INT. J. GEOGRAPHICAL INFORMATION SCIENCE, 2002
VOL. 16, NO. 2, 151–166

Taylor & Francis
Taylor & Francis Group

Research Article

Modelling a discrete spatial response using generalized linear mixed models: application to Lyme disease vectors

ABHIK DAS
Statistics Research Division, Research Triangle Institute, 6110 Executive Blvd, Suite 420, Rockville, MD 20852, USA; e-mail: adas@rti.org

SUBHASH R. LELE
Department of Mathematical Sciences, University of Alberta, Canada

GREGORY E. GLASS, TIMOTHY SHIELDS
Department of Molecular Microbiology and Immunology, Johns Hopkins University School of Hygiene & Public Health, Baltimore, MD, USA

and JONATHAN PATZ
Division of Environmental and Occupational Health, Johns Hopkins University School of Hygiene & Public Health, Baltimore, MD, USA

Abstract. Predicting disease risk by identifying environmental factors responsible for the geographical distribution of disease vectors can help target control strategies and optimize preventive measures. In this study we present a hierarchical approach to model the distribution of Lyme disease ticks as a function of environmental factors. We use the Poisson framework natural for count data while allowing for spatial correlations. To help identify environmental factors that best explain tick abundance, we develop an intuitive procedure for covariate selection in the spatial context. These methods could be useful in analysing effects of environmental and climatological changes on the distribution of disease vectors, and the spatial extrapolation of vector abundance under such scenarios.

1. Introduction

Estimating the distribution and abundance of biological populations remains a critical, but difficult, task. This is especially true when we attempt to account for the effects of environmental heterogeneity on population sizes, within a portion of a species range. Methods allowing researchers to assess population sizes from sampling done in diverse environmental conditions are rare. However, such attempts are important in assessing, for example, the disease risk from infectious agents. They are also pertinent to climate change and disease analysis, as altered geographic distribution of disease vectors is anticipated under future climate scenarios. As an illustration, consider the emergence and transmission of vector-borne diseases such as Lyme disease, which is of considerable public health concern. Projections of climate change are expected to cause an upward geographic shift in both altitude and latitude for

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2002 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810110099134

many vector-borne diseases (Patz *et al.* 1996). It is thus important that we have the methods to quantify the future risk of such diseases. While the future risk of a disease is not directly observable, it is known that vector abundance can serve as a surrogate for disease risk (Wallis *et al.* 1978). The modelling of vector abundance can thus be an essential tool for enhancing the accuracy of disease risk predictions.

In this study, we focus on Lyme disease, the most prevalent vector-borne disease in the USA. The black-legged deer tick, *Ixodes scapularis*, is the main vector for *Borrelia burgdorferi*, the causative agent for Lyme disease in eastern USA (Steere 1989). It has been shown that the emergence of Lyme disease follows deer tick introduction (Lastavica *et al.* 1989). There is also evidence that tick abundance, defined as the number of *I. scapularis* ticks present per deer, is directly related to disease risk (Wallis *et al.* 1978). Further, note that though these ticks most often use the white-tailed deer as a host (Apperson *et al.* 1990), they do not require such a host per se, and prevalence of *B. burgdorferi* has been documented in ticks collected from other sources (Solberg *et al.* 1995). Therefore, the abundance of ticks on deer, which is an indicator of the vector's population size, rather than deer abundance in itself, can be considered a proxy for the risk of Lyme disease.

Identifying environmental factors responsible for the distribution of tick abundance and using them to predict the risk of Lyme disease is a particularly useful approach for targeting control strategies and optimizing preventive measures. In this paper we present an approach to modelling tick abundance as a function of environmental factors, such as land use categories, soil characteristics, vegetation and slope of land.

Consider a study of the Lyme disease tick population reported in Glass *et al.* (1994). The ticks were collected from hunter-killed deer in November 1990 at three deer check stations in Kent County, Maryland. For each of 18 locations spread across the county, that were identified by the hunters, the data consisted of the number of male and female ticks recovered from each deer killed at that location. In addition, environmental data on land-use/land-cover patterns (urban, agricultural, forest, etc.), watershed distributions, soil types, land ownership (private and public) and steepness of land (flat to hilly) for each of the locations was derived from 1985 digital maps of Maryland using Geographical Information System (GIS) techniques. In order to use such data to predict the risk of Lyme disease properly, we should take into account that the response (number of ticks recovered from deer) is in the form of counts that are spatially spread over the study area. The covariate information on environmental characteristics is also location specific and thus spatial in nature.

The traditional technique for prediction in spatial statistics is kriging (Cressie 1993, Ribeiro *et al.* 1996). This procedure produces a smooth interpolated map of the response (such as vector abundance) over the area of interest, which can help policy makers make decisions. Unfortunately, in the context of predicting vector abundance, the kriging formulation has some problems. First, it is ideally suited for spatial processes that are implicitly assumed to be continuous. However, in vector abundance, the response is usually in the form of counts. Second, kriging is not a very useful tool for extrapolation, i.e. for predicting the response in areas outside the current study area. This is because it involves using information from neighbouring sites where the response is known, to *interpolate* values for sites *within the study area*. Thus, for instance, if we wish to use tick data from Kent County, Maryland to predict Lyme disease risk in another area, kriging would not be an appropriate

approach. Regression models, however, allow for *spatial extrapolations*, so that we can predict tick abundance in a new area for which the relevant environmental features are known, but no data on ticks at neighbouring sites are available. Since this study is motivated by the need to develop methods for disease risk extrapolation under climate change scenarios, in this paper we adopt a regression modelling approach that is suitable for spatial extrapolation of count responses.

The standard Poisson regression framework used to model count data (McCullagh and Nelder 1989) assumes equality between the mean and the variance of the response and does not allow for the presence of spatial correlations. These implicit assumptions are restrictive and may not be supported by the data, since spatial dependence may be present in the observed tick abundance rates. This may be due to environmental factors such as land-use features and forest cover that have spatial structure (Cressie and Chan 1989, Breslow and Clayton 1993, Yasui and Lele 1997). If such information is available, identifying and including these factors as covariates in the regression model may account for spatial dependence to some extent. However, even in the absence of a spatial structure, variances frequently exceed the mean (over-dispersion) among outcomes that nominally have a Poisson distribution (Breslow 1984). Ignoring such over-dispersion and correlation produces falsely high precision for the regression coefficients and confidence intervals that are too narrow, leading to inaccurate and biased risk projections.

There are several ways to deal with spatial auto-correlation and over-dispersion. Besag's (1974) approach, based on the auto-Poisson model implies severe restrictions on the parameter space. In fact, the auto-Poisson framework allows for only negative spatial correlations, whereas most geographically distributed data exhibit some degree of clustering (i.e. positive correlation). To allow for positive as well as negative spatial dependence, Cressie and Chan (1989) transformed the count data so that it was approximately normally distributed. Incorporation of a flexible covariance structure into the modelling framework of count data can also be attained using mixed models, which introduce an additional level of variability over standard generalized linear models (e.g. Clayton and Kaldor 1987, Breslow and Clayton 1993).

The central purpose of the spatial modelling exercise undertaken in this study is twofold: (*i*) to identify risk factors for vector abundance, which is a proxy for disease risk, and (*ii*) to enable extrapolation of disease risk. In this study we do this by using a hierarchical Poisson regression model (Breslow and Clayton 1993, McCulloch 1997, Yasui and Lele 1997). This method uses the Poisson framework, which is natural for count data, while the hierarchical aspect allows for the likely presence of (positive) spatial correlation and over-dispersion. Although we need to account for this spatial correlation, it is not the focus of our attention, and is essentially a nuisance parameter that hinders extrapolation of vector abundance to new areas. The regression formulation we adopt here enables us to reduce this spatial correlation by the inclusion of relevant environmental covariates in the model. This approach achieves the twin objectives of identifying risk factors (i.e. environmental covariates) for vector abundance as well as enabling extrapolation through the fitted regression model. In addition, we utilize this framework to devise a novel and intuitively appealing procedure for model selection, which identifies the environmental factors that best explain tick abundance.

2. Methods

2.1. *Model development*

Let n be the number of sampled locations and Y_i be the number of female ticks found at the ith location ($i = 1, \ldots, n$). Let D_i be the number of deer hunted at the

*i*th location. For each location *i*, we denote the true, but unknown, tick abundance, by θ_i. These θ_is are influenced by environmental covariates, denoted by x_i, such as land-use characteristics, vegetation, forest cover and soil type. This relationship is of scientific interest. Unfortunately the θ_is themselves are unobservable, but Y_is, the observed tick counts on deer, can serve as their surrogates.

We model the Y_is using the following two-step hierarchical model. We first relate θ_is to the covariates X. We assume that the vector $\theta=(\theta_1, \ldots, \theta_n)$ has a multivariate log-normal distribution, i.e. $\theta \sim$ log-normal$[X\beta, V(\sigma^2, \rho)]$, where β are regression coefficients, σ^2 is the variance of $\log(\theta_i)$ at each location i and ρ is a measure of spatial dependence. Also, the *(i, j)*-th element of the covariance matrix V is given by $\sigma^2\rho^{d(i,j)}$, where *d(i, j)* is the distance between locations i and j, $(i, j)=1, \ldots, n$. This covariance structure entails some assumptions. First, we assume that the θs have the same variance at all locations. Second, covariance of the process between any two locations is assumed to be isotropic, which means that it just depends on the distance between those two locations. These assumptions are standard in spatial statistics (Cressie 1993, p. 105). Note that, although ρ is restricted to be positive for this particular model, any other structure for $V(\sigma^2, \rho)$ with $\rho<0$, can also be accommodated. The specification of a log-normal distribution for the θs properly forces the abundance rates to be positive. This model also implicitly assumes that the association between θ and the covariates X can be adequately characterized by a linear relationship in the logarithmic scale. However, other forms of relationship may also be entertained quite easily.

In the next stage of modelling, we relate the vector of observed counts Y to θ, the underlying population abundance, by assuming that $(Y_i|\theta_i, D_i) \sim$ Poisson$(D_i\theta_i)$, independent of each other $(i=1, \ldots, n)$. Thus, the average number of deer ticks is $D_i\theta_i$, number of hunted deer times the underlying abundance.

This two-step hierarchical model has several advantages. First, it models the observed number of ticks as Poisson random variables, which is appropriate for count data. Second, it allows for both positive and negative spatial dependence, as well as over-dispersion in the data (Breslow and Clayton 1993, Yasui and Lele 1997). Third, it uses the regression framework, which enables the extrapolation of the response through the fitted model.

Interpretation of regression parameters β in the above Poisson-log normal mixed model is the same as in standard Poisson regression. For any particular covariate, the associated regression coefficient β captures the magnitude and direction of the effect of that covariate on tick abundance. Specifically, $\exp(\beta)$is the *factor* by which abundance of female ticks is expected to increase, for a unit increase in the value of that covariate. This quantity is frequently referred to as a 'rate ratio' (RR) because it is the units-free ratio of two (expected) abundance rates (Breslow and Day 1980).

Point estimates of the regression parameters are expected to be similar to those produced by Poisson regression. However, the Poisson-log normal model has a more general covariance structure that allows for the presence of spatial auto-correlation and over-dispersion in the data. Hence, estimates of the precision of the regression coefficients (i.e. confidence intervals) and predicted disease risk at new locations (i.e. prediction intervals) would be more accurate.

2.2. *Estimation: The Monte Carlo Newton Raphson (MCNR) procedure*

The hierarchical model we have presented here belongs to the class of Generalized Linear Mixed Models (GLMM) (Breslow and Clayton 1993). There are several

different approaches for obtaining estimates of model parameters in a GLMM. Clayton and Kaldor (1987) use the EM Algorithm (Dempster *et al.* 1977), Breslow and Clayton (1993) develop a quasi-likelihood-based approach, while Yasui and Lele (1997) use estimating functions. A fully Bayesian treatment of such models through the use of Markov Chain Monte Carlo (MCMC) techniques, is given by Diggle *et al.* (1998).

A straightforward EM approach for our model entails the calculation of some complicated integrals. Since the Penalized Quasi-Likelihood (PQL) method does not seem to perform well, especially for sparse data (McCulloch 1997), and an estimating functions approach also has similar problems (Yasui and Lele 1997), in this study, we use the Monte Carlo Newton Raphson (MCNR) method proposed by McCulloch (1997) to estimate the model parameters (β, σ^2, ρ).

In this section we provide a brief overview of the MCNR estimation procedure. To do this, first note that the model presented here can be equivalently expressed as $(Y_i|\theta_i)$ being independently distributed Poisson random variables with mean $D_i e^{\theta_i}$, $i=1, 2, \ldots, n$, where the vector $\boldsymbol{\theta}$ has a multivariate normal distribution with mean $\mathbf{X}\beta$ and covariance $V(\sigma^2, \rho)$. Estimation is then based on the likelihood function for the observed Ys, which is given by

$$L(\beta, \sigma^2, \rho|Y)=\int_{\theta} f_{\beta,\sigma^2,\rho}(Y, \theta)\,d\theta=\int_{\theta}\left\{\prod_{i=1}^{n} f_{D_i}(Y_i|\theta_i)\right\}f_{\beta,\sigma^2,\rho}(\theta)\,d\theta \qquad (1)$$

where the distributions of $(Y_i|\theta_i)$ and θ are as above. Since multidimensional integrals can be difficult to work with, we use the full data likelihood instead. Given the full data $W=(Y, \theta)$, the complete data likelihood is

$$L(\beta, \sigma^2, \rho|W)=\left\{\prod_{i=1}^{n} f_{D_i}(Y_i|\theta_i)\right\}f_{\beta,\sigma^2,\rho}(\theta)$$

In the spirit of the EM algorithm (Dempster *et al.* 1977), our estimates of the parameters $\varphi=(\beta, \sigma^2, \rho)$ would be those values of φ which maximize $E[\ln L(\varphi|W)|Y]$, the conditional expectation of the complete data log likelihood, given the available data. In order to use a Newton-Raphson scoring type algorithm for estimating φ, we first note that the maximum likelihood score equations for φ would be

$$E\left[\frac{\partial \ln L(\varphi|W)}{\partial \varphi}\Big|Y\right]=0$$

Taking $\varphi=\beta$, and expanding the quantity inside the conditional expectation as a function of β around the value β_0 gives a scoring-type algorithm of the form

$$\frac{\partial \ln L(\beta|W)}{\partial \beta}\approx X'V^{-1}(\theta-X\beta_0)-X'V^{-1}X(\beta-\beta_0)$$

This gives the stage-m ($m=0, 1, \ldots$) iterative equation

$$\beta_{m+1}=\beta_m+(X'V^{-1}X)^{-1}(X'V^{-1}E[(\theta-X\beta)|Y]). \qquad (2)$$

Since the conditional expectation above may be difficult to evaluate, at each step of iteration we generate B values $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(B)})$ from the conditional distribution of $(\theta|Y, \varphi_m)$ using the Metropolis Algorithm (Hastings 1970). We use McCulloch's recommendation for using the marginal distribution of θ as a proposal distribution

to generate candidate values of θ in the Metropolis sampling, since it considerably simplifies the acceptance function for the Metropolis Algorithm (McCulloch 1997).

Once we have $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(B)})$, we then use $\Sigma_{j=1}^{B}(\theta^{(j)} - X\beta_m)B$ as an approximation to $E[(\theta - X\beta_m)|Y]$. Similarly, we estimate σ^2 and ρ by choosing values for $(\sigma^2_{m+1}, \rho_{m+1})$ that maximize $E[\log L(\varphi|W)|Y]$. In practice, this conditional expectation is again approximated by its Monte Carlo equivalent $\Sigma_{j=1}^{B} \log L(\sigma_m^2, \rho_m|W^{(j)})B$, where $W^{(j)} = (\theta^{(j)}, Y)$. For a fuller discussion of the MCNR estimation procedure, see McCulloch (1997).

The algorithm for MCNR estimation of the parameters for the Poisson log normal model is thus given by:

1. Set initial parameter estimates φ_0. Let $m = 0$.
2. Generate B values $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(B)})$ from the conditional distribution of $(\theta|Y, