i) \;\; 1 \;\; x_i\} \begin{pmatrix} 1 \\ \alpha \\ \beta \end{pmatrix} + (Z_u \;\; Z_v^*) \begin{pmatrix} \theta_u \\ \theta_v^* \end{pmatrix}, \tag{13}$$

where Z_u is the identity matrix and $Z_v^* = \{z_{ij}\}$, with Z and θ from equation (1) partitioned into heterogeneity effects and spatial effects to give $(Z_u \;\; Z_v^*)$ and $\left(\begin{smallmatrix}\theta_u \\ \theta_v^*\end{smallmatrix}\right)$ respectively. With a variance structure such as

$$\operatorname{var}\left\{\begin{pmatrix} \theta_u \\ \theta_v^* \end{pmatrix}\right\} = \begin{pmatrix} \sigma_u^2 I & \sigma_{uv} I \\ \sigma_{uv} I & \sigma_v^2 I \end{pmatrix}, \tag{14}$$

which is equivalent to

$$\operatorname{var}\left\{\begin{pmatrix} u_i \\ v_i^* \end{pmatrix}\right\} = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix},$$

the overall variance from equation (1), conditional on the fixed parameters, is given by

$$\operatorname{var}(\pi|X\beta) = Z\Sigma_\theta Z^{\mathrm{T}}, \tag{15}$$

where Σ_θ is the variance–covariance matrix of the random terms in θ. The structure of Σ_θ will often lead to simplifications. For example, in a random effects model when $\theta = \{u_i\}$ and $\operatorname{var}(u_i) = \sigma_u^2$, $\operatorname{cov}(u_i,\; u_j) = 0$; then $\Sigma_\theta = \sigma_u^2 I$ and so $\operatorname{var}(\pi|X\beta) = \sigma_u^2 ZZ^{\mathrm{T}}$. Similarly, in the spatial model defined by the partitions in θ and Z given by equation (13) and the variance structure of equation (12),

$$\operatorname{var}(\pi|X\beta) = \sigma_u^2 Z_u Z_u^{\mathrm{T}} + \sigma_{uv}(Z_u Z_v^{*\mathrm{T}} + Z_v^* Z_u^{\mathrm{T}}) + \sigma_v^2 Z_v^* Z_v^{*\mathrm{T}}. \tag{16}$$

There are many ways in which the z_{ij} can be formulated; in general we can write

$$z_{ij} = w_{ij}/w_{i+} \tag{17}$$

where the $w_{ii} = 0$. Common choices for the w_{i+} would be $w_{i+} = (\Sigma_{j\neq i}\, w_{ij})^{0.5}$, which ensures that the variance contribution is the same for all areas, or $w_{i+} = \Sigma_{j\neq i}\, w_{ij}$ indicating that the variance of an area decreases as the information about that area (e.g. in terms of the number of neighbours in an adjacency model) increases.

The simplest form of adjacency matrix is such that $w_{ij} = 1$ if areas i and j share a common boundary and $w_{ij} = 0$ otherwise, although other formulations are possible such as the use of distance decay functions (Bailey and Gatrell, 1995). The choice of such functions is largely user dependent and should ideally be based on some prior hypothesis about the data. Here we have used adjacency matrices and have also considered a simple exponential decay model where we define the w_{ij} as

$$w_{ij} = \exp(-\lambda d_{ij}) \tag{18}$$

with d_{ij} the Euclidean distances between the centroids of areas i and j, and λ a constant to be estimated from the data. The estimation of λ is problematical, as it is non-linear in the random part of the model. Goldstein *et al.* (1994) showed that maximum likelihood estimates can be obtained by using a Taylor series expansion for the normal distribution model. However, estimation becomes more complicated for a Poisson model; an alternative is to fit a series of models with differing values of λ_k, and to determine the residual deviance D_k from each model. We can then regress the deviance against the distance decay parameter so that

$$D_k = a + b\lambda_k + c\lambda_k^2 + e_k \tag{19}$$

where e_k is an error term. Differentiating, the approximate solution is $\lambda = -b/2c$. Successive approximations then converge towards the best estimate.

