Journal of Royal Statistical Society, Series C, (1999). Vol. 48, pps. 253-268

MULTILEVEL MODELLING OF THE GEOGRAPHICAL DISTRIBUTIONS OF DISEASES

Ian H. Langford[1,2], Alastair H. Leyland [2,3], Jon Rasbash[2] and Harvey Goldstein[2]

[1] Centre for Social and Economic Research on the Global Environment (CSERGE), University of East Anglia, Norwich, and University College, London, UK

[2] Multilevel Models Project, Institute of Education, University of London, UK

[3] Public Health Research Unit, University of Glasgow, UK

Address for Correspondence:

Dr Ian H. Langford
CSERGE
School of Environmental Sciences
University of East Anglia
Norwich NR4 7TJ
UNITED KINGDOM

Tel: +44 1603 593314
Fax: +44 1603 593739

email: i.langford@uea.ac.uk

Summary

This paper addresses some of the theoretical and methodological problems of modelling the distribution of diseases, such as cancer, in discrete geographical areas. Theoretically, it is necessary to examine in detail the various processes, both artefactual and causative, which may affect the number of cases occurring within a certain area, and the distribution of relative risks between areas. A methodological framework based on multilevel modelling is developed, with spatial and nonspatial relationships being considered as random effects occurring at different levels within a population data structure. Examples of exploratory and inferential analyses are given, and discussion focuses on the issues raised by complex spatial modelling of geographically distributed health data.

Keywords: multilevel modelling; geographical epidemiology; spatial analysis; random coefficient models; cancer epidemiology

1. Introduction

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

(a) exploratory analysis: maps of disease distribution are produced to provide health researchers with a visual display which can suggest, via patterns and spatial trends, useful avenues of research into causal processes. Atlases containing collections of such maps attempt to reflect the distribution of a range of diseases over a large geographical area (e.g. Kemp *et al*., 1985; Statistics Canada, 1991). The main aim in this type of analysis is to provide a picture which, in some respects at least, reflects the true distribution of cases of disease over the area depicted. However, if one simply uses the relative risk of a disease (defined as the number of observed cases divided by number of expected cases for each area) for this purpose, then problems may occur as areas with small populations, usually in rural locations, will tend to have extreme relative risks as the number of expected cases in the denominator is low. Conversely, if one maps probability values, say $p < 0.05$ for a particular relative risk being equal to unity, then these will tend to occur in areas with large populations, usually in urban areas, as the probability value is related to sample size. A more complete discussion is given in Clayton and Kaldor (1987) and Langford (1994), but 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, and/or the local distribution of relative risks in areas geographically close to each other. The example given in section 3.1 gives a simple example of how this technique can be implemented as part of a multilevel modelling analysis using data collected on deaths from all causes in Greater Glasgow. Section 3.3 then uses the flexibility of the multilevel model to develop a multivariate spatial analysis, where deaths from two different causes, namely cancer and circulatory diseases are modelled together. We discuss in section 2, and also section 4 how residuals may be extracted to provide information for mapping the distributions of these diseases;

(b) inferential analysis: in this case, 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 allows for more reliable inferences to be made, although we demonstrate that choosing between different spatial models is not always straightforward using data collected on prostate cancer incidence in the 56 counties of Scotland. 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 higher incidence of prostate cancer.

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

(a) within-area effects, such as population at risk, individual characteristics, and so on;

(b) hierarchical effects. These are due to the fact that small areas are grouped into larger areas, for administrative purposes, or for cultural and geographical reasons. For example, a number of small areas, such as 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 or demographic factors which influence the incidence or outcome of disease. In addition, as areas are usually formed from geopolitical boundaries which have nothing to do with the disease we are interested in, we may wish to spatially smooth the distribution or relative risks to remove any artefactual variation brought into the data by the method of aggregating the data.

The use of empirical Bayes and fully Bayesian techniques has allowed for alternative models of spatial and environmental processes affecting the distribution of a disease which rely on different underlying beliefs or assumptions about aetiology (Bernardinelli *et* al., 1995; Bernardinelli and Montomoli, 1992; Cisaglhi *et* al., 1995; Clayton and Kaldor, 1987; Langford *et al.*, in press; Langford, 1994; 1995; Lawson, 1994; Lawson and Williams, 1994; Mollie and Richardson, 1991; Schlattmann and Bohning, 1993). Two main statistical methodologies have been used to model geographically distributed health data in this way. The first are Markov chain Monte Carlo (MCMC) methods, using Gibbs sampling (Gilks *et al*., 1993) often fitted using the BUGS software (Spiegelhalter *et al*., 1995). The second set of methods are multilevel modelling techniques based on iterative generalised least squares procedures (IGLS: Goldstein, 1995) and are the focus of this paper. These methods can be described as using the Bayesian and empirical Bayesian models respectively, and we discuss the differences between the two approaches in the discussion. In the following section, we detail the methodology and computational algorithms necessary to model the three types of effect described above within the IGLS framework. We then present three brief examples of analyses of geographical health data, and the discussion focuses on issues surrounding both the theory and methodology of building complex spatial models, and provides pointers for future research. The models were all fitted using the multilevel modelling software, MLn (Rasbash and Woodhouse, 1995).

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:

$$Y = X\beta + Z\theta \qquad (1)$$

with a vector of observations Y being modelled by explanatory variables X and associated fixed parameters β, and explanatory variables Z with random coefficients $Z\theta$. The fixed and random part design matrices X and Z need not be the same. Goldstein (1995) describes 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. First, we estimate the fixed parameters in an initial ordinary least squares regression, assuming the variance at higher levels on the model to be zero. From the vector of residuals

from this model we can construct initial values for V. Then, we iterate the following procedure, first estimating fixed parameters in a generalised least squares regression as:

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

and again calculating residuals $\tilde{Y} = Y - X\hat{\beta}$. By forming the matrix product of these residuals, and stacking them into a vector, i.e. $Y^* = vec(\tilde{Y}\tilde{Y}^T)$ we can estimate the random parameters θ as:

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

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

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

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

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

2.2 The Poisson model

In order to model the distribution of rare diseases, we use a model developed for the distribution of relative risks of a disease by Besag *et al.* (1991). If we consider a population of areas with 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:

$$E_i = N_i . \frac{\sum O_i}{\sum N_i} \tag{6}$$

and may be additionally divided into different age and sex bands. We can write the basic Poisson model as:

$$\begin{aligned} &O_i \sim Poisson(\mu_i), \\ &\log(\mu_i) = \log(E_i) + \alpha + X_i\beta + u_i + v_i \end{aligned} \tag{7}$$

where $\log(E_i)$ is treated as an offset, and α is a constant. We take account of the distribution of cases *within* each area by assuming the cases have 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. 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 nonlinear (logarithmic) relationship between the outcome variable and the predictor part of the model. There are two options:

(a) if the cases in each area 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 these

follow a Normal distribution. In this case, heterogeneity effects can be accommodated by weighting the random part of the model by some function of the population at risk in each area;

(b) in cases of very rare diseases, we can make a linearising approximation for estimating the random parameters. If we take the case of having heterogeneity effects only for the sake of simplicity, we can estimate residuals $\hat{u}_i$ from the model using penalized quasi-likelihood (PQL) estimation with a second order Taylor series approximation (Breslow and Clayton, 1993; Goldstein, 1995; Goldstein and Rasbash, 1996). After each iteration, *t*, we make predictions H_t from the model, where $H_t = X_i\hat{\beta}_t + \hat{u}_i$,and use these to calculate new predictions for iteration *t*+1, so that:

$$\begin{aligned} f(H_{t+1}) = f(H_t) + X_i(\hat{\beta}_{t+1} - \hat{\beta}_t)f'(H_t) \\ +(u_i - \hat{u}_i)f'(H_t) + (u_i - \hat{u}_i)^2 f''(H_t)/2 \end{aligned} \qquad (8)$$

with the first two terms on the right hand side of (8) provides the updating function for the fixed part of the model. The third term comprises a linear random component created by multiplying the first differential of the predictions by the random part of the model, and the fourth term is the next term in the Taylor expansion about H_t. For the Poisson distribution:

$$f(H) = f'(H) = f''(H) = \exp(X_i\hat{\beta}_t + \hat{u}_i) \qquad (9)$$

Hence, at each iteration we estimate about the fixed part of the model plus the residuals. A full description of this procedure can be found in Goldstein (1995) and Goldstein and Rasbash (1996). This can lead to problems with convergence, or with the model "blowing up" if some of the residuals are particularly large. In these cases, the second order term in (8) can be omitted, or, in extreme cases, estimates can be based on the fixed part of the model only. This latter case 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). For equation (7) we substitute $\hat{u}_i + \hat{v}_i$ for $\hat{u}_i$ in (8) and (9).

2.3 Defining the spatial structure

There are several possibilities for specifying the structure of the random effects in the model(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 (7) with intrinsic Gaussian distributions for each type of effect.

We adopt a somewhat 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 1's 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, the case of the spatial effects is more complex, as we require off-diagonal

terms in the variance covariance matrix. First, we can define a set of random explanatory variables, Z_v, one for each area, which contain the square roots of the spatial weights linking areas, w_{ij}. Our formulation of the spatial model is therefore altered from the basic spatial Poisson model given in (7) so that:

$$O_i \sim Poisson(\mu_i),$$
$$\log(\mu_i) = \log(E_i) + \alpha + X_i\beta + u_i + \sum_{j \neq i} z_{ij} v_j \qquad (14)$$

We then need to form the square of the matrix containing the z_{ij} to form the spatial component of V in order to estimate the associated variance parameter σ_v^2 (see equations (17) and (18) below). For example, to model (11), we require:

$$Z_v Z_v^{/} = \begin{bmatrix} 0 & w_{12}/w_{1+} & w_{13}/w_{1+} \\ w_{21}/w_{2+} & 0 & w_{23}/w_{2+} \\ w_{31}/w_{3+} & w_{32}/w_{3+} & 0 \end{bmatrix}^{0.5} \begin{bmatrix} 0 & w_{12}/w_{1+} & w_{13}/w_{1+} \\ w_{21}/w_{2+} & 0 & w_{23}/w_{2+} \\ w_{31}/w_{3+} & w_{32}/w_{3+} & 0 \end{bmatrix}^{T0.5} \qquad (15)$$

for an adjacency matrix of three areas. For example, if:

$$Z_v = \begin{bmatrix} 0 & \sqrt{0.5} & \sqrt{0.5} & \cdots \\ 1 & 0 & 0 & \\ 1 & 0 & 0 & \\ \vdots & & & \ddots \end{bmatrix}$$

this becomes:

$$Z_v Z_v^{/} = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 1 & \\ 0 & 1 & 1 & \\ \vdots & & & \ddots \end{bmatrix}$$

where area 1 is adjacent to both areas 2 and 3, and areas 2 and 3 are only adjacent to area 1. The structure of the variance-covariance matrix associated with the spatial effects is that each row represents an area, and the weights sum to 1. Of course, for the distance decay function, each cell will have a value equal to $\exp(-\lambda d_{ij})$, and zeros again on the diagonal.

Finally, there is the problem of specifying the random effects for heterogeneity and spatial effects within a generalised linear modelling framework, in this case using IGLS estimation within the MLn software. There are two basic options for fitting the random effects within MLn which demonstrate some more general issues for spatial modelling:

(a) 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},..., z_{vn}. In the case of the distance decay model, $z_{v1} = \{\exp(-\lambda d_{i1})\}^{0.5}$ as described in (14) above, and so on. 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, there is a problem, because we

only wish to estimate a single variance parameter for *all areas* for heterogeneity effects, one for spatial effects, and a single covariance term. Hence, we need to constrain the parameter estimates for each area to be same for each set of effects, e.g. σ_v^2 is constrained to be the same for all the z_v's. This means introducing these complex constraints into the model via a set of linear equations. A discussion of this procedure can be found in Goldstein (1995) but the important point to note is that we have to include 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, this formulation easily allows for the calculation of residuals from the model, as these can be estimated for each of the random explanatory variables. This is important if the focus of our investigation is upon comparison of relative risks between areas in the data set. It is less of an issue if we are only interested in the global parameters such as σ_u^2 and σ_v^2 which describe the size of the overall heterogeneity and spatial effects present in the model;

(b) in contrast to the above, we can build the weights matrices associated with the random effects and fit these directly into the model. Consider the variance of

Y given $X\beta$ from equation (1) written as:

$$Var(Y|X\beta) = Z\Sigma_\theta Z^T \tag{16}$$

where Σ_θ is the variance of the random parameters θ. The structure of Σ_θ will often lead to simplifications; for example, in a random effects model when $\theta = \{u_i\}$ and $Var(u_i) = \sigma_u^2$, $Cov(u_i, u_j) = 0$ then $\Sigma_\theta = \sigma_u^2 I$ and so $Var(Y|X\beta) = \sigma_u^2 ZZ^T$. Similarly, in a spatial model, if θ and Z are partitioned such that equation (1) may be rewritten

$$Y = X\beta + \begin{bmatrix} Z_u & Z_v \end{bmatrix} \begin{bmatrix} \theta_u \\ \theta_v \end{bmatrix} \tag{17}$$

then, with $Var\left(\begin{bmatrix} \theta_u \\ \theta_v \end{bmatrix}\right) = \begin{bmatrix} \sigma_u^2 I & \sigma_{uv} I \\ \sigma_{uv} I & \sigma_v^2 I \end{bmatrix}$ which is equivalent to $Var\left(\begin{bmatrix} u_i \\ v_i \end{bmatrix}\right) = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}$. An important point to note is that in our procedure, we are expressing the spatial correlation in terms of the explanatory variables, Z_v, as described above in (14). Hence, for the residuals for each area we obtain from the model it is true that $\text{cov}(u_i, u_j) = \text{cov}(v_i, v_j) = \text{cov}(u_i, v_j) = 0$, and hence we can write:

$$Var(Y|X\beta) = \sigma_u^2 Z_u Z_u^T + \sigma_{uv}\left(Z_u Z_v^T + Z_v Z_u^T\right) + \sigma_v^2 Z_v Z_v^T \tag{18}$$

Expressing the model in terms of these design matrices overcomes the need to place multiple equality constraints upon the random parameters. This is generalisable to the non-linear model expressed in equations (7) to (9). A penalised quasi-likelihood (PQL) estimation procedure requires the estimation of the residuals and their associated variances at each iteration. The estimation of the residuals is described in the Appendix.

3. Applications

In this section, we give three examples of results from health data sets in order to comment on methodological issues raised in the discussion and show how substantive interpretations can made of spatial multilevel models.

3.1 Greater Glasgow Health Board mortality data

The data for this example refer to deaths from all causes in 143 postcode sectors within Greater Glasgow Health Board (GGHB) in 1993 obtained from the Registrar General for Scotland. Hence, as postcode sectors are relatively small (average population ≈ 5000), and the data are only for one year, we formulate the model in a similar way to equation (7):

$$\begin{aligned} &O_i \sim Poisson(\mu_i), \\ &\log(\mu_i) = \log(E_i) + \alpha + u_i + \sum_{j \neq i} z_{ij} v_j \end{aligned} \tag{19}$$

where the E_i are age-sex standardised for the Greater Glasgow Health Board area. For a *first order autocorrelation model* we consider $z_{ij} = 1/n_i$ if area j is a neighbour of area i and 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 upon its neighbours with the summation term $\sum_{j \neq i} z_{ij} v_j$ 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 upon its neighbours:

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix} \right) \tag{20}$$

Then this may be expressed in the terms of equation (16) by writing

$$\begin{aligned} X &= \{\log(E_i) \quad 1\} \\ \beta &= \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \\ Z_u &= I \\ Z_v &= \{z_{ij}\} \end{aligned} \tag{21}$$

and so estimation may proceed as described in equation (18) and the Appendix. The parameter estimates for this model are shown in table 1. To aid convergence, $\log(E_i)$ were centred around zero, 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 of 13.54 with 2 d.f. (p = 0.0011). The correlation between the random effects and spatial effects is 0.786, 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 the number of neighbours it has 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 would imply a total variance of 0.0328, of

which 48% arises from the spatial effects. The estimated covariance between two areas depends on;

(i) whether the two areas border each other, and;

(ii) the number of shared neighbours.

In terms of the z_{ij} used in equation (22) the covariance between areas i and j can be expressed as $(z_{ij} + z_{ji})\sigma_{uv} + \sum_{k \neq i,j} z_{ik} z_{jk} \sigma_v^2$.

Substantively, this is the kind of basic application of spatial modelling that is required for preliminary exploratory analysis, or the production of maps of relative risk. In this case, the model shows that there are significant parameters for both heterogeneity and spatial autocorrelation (using a Wald test with significance level $\alpha = 0.05$). This makes sense, as postcode sectors are quite variable in population size, and this effect is summarised 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 comparisons to be made with σ_u^2 in each area) and there is a significant covariance between the two effects. This may be because mortality rates are similar in social areas larger than the postcode sectors analysed here. A further analysis could place larger units such as social neighbourhoods as a higher level in the model to test for this effect, and covariates such as social and housing status could be included. The significant covariance occurs because areas whose populations have similar socio-demographic characteristics (and also large populations) tend to cluster together and also have similar mortality rates.

3.2 Prostate cancer incidence in Scottish districts

In this example, we examine data covering six years, from 1975-80 on the incidence of prostate cancer in 56 districts in Scotland (Kemp *et al.*, 1985). As the data are collected in relatively large geographical units for a longer time period than the first example, the numbers of cases occurring in each district are sufficiently large (between 10 and 627 cases) that we can model the relative risks of disease incidence (based on crude rates) and assume that log relative risk follows an approximately Normal distribution. In this case, we wish to investigate the hypothesis that the relative risk of prostate cancer is higher in rural than urban areas, as previous research has indicated an association between agricultural employment and incidence of prostate cancer (Key, 1995). In this case, we use a variable measuring the percentage of the male workforce employed in agriculture, fishing and forestry industries as a surrogate measure of the rurality of an area. However, we have to look not only at the incidence of prostate cancer within districts, but account for a potential artefactual effect caused by differential diagnosis rates between Health Board areas in Scotland. Hence, we are modelling spatial effects caused by different processes at two different scales, namely:

(a) a spatial autocorrelation model at district scale, where we are accounting for the possibility that areas closer in geographical space have similar incidence of prostate cancer;

(b) a variance components model at Health Board scale, where we investigate the possibility that different Health Boards have different relative risks of prostate cancer, potentially because diagnostic criteria are variable.

Hence, we can extend equation (16) so that:

$$Y = X\beta + \begin{bmatrix} Z_u & Z_v \end{bmatrix} \begin{bmatrix} \theta_u \\ \theta_v \end{bmatrix} + Z_{hb}\theta_{hb} \qquad (22)$$

In this case, we have three explanatory variables in the fixed part of the model ($X\beta$) in addition to the intercept term (CONS), namely: the proportion of the population in higher social classes (SC12); the estimated incidence to ultraviolet light at the earth's surface (UVBI), and; the percentage of the male employment in agriculture, fishing and forestry (AGRI). Social class and ultraviolet light exposure have been included as these have been previously postulated as risk factors for prostate cancer. In this case, Z_v is calculated using distances between district centroids, and a distance decay parameter λ is estimated from the spatial linkage described in equation (12). Z_{hb} is a vector of 1's which allows for a variance component for each Health Board to be estimated, and hence a measure of the variance at this scale, σ^2_{hb} . Table 2 presents the results when we take Z_u is taken as a vector of 1's, and hence equal weight is given to each district. Four models are shown, representing: a simple, single level model with no spatial effects; a model with district scale spatial effects, but no Health Board effects; a model with only Health Board effects; a model with both district and Health Board effects as given in equation (22).

The simple model presented in Table 2 seems to indicate a strong, and significant effect of rurality, as measured by percentage male agricultural employment (AGRI). However, this is weakened by fitting a spatial autocorrelation parameter, which suggests that some of the effect of AGRI may be due to adjacent areas having similar mortality rates, although the parameter for AGRI is still statistically significant. The change in deviance between the two models is 14.9 on two degrees of freedom ($p < 0.01$: we have fitted a covariance parameter as well as a variance term). The third model, 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 (in this case districts) could lead to misleading results at higher levels (in this case, Health Boards). Unexplained random variation at district level can appear spuriously at Health Board level, although the final model, with both Health Board and spatial effect between districts, suggest that Health Board areas are significantly different from each other. The parameter estimate for AGRI becomes insignificant in the Health Boards only model, but becomes stronger again in the combined model. 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 analysis of residuals in these complex models: a forthcoming paper by Langford and Lewis (1997) details some procedures for the general analysis of outliers in multilevel models.

However, the models in Table 2 have not taken into account the different populations at risk in each district. We can also specify $Z_u = n^{-0.5}$, where n is the vector of population sizes for the districts in the study area, and the districts are hence weighted in the random part of the model by their population size. When this is done (results not shown here), the fixed parameter for AGRI becomes statistically insignificant in all models except the simple one,

and the Health Board effects become less important relative to spatial autocorrelation between districts. The substantive conclusions to be drawn from this dataset are left to elsewhere, but it is interesting to note that a range of potentially interesting models can be generated from what appears to be a simple issue, namely testing the relationship of one variable, AGRI, with prostate cancer incidence.

3.3 Multivariate spatial analysis of mortality in GGHB postcodes

In the example above, we have extended the scale of spatial analysis to include Health Board as well as district level effects. We can extend the methods described to be applicable to more than one disease within the same model. For example, we can look at multiple causes of death from the GGHB postcoded 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 "A") and deaths from circulatory diseases (indexed as "B") we can write the model:

$$\begin{bmatrix} O_{A,i} \\ O_{B,i} \end{bmatrix} \sim Poisson \begin{bmatrix} \lambda_{A,i} \\ \lambda_{B,i} \end{bmatrix} \tag{26}$$

where

$$\log\left(\begin{bmatrix} \lambda_{A,i} \\ \lambda_{B,i} \end{bmatrix}\right) = \log\left(\begin{bmatrix} E_{A,i} \\ E_{B,i} \end{bmatrix}\right) + \begin{bmatrix} \alpha_A \\ \alpha_B \end{bmatrix} + \begin{bmatrix} u_{A,i} \\ u_{B,i} \end{bmatrix} + \begin{bmatrix} \sum_{j \neq i} d_{ij} v_{A,j} \\ \sum_{j \neq i} d_{ij} v_{B,j} \end{bmatrix} \tag{27}$$

This gives a possible 16 random parameters to be estimated. However, we do estimate all 16 because of the difficulty in interpreting some of the parameters. Specifically, the covariance between the spatial parts of the two causes $\sigma_{v,CA,CIRC}$ and between the random effect of one cause and the spatial part of the other cause $\sigma_{uv,CA,CIRC}$ and $\sigma_{uv,CIRC,CA}$ have all been set to zero. Hence, we estimate:

$$\begin{bmatrix} u_{CA,i} \\ u_{CIRC,i} \\ v_{CA,i} \\ v_{CIRC,i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u,CA} & \sigma_{u,CA,CIRC} & \sigma_{uv,CA} & 0 \\ \sigma_{u,CA,CIRC} & \sigma^2_{u,CIRC} & 0 & \sigma_{uv,CIRC} \\ \sigma_{uv,CA} & 0 & \sigma^2_{v,CA} & 0 \\ 0 & \sigma_{uv,CIRC} & 0 & \sigma^2_{v,CIRC} \end{bmatrix}\right) \tag{28}$$

The results for this model are given in Table 4. As can be seen, considering only two causes of death, and including no covariates in the model still leads to the estimation of 10 parameters to account for heterogeneity and spatial effects. However, it is interesting to note that both the heterogeneity and 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 postcodes. One author (AHL) is currently investigating the effect of entering a covariate measuring deprivation into the model. Computationally, there were a number of problems which needed to be overcome in the estimation of the multivariate model which will be dealt with in the Discussion.

4. Discussion

The Glasgow Health Board data show how a simple analysis can be achieved relatively 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 large memory is required if the number of areas is large. However, given a suitable software platform, in this case the MLn software (Rasbash and Woodhouse, 1995), which allows for flexible random coefficient modelling, and some modifications using macros, it is possible to make the modelling process relatively simple. A version of the spatial analysis macros suitable for general use are 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 test these in a number of scenarios. The data set here is small, but with sufficient numbers of areas, and hence information, 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 in these cases 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.

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 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 -1 to +1). Careful consideration of the influence of individual areas upon the global statistics reported here obviously needs to be made, and some adjustment for outliers undertaken.

The theoretical basis for the spatial multilevel models we have specified is frequentist, and could be labelled as an empirical Bayes procedure because we estimate our random parameters directly from the data. By specifying random explanatory variables to define the spatial effects - a diagonal matrix of 1's 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; Langford, Bentham and McDonald, 1997). 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, whilst 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. It would not be impossible to modify our procedure 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 demonstrated the theory behind spatial multilevel modelling using an iterative generalised least squares procedure, and have given a couple of brief examples to show the possibilities the technique may bring to analysis of geographical data. However, the process is far from complete, and a number of problems and further possibilities are currently under investigation, namely:

(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, be dependent on the data and hypotheses being tested;

(b) the deviance statistic for the nonlinear models cannot be easily calculated, and a simulation method for producing a quasi-likelihood ratio statistic is presently 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 is computationally intensive (Langford and Jones, submitted paper). 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 nonlinear models tend to "crash" 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 is important if a program for general users is to be developed. It can be avoided by using iterative bootstrapping of the MQL procedure.

Conceptually, the clear message is that one must take a decision prior to 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 is 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 nonspatial 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.

Appendix

Following Goldstein (1995), the residuals for a model with heterogeneity and spatial effects may be estimated by

$$\begin{bmatrix} \hat{r}_u \\ \hat{r}_v \end{bmatrix} = \begin{bmatrix} \sigma_u^2 Z_u^T + \sigma_{uv} Z_v^T \\ \sigma_{uv} Z_u^T + \sigma_v^2 Z_v^T \end{bmatrix} V^{-1} \left(Y - X\beta \right) \tag{A1}$$

and their variances are given by:

$$Var\left(\begin{bmatrix} \hat{r}_u \\ \hat{r}_v \end{bmatrix} \right) = \begin{bmatrix} \sigma_u^2 \otimes I - (\sigma_u^2 Z_u^T + \sigma_{uv} Z_v^T) M (\sigma_u^2 Z_u + \sigma_{uv} Z_v) & \sigma_{uv} \otimes I - (\sigma_u^2 Z_u^T + \sigma_{uv} Z_v^T) M (\sigma_{uv} Z_u + \sigma_v^2 Z_v) \\ \sigma_{uv} \otimes I - (\sigma_{uv} Z_u^T + \sigma_v^2 Z_v^T) M (\sigma_u^2 Z_u + \sigma_{uv} Z_v) & \sigma_v^2 \otimes I - (\sigma_{uv} Z_{u1}^T + \sigma_v^2 Z_v^T) M (\sigma_{uv} Z_v + \sigma_v^2 Z_v) \end{bmatrix} \tag{A2}$$

where:

$$M = V^{-1}[V - X(X^T V^{-1} X)^{-1} X^T] V^{-1} \tag{A3}$$

and:

$$V = \sigma_e^2 \otimes I + \sigma_u^2 Z_u Z_u^T + \sigma_{uv} \left(Z_u Z_v^T + Z_v Z_u^T \right) + \sigma_v^2 Z_v Z_v^T \tag{A4}$$

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) to (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 variance components in generalised linear models. *Statistics and Computing*, **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? *Statistics in Medicine*, **14**, 2411-32.

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

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with Discussion) . *Annals of the Institute of Statistical Mathematics*, **43.1**, 1-75.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Cisaghli C., Biggeri A., Braga M., Lagazio C. and Marchi M. (1995). Exploratory tools for disease mapping in geographical epidemiology. *Statistics in Medicine*, **14**, 2363-82.

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

Clayton, D. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk. In

Elliott, P., Cuzick J. and English D. *Geographical and Environmental Epidemiology: Methods for Small Area Studies*. New York: Open 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. *Statistical Methods in Medical Research*, **4**, 137-159.

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

Kirby, A.J. (1993). Modelling complexity: applications of Gibbs sampling in medicine (with Discussion). *Journal of the Royal Statistical Society Series B*, **55**, 39-102.

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

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

Goldstein, H., Healy, M. and Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, **13**, 1643-55.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A*, **159**, 505-514.

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

Kuk, A.Y.C. (1995) Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of the Royal Statistical Society*, *Series B*, **57**, 395-407.

Kemp, I., Boyle, P., Smans, M. and Muir, C. (1985). *Atlas of cancer in Scotland 1975-1980: incidence and epidemiological perspective*. Lyon: IARC Scientific Publications.

Langford, I.H. (1995) A log-linear multi-level model of childhood leukaemia mortality, *Journal of Health and Place*, **1.2**, 113-120.

Langford, I.H. (1994) Using empirical Bayes estimates in the geographical analysis of disease risk. *Area*, **26.2**, 142-149.

Langford, I.H. and Bentham, G. (1996) Regional variations in mortality rates in England and Wales: an analysis using multi-level modelling. *Social Science and Medicine*, **42.6**, 897-908.

Langford, I.H., Bentham, G. and McDonald, A-L. Multilevel modelling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European community. *Statistics in Medicine*, in press.

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

Langford, I.H. and Lewis, T. (1997). Outliers in multilevel models. *Journal of the Royal Statistical Society Series A*, in press.

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

Lawson A.B. and Williams F.L.R. (1994). Armadale: a case study in environmental epidemiology. *Journal of the Royal Statistical Society Series A*, **157**, 285-298.

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

Mollie A. and Richardson S. (1991). Empirical Bayes estimates of cancer mortality rates using spatial models. *Statistics in Medicine*, **10**, 95-112.

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

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

Schlattmann P. and Böhning D. (1993). Mixture models and disease mapping. *Statistics in Medicine*, **12**, 1943-50.

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

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

Parameter	Estimate	Standard error
α	4.198	0.0371
σ_u^2	0.0172	0.00530
σ_{uv}	0.0299	0.00841
σ_v^2	0.0846	0.0328

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

	(A) simple model		(B) spatial effects		(C) Health Board effe	
	estimate	st. error	estimate	st. error	estimate	st. error
Fixed part						
intercept	-0.0257	0.584	-0.513	0.605	-0.0108	0.636
SC12	-0.000645	0.00524	0.00145	0.00389	-0.00339	0.00477
UVBI	-0.0141	0.0635	0.0565	0.0704	-0.00112	0.0705
AGRI	0.0272	0.00603	0.0163	0.00636	0.180	0.00634
Random part						
σ^2_{hb}					0.0327	0.0183
σ^2_u	0.0822	0.0155	0.0665	0.0141	0.0530	0.0117
σ_{uv}			0.000805	0.000414		
σ^2_v			0.0000159	0.0000167		
λ			7.23			
residual deviance	18.98		4.09		10.93	

Table 3 Parameter estimates and standard errors for the Glasgow Health Board mortality data for cancer (CA) and circulatory (CIRC) deaths

Parameter	Estimate	Standard error
α_{CA}	2.820	0.0310
α_{CIRC}	3.397	0.0377
$\sigma^2_{u,CA}$	0.00205	0.00530
$\sigma_{u,CA,CIRC}$	0.00224	0.00841
$\sigma^2_{u,CIRC}$	0.00472	0.0328
$\sigma_{uv,CA}$	0.0112	0 †
$\sigma_{uv,CIRC,CA}$	0	0
$\sigma^2_{v,CA}$	0.0606	0.0389
$\sigma_{uv,CA,CIRC}$	0	0
$\sigma_{uv,CIRC}$	0.0237	0.0168
$\sigma_{v,CA,CIRC}$	0	0
$\sigma^2_{v,CIRC}$	0.122	0.0439

† This parameter has been constrained so that the correlations between parameters lie in the range -1 to +1 : see Discussion.