Finally, the random effects for heterogeneity and spatial effects must be specified within a generalized linear modelling framework, in this case using IGLS estimation within the MLn software. There are two approaches for fitting the random effects within MLn which demonstrate some more general issues for spatial modelling.

First, a suitable set of explanatory variables may be defined with random coefficients. For example, for the spatial part of the model, we may define a set of variables $z_{v1}, z_{v2}, \ldots, z_{vn}$ whose values form the columns of Z_v^*, with $z_{vj} = \{z_{ij}\}$. A similar set of variables can be defined for the heterogeneity effects, and a covariance term can be fitted between the two sets of effects. However, a problem arises because we only wish to estimate a single variance parameter for *all areas* for heterogeneity effects, a single variance parameter for spatial effects and a single covariance term. Hence, the parameter estimates for each area need to be constrained to be the same for each set of effects, e.g. σ_v^2 constrained to be the same for all the z_vs. These complex constraints are introduced into the model via a set of linear equations. A discussion of this procedure is given by Goldstein (1995), pages 57–58; it requires the inclusion of a large number of explanatory variables in the model — far more than the number of data points — and a large number of constraint vectors. These add to the complexity of the model, the computational time required and the stability of the model in terms of convergence properties. However, the calculation of residuals from the model is straightforward, as these can be estimated for each of the random explanatory variables. This is important when the focus of our investigation is the comparison of relative risks between the areas in the data set. It is less important when only the global parameters are of interest, such as σ_u^2 and σ_v^2 which describe respectively the size of the overall heterogeneity and spatial effects in the model.

An alternative approach is to build the weights matrices associated with the random effects and to fit these directly into the model. The variance of the data conditional on the fixed part of the model, as given in equation (16), is formed from three matrices: $Z_u Z_u^{\mathrm{T}}$, $Z_u Z_v^{*\mathrm{T}} + Z_v^* Z_u^{\mathrm{T}}$ and $Z_v^* Z_v^{*\mathrm{T}}$. Expressing the model in terms of these design matrices overcomes the need to place multiple equality constraints on the random parameters. This method is generalizable to the non-linear model of equation (9). A PQL estimation procedure requires the estimation of the residuals and their associated variances at each iteration. The estimation of the residuals is described in Appendix A.

3. Applications

In this section, we give three examples of results from health data sets which raise particular methodological issues addressed in the discussion and show how substantive interpretations can be made of spatial multilevel models. The data sets are available as MLn worksheets at

`http://www.blackwellpublishers.co.uk/rss/`

3.1. Greater Glasgow Health Board mortality data

The data for this example are deaths from all causes in 143 postcode sectors within Greater

Table 1. Parameter estimates and standard errors for the Glasgow Health Board mortality data

Parameter	*Estimate*	*Standard error*
α	4.2047	0.0374
σ_u^2	0.0174	0.0054
σ_{uv}	0.0300	0.0085
σ_v^2	0.0865	0.0334

Glasgow Health Board in 1993 obtained from the Registrar General for Scotland. Hence, as postcode sectors are fairly small (average population about 6500), and the data are only for 1 year, we formulate the model in a similar manner to that implied by equations (7) and (10):

$$\begin{aligned} O_i &\sim \text{Poisson}(\mu_i), \\ \log(\mu_i) &= \log(E_i) + \alpha + u_i + \sum_{j\neq i} z_{ij} v_j^*, \end{aligned} \tag{20}$$

where the E_i are age and sex standardized for the Greater Glasgow Health Board area. For a *first-order autocorrelation model* (Bailey and Gatrell, 1995) we define $z_{ij} = 1/n_i$, if area j is a neighbour of area i and $z_{ij} = 0$ otherwise, with area i having a total of n_i neighbours. The u_i are the random effects for each area; the v_i^*, by contrast, are the effects of each area on its neighbours with the summation term $\Sigma_{j\neq i} z_{ij} v_j^*$ giving the spatial effect for area i. We can specify a joint distribution for the u_i and v_i^* to model a correlation between the random effect of an area and its effect on its neighbours as in equation (14):

$$\begin{pmatrix} u_i \\ v_i^* \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} \right\}. \tag{21}$$

This may then be expressed in the terms of equations (1) and (13) by writing

$$\left. \begin{aligned} X &= \{\log(E_i)\ \ \mathbf{1}\}, \\ \beta &= \begin{pmatrix} 1 \\ \alpha \end{pmatrix}, \\ z_u &= I, \\ Z_v^* &= \{z_{ij}\} \end{aligned} \right\} \tag{22}$$

where **1** is the unit vector. Estimation may proceed as described in equation (16) and Appendix A. The parameter estimates for this model are shown in Table 1. To aid convergence, the $\log(E_i)$ were centred around 0, and hence $\alpha \neq 0$ even though the relative risks have a mean of 1.

The spatial variance and covariance terms are highly significant with a χ^2-value of 13.44 with 2 degrees of freedom ($p = 0.001$). The correlation between the random effects and spatial effects is 0.774, indicating that the neighbours of an area with high mortality also tend to have high mortality. The total estimated variance for an area is dependent on its number of neighbours and is given by $\sigma_u^2 + \sigma_v^2/n_i$. The mean number of neighbours for a postcode sector within Greater Glasgow Health Board is 5.4; this implies a total variance of 0.0333, of which 49.0% arise from the spatial effects. The estimated covariance between any two areas depends on

(a) whether the two areas border each other and
(b) the number of common neighbours.

In terms of the z_{ij} used in equation (20) the covariance between areas i and j can be expressed as

$$(z_{ij}+z_{ji})\sigma_{uv}+\sum_{k\neq i,j} z_{ik}z_{jk}\sigma_v^2.$$

In this example, the fitted model shows that there are significant parameters for both heterogeneity and spatial autocorrelation (using a Wald test with significance level $\alpha = 0.05$). This has a sensible interpretation, as postcode sectors are quite variable in population size, and this effect is summarized by σ_u^2, the mean variance between areas. However, the spatial effects parameter σ_v^2 is larger (although it needs to be scaled by the number of adjacent areas for comparison with σ_u^2 in each area). This may be because mortality rates are similar in social areas that are larger than the postcode sectors analysed here. A further analysis could place larger units such as social neighbourhoods at a higher level in the model to test for their effect, and covariates such as social and housing status could be included. The significant covariance between the heterogeneity and spatial effects parameters occurs because areas whose populations have similar sociodemographic characteristics (and also large populations) tend to cluster and also have similar mortality rates.

3.2. Prostate cancer incidence in Scottish districts

In this example, we examine data covering 6 years, from 1975 to 1980, on the incidence of prostate cancer in 56 districts in Scotland (Kemp *et al.*, 1985). As the data were collected in larger geographical units and for a longer time period than the first example, the numbers of cases in each district are sufficiently large (between 10 and 627 cases) for us to model the relative risks of disease incidence (based on crude rates) and to assume that log(relative risk) follows an approximately normal distribution. Here, we wish to investigate the hypothesis that the relative risk of prostate cancer is higher in rural than in urban areas, as previous research has indicated an association between agricultural employment and incidence of prostate cancer (Key, 1995). A variable which is the percentage of the male workforce employed in agriculture, fishing and forestry industries is used as a surrogate measure of the rurality of an area. However, we must not only look at the incidence of prostate cancer within districts but also account for a potential artefactual effect caused by differential diagnosis rates between health board areas in Scotland. Hence, spatial effects caused by different processes at two different scales need to be modelled, namely

(a) a spatial autocorrelation model at district scale, which accounts for the possibility that areas closer in geographical space have similar incidences of prostate cancer and
(b) a variance components model at health board scale, which allows for the possibility that different health boards have different relative risks of prostate cancer, because diagnostic criteria are potentially variable.

Hence, equations (1) and (11) can be extended so that

$$Y = X\beta + (Z_u \;\; Z_v^*)\begin{pmatrix}\theta_u\\ \theta_v^*\end{pmatrix} + Z_{\text{hb}}\theta_{\text{hb}}, \tag{23}$$

where Z_{hb} and θ_{hb} are a design matrix and parameters for health board level random effects.

Table 2. Parameter estimates and standard errors for the prostate cancer models

	Estimates and standard errors for the following models:							
	Model A: simple model		*Model B: spatial effects*		*Model C: health board effects*		*Model D: both effects*	
	Estimate	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
Fixed part								
Intercept	0.4312	0.8793	-0.6199	0.7542	0.3011	0.8180	-0.1981	0.7642
SC12	0.0046	0.0052	0.0045	0.0044	-0.0022	0.0038	0.0006	0.0041
UVBI	-0.0710	0.0952	0.0896	0.0787	-0.0353	0.0904	0.0537	0.0830
AGRI	0.0219	0.0083	0.0084	0.0066	0.0119	0.0070	0.0068	0.0064
Random part								
σ^2_{hb}					0.0406	0.0196	0.0116	0.0094
σ^2_u	0.1338	0.0253	0.0474	0.0160	0.0588	0.0129	0.0447	0.0176
σ_{uv}			0.0053	0.0048			0.0056	0.0055
σ^2_v			0.0036	0.0017			0.0031	0.0018
λ			3.15				2.93	
Residual deviance	35.00		4.32		12.50		1.30	

Here, we have three explanatory variables in the fixed part of the model ($X\beta$), in addition to the intercept term, namely the proportion of the population in higher social classes (SC12); the estimated biologically active incidence of ultraviolet light at the earth's surface (UVBI) and the percentage of males employed in agriculture, fishing and forestry (AGRI). Social class and exposure to ultraviolet light were included as these have been postulated as risk factors for prostate cancer. In this model, Z^*_v is calculated using distances between district centroids, and a distance decay parameter λ is estimated from the spatial linkage described in equation (18); Z_{hb} is a vector of 1s which allows for a variance component for each health board to be estimated, and hence we can measure the variance at this scale, σ^2_{hb}. Table 2 presents the results of fitting the model given in equation (23) to the data with $Z_u = n^{-0.5}$, where n is the vector of population sizes for the districts in the study area, so that the districts are weighted by their population size in the random part of the model. Parameter estimates and standard errors are shown for four models: a simple, single-level model (A) with no spatial effects; a model with district scale spatial effects, but no health board effects (B); a model with only health board effects (C); a model with both district and health board effects (D) as given in equation (23).

The results for the simple model A seem to indicate a strong and significant effect of rurality, as measured by percentage of males in agricultural employment (AGRI). However, this is weakened by fitting a spatial autocorrelation parameter in model B, suggesting that the effect of AGRI may be because adjacent areas have similar mortalities. The change in deviance between the two models is 31.68 on 2 degrees of freedom ($p < 0.001$: with a covariance parameter fitted as well as a variance term). The third model (C), using health boards as a level with no spatial autocorrelation between districts, shows how ignoring autocorrelation between residuals at a lower level of a multilevel model (districts) could lead to misleading results at higher levels (health boards), as the parameter for the variance between health boards is statistically significant at $p < 0.05$, but the deviance statistic suggests that the model is not as good a fit to the data as model B is. Unexplained random variation at district level can appear spuriously at health board level, and model D, with both health board effects and spatial effects between districts, suggests that this may be happening in this example. The parameter estimate for AGRI becomes statistically insignificant in models B, C and D. Hence, misspecification of the random part of a model can noticeably affect the fixed as well as the random parameters. Further work needs to be done on the analysis of residuals in these complex models: Langford and Lewis (1998) details some procedures for the general analysis of outliers in multilevel models.

3.3. Multivariate spatial analysis of mortality in Greater Glasgow Health Board postcodes

In the example of Section 3.2, the scale of spatial analysis was extended to include health board as well as district level effects. We can further extend the methods to more than one disease within the same model. For example, we can look at multiple causes of death from the Greater Glasgow Health Board postcode mortality data and assess the degree to which different causes of death are related. In addition, we can examine the possibility of a spatial element to the distribution of each cause and assess whether these spatial elements are related for different causes. For example, if we take deaths from cancer (denoted by 'P') and deaths from circulatory diseases (indexed as 'Q') we can write the model

$$\begin{pmatrix} O_{P,i} \\ O_{Q,i} \end{pmatrix} \sim \text{Poisson}\begin{pmatrix} \lambda_{P,i} \\ \lambda_{Q,i} \end{pmatrix} \tag{24}$$

where

$$\log\left\{\begin{pmatrix} \lambda_{P,i} \\ \lambda_{Q,i} \end{pmatrix}\right\} = \log\left\{\begin{pmatrix} E_{P,i} \\ E_{Q,i} \end{pmatrix}\right\} + \begin{pmatrix} \alpha_P \\ \alpha_Q \end{pmatrix} + \begin{pmatrix} u_{P,i} \\ u_{Q,i} \end{pmatrix} + \begin{pmatrix} \sum\limits_{j\neq i} z_{ij} v^*_{P,j} \\ \sum\limits_{j\neq i} z_{ij} v^*_{Q,j} \end{pmatrix}. \tag{25}$$

This gives a possible 16 random parameters to be estimated. However, we do not estimate all 16 because of the difficulty in interpreting some of the parameters. Specifically, the covariance between the spatial parts of the two causes $\sigma_{v,P,Q}$ and between the random effect of one cause and the spatial part of the other cause $\sigma_{uv,P,Q}$ and $\sigma_{uv,Q,P}$ have all been set to 0. Hence, we estimate

$$\begin{pmatrix} u_{P,i} \\ u_{Q,i} \\ v_{P,i} \\ v_{Q,i} \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u,P} & \sigma_{u,P,Q} & \sigma_{uv,P} & 0 \\ \sigma_{u,P,Q} & \sigma^2_{u,Q} & 0 & \sigma_{uv,Q} \\ \sigma_{uv,P} & 0 & \sigma^2_{v,P} & 0 \\ 0 & \sigma_{uv,Q} & 0 & \sigma^2_{v,Q} \end{pmatrix}\right\}. \tag{26}$$

The results for this model are given in Table 3. As can be seen, considering only two causes of death and including no covariates in the model still lead to the estimation of 10 parameters to account for heterogeneity and spatial effects. However, both the heterogeneity and the spatial effects for the circulatory diseases are greater than those for cancers, suggesting a greater variability between areas for circulatory diseases, and more spatial clustering of mortality rates in adjacent postcode areas, although it must be borne in mind that some of the standard errors of the parameter estimates are large. We are currently investigating the effect of entering a covariate measuring deprivation into the model. Computationally, there were

Table 3. Parameter estimates and standard errors for the Glasgow Health Board mortality data for cancer and circulatory deaths

Parameter	*Estimate*	*Standard error*
α_P	2.8187	0.0310
α_Q	3.3978	0.0367
$\sigma^2_{u,P}$	0.0032	0.0073
$\sigma_{u,P,Q}$	0.0021	0.0037
$\sigma^2_{u,Q}$	0.0077	0.0065
$\sigma_{uv,P}$	0.0472	0†
$\sigma_{uv,Q,P}$	0	0
$\sigma^2_{v,P}$	0.0300	0.0367
$\sigma_{uv,P,Q}$	0	0
$\sigma_{uv,Q}$	0.0300	0.0177
$\sigma_{v,P,Q}$	0	0
$\sigma^2_{v,Q}$	0.1044	0.0421

†This parameter has been constrained so that the correlations between parameters lie in the range from −1 to 1: see Section 4.

some problems which needed to be overcome in the estimation of the multivariate model which will be dealt with in the next section.

4. Discussion

The Glasgow Health Board data show how a simple analysis can be achieved quickly by setting up a partitioned variance–covariance matrix to describe extra-Poisson variation in a log-linear model. The theory behind the model is quite complex, requiring the calculation of residuals at each iteration, and hence a powerful computer with a large memory is required if the number of areas is large. However, given a suitable software platform, here the MLn software (Rasbash and Woodhouse, 1995), which allows for flexible random coefficient modelling, and some modifications using macros, the modelling process can be made relatively simple. A version of the spatial analysis macros that is suitable for general use is planned in the future. The second example on prostate cancer shows a more complex series of models, which require more computing time as an extra parameter for distance decay needs to be estimated where spatial autocorrelation is included. The models show how care must be taken when investigating geographically distributed health data to formulate realistic hypotheses, and then testing these in various scenarios. Given sufficient data, it is possible to add covariates into the random part of the model at either level. Hence, models can easily become very complex, and this is why we emphasize the need for hypotheses to be properly specified before modelling begins. However, it must also be noted that a single final model may not be the optimal solution to the problem, and a range of possible scenarios may warrant presentation, as here. This is because of the complex nature of the interactions between variables and geographical space, and choices between competing models may be made on epidemiological as well as statistical grounds.

The multivariate model introduces a further set of issues concerning the complexity of the model to be analysed, concerning computational requirements and problems of interpretation. The first set of problems concerned the size of the workspace that is required, as separate design matrices need to be stored and manipulated for each of the random terms estimated. For large data sets, with several causes of death, this problem becomes intractable, even with powerful computers using large memories with the current method of estimation. The second set of problems involves obtaining estimates for variances, or correlations between parameters which are out of range (e.g. negative variance estimates and correlations outside the range from -1 to 1). A careful consideration of the influence of individual areas on the global statistics reported here obviously needs to be made, and some adjustment for outliers to be undertaken.

The theoretical basis for the spatial multilevel models that we have specified can be labelled as an empirical Bayes procedure because the random parameters are estimated directly from the data. By specifying random explanatory variables to define the spatial effects — a diagonal matrix of 1s for the global heterogeneity effects and a matrix of weights for the local spatial effects — and estimating variance and covariance parameters associated with these variables, we have produced a flexible modelling strategy which can be used in conjunction with more conventional hierarchical models (e.g. Langford and Bentham (1997) and Langford *et al.* (1998)). By comparing the size of the estimated variance parameters associated with heterogeneity and spatial effects, we judge the relative importance of these processes in explaining the variance seen in the dependent variable. This is similar to the method of Clayton and Kaldor (1987), where a parameter ρ is estimated to give the relative weight attached to heterogeneity and spatial effects in an autoregression model. However, the fully Bayesian approach (e.g. Bernardinelli and Montomoli (1992)) allows for prior distributions to be placed on the parameters in the spatial model. For example, whereas we estimate the heterogeneity parameter directly from the data, assuming normality for the random effects, it may be reasonable to assume a gamma or t-distribution as a prior for the relative risks. Our procedure could be modified to allow for this, but it is easier to implement in the BUGS software which uses Gibbs sampling (Spiegelhalter *et al.*, 1995). A further avenue which we are currently exploring is the use of nonparametric maximum likelihood procedures for estimating the distribution of relative risks (Aitkin, 1996).

In summary, we have explored the theory behind spatial multilevel modelling by using an IGLS procedure, and we have given three brief examples to show the possibilities that the technique may bring to the analysis of geographical data. However, the process is far from complete, and several problems and further possibilities are currently under investigation, as follows.

(a) Some of the models are inherently unstable, and the log-likelihood curves show several maxima and minima, or else bifurcate, with models oscillating between two stable states. This is particularly true of the distance decay models. One solution is to introduce a kernel around each district centroid to restrict its sphere of influence to a realistic distance. This will, of course, depend on the data and hypotheses being tested.

(b) The deviance statistic for the non-linear models cannot be easily calculated, and a simulation method for producing a quasi-likelihood ratio statistic is at present being investigated (Goldstein, 1996).

(c) Residuals can be taken from the model and posterior estimates of relative risk calculated. Bootstrapping can be used to develop an empirical distribution of the posterior relative risk for each area, but it is computationally intensive (Langford and Jones, 1998). Iterative bootstrapping to correct for bias may also be used with the MQL procedure, although this can further increase the effort required (Kuk, 1995; Goldstein, 1996).

(d) The non-linear models tend to fail to converge quite regularly. This is due to the PQL procedure, where predicted residuals (and their variances if the second-order term of the Taylor expansion is included) for each area are added back onto the fixed part of the model. If one or more of these is very large, then it invokes an arithmetic overflow when exponentiated. This is a technical detail, but it is important if a program for general users is to be developed. It can potentially be avoided by using iterative bootstrapping of the MQL procedure.

Conceptually, the clear message is that one must take a decision before analysis on whether an exploratory or inferential analysis is being conducted. For exploratory analyses, it is best to keep the models simple, with a heterogeneity and spatial term included in the model, perhaps at more than one level if this is justified. For inferential analysis, it is important to have specific hypotheses to test via competing models, as spatial effects tend to be rather poorly determined, and interact with covariates, and other non-spatial effects in the model. Complex models can easily be built, but less easily interpreted, and often it is not possible to judge meaningfully between competing models. However, the tools developed here provide a methodological and data analytic framework for the exploration of hypotheses where spatially distributed factors are of potential importance in understanding the aetiology of a disease.

Acknowledgements

This work was supported by the Economic and Social Research Council, UK, via two Visiting Fellowships under the Analysis of Large and Complex Datasets Programme to Ian Langford and Alastair Leyland at the Institute of Education, London. The Public Health Research Unit is financially supported by the Chief Scientist Office of the Scottish Office Department of Health. The opinions expressed in this paper are not necessarily those of the Chief Scientist Office. We would like to thank the Associate Editor and two referees for their helpful comments and advice on our paper.

Appendix A

Following Goldstein (1995), the residuals for the model with heterogeneity and spatial effects given in equation (11) may be estimated by

$$\begin{pmatrix} \hat{\theta}_u \\ \hat{\theta}_v^* \end{pmatrix} = \begin{pmatrix} \sigma_u^2 Z_u^{\mathrm{T}} + \sigma_{uv} Z_v^{*\mathrm{T}} \\ \sigma_{uv} Z_u^{\mathrm{T}} + \sigma_v^2 Z_v^{*\mathrm{T}} \end{pmatrix} V^{-1}(\pi - X\beta).$$

The variance–covariance matrix for the estimators is

$$\operatorname{var}\left\{\begin{pmatrix} \hat{\theta}_u \\ \hat{\theta}_v^* \end{pmatrix}\right\} = \begin{pmatrix} \sigma_u^2 \otimes I - (\sigma_u^2 Z_u^{\mathrm{T}} + \sigma_{uv} Z_v^{*\mathrm{T}}) M (\sigma_u^2 Z_u + \sigma_{uv} Z_v^*) & \sigma_{uv} \otimes I - (\sigma_u^2 Z_u^{\mathrm{T}} + \sigma_{uv} Z_v^{*\mathrm{T}}) M (\sigma_{uv} Z_u + \sigma_v^2 Z_v^*) \\ \sigma_{uv} \otimes I - (\sigma_{uv} Z_u^{\mathrm{T}} + \sigma_v^2 Z_v^{*\mathrm{T}}) M (\sigma_u^2 Z_u + \sigma_{uv} Z_v^*) & \sigma_v^2 \otimes I - (\sigma_{uv} Z_u^{\mathrm{T}} + \sigma_v^2 Z_v^{*\mathrm{T}}) M (\sigma_{uv} Z_v + \sigma_v^2 Z_v^*) \end{pmatrix}$$

where

$$M = V^{-1}\{V - X(X^{\mathrm{T}} V^{-1} X)^{-1} X^{\mathrm{T}}\} V^{-1},$$

$$V = \sigma_e^2 \otimes I + \sigma_u^2 Z_u Z_u^{\mathrm{T}} + \sigma_{uv}(Z_u Z_v^{*\mathrm{T}} + Z_v^* Z_u^{\mathrm{T}}) + \sigma_v^2 Z_v^* Z_v^{*\mathrm{T}}$$

and σ_e^2 is the lower level variance. The estimation for non-linear models remains basically unchanged following the transformations described in equations (7)–(9) with the addition of offset terms to the V-

and M-matrices. Although the equations presented here are in terms of one random and one spatial effect for each area, they may easily be extended to include further random coefficients and associated parameters.

References

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalised linear models. *Statist. Comput.*, **6**, 251–262.

Bailey, T. C. and Gatrell, A. C. (1995) *Interactive Spatial Data Analysis*. Harlow: Longman.

Bernardinelli, L., Clayton, D. and Montomoli, C. (1995) Bayesian estimates of disease maps: how important are priors? *Statist. Med.*, **14**, 2411–2432.

Bernardinelli, L. and Montomoli, M. (1992) Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statist. Med.*, **11**, 983–1007.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.*, **43**, 1–75.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Cisaghli, C., Biggeri, A., Braga, M., Lagazio, C. and Marchi, M. (1995) Exploratory tools for disease mapping in geographical epidemiology. *Statist. Med.*, **14**, 2363–2382.

Clayton, D. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology: Methods for Small Area Studies* (eds P. Elliott, J. Cuzick and D. English). New York: Open University Press.

Clayton, D. and Hills, M. (1993) *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Clayton, D. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Elliott, P., Cuzick, J. and English, D. (1992) *Geographical and Environmental Epidemiology: Methods for Small Area Studies*. New York: Open University Press.

Elliott, P., Martuzzi, M. and Shadick, G. (1995) Spatial statistical methods in environmental epidemiology: a critique. *Statist. Meth. Med. Res.*, **4**, 137–159.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993) Modelling complexity: applications of Gibbs sampling in medicine (with discussion). *J. R. Statist. Soc.* B, **55**, 39–102.

Goldstein, H. (1995) *Multilevel Statistical Models*. London: Arnold.

——— (1996) Likelihood computations for discrete response multilevel models. *Technical Report*. Multilevel Models Project, Institute of Education, London.

Goldstein, H., Healy, M. and Rasbash, J. (1994) Multilevel time series models with applications to repeated measures data. *Statist. Med.*, **13**, 1643–1655.

Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *J. R. Statist. Soc.* A, **159**, 505–513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, H. (1998) *A User's Guide to MLwiN*. London: Institute of Education.

Kemp, I., Boyle, P., Smans, M. and Muir, C. (1985) *Atlas of Cancer in Scotland 1975–1980: Incidence and Epidemiological Perspective*. Lyon: International Agency for Research on Cancer.

Key, T. (1995) Risk factors for prostate cancer. *Cancer Surv.*, **23**, 63–76.

Kuk, A. Y. C. (1995) Asymptotically unbiased estimation in generalized linear models with random effects. *J. R. Statist. Soc.* B, **57**, 395–407.

Langford, I. H. (1994) Using empirical Bayes estimates in the geographical analysis of disease risk. *Area*, **26**, 142–149.

——— (1995) A log-linear multi-level model of childhood leukaemia mortality. *J. Hlth Place*, **1**, 113–120.

Langford, I. H. and Bentham, G. (1996) Regional variations in mortality rates in England and Wales: an analysis using multi-level modelling. *Socl Sci. Med.*, **42**, 897–908.

——— (1997) A multilevel model of sudden infant death syndrome in England and Wales. *Environ. Planng* A, **29**, 629–640.

Langford, I. H., Bentham, G. and McDonald, A.-L. (1998) Multilevel modelling of geographically aggregated health data: a case study on malignant melanoma mortality and uv exposure in the european community. *Statist. Med.*, **17**, 41–57.

Langford, I. H. and Jones, A. P. (1998) Comparing area mortality rates using a random effects model and simulation methods. Submitted to *Statistician*.

Langford, I. H. and Lewis, T. (1998) Outliers in multilevel data (with discussion). *J. R. Statist. Soc.* A, **161**, 121–160.

Lawson, A. (1994) Using spatial Gaussian priors to model heterogeneity in environmental epidemiology. *Statistician*, **43**, 69–76.

Lawson, A. B. and Williams, F. L. R. (1994) Armadale: a case-study in environmental epidemiology. *J. R. Statist. Soc.* A, **157**, 285–298.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Mollié, A. and Richardson, S. (1991) Empirical Bayes estimates of cancer mortality rates using spatial models. *Statist. Med.*, **10**, 95–112.

Rasbash, J. and Woodhouse, G. (1995) *MLn Command Reference*. London: Institute of Education.

Schlattmann, P. and Böhning, D. (1993) Mixture models and disease mapping. *Statist. Med.*, **12**, 1943–1950.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1995) BUGS: Bayesian inference using Gibbs sampling. *Technical Report*. Medical Research Council Biostatistics Unit, Cambridge.

Statistics Canada (1991) *Mortality Atlas of Canada*. Ottawa: Canada Communications Group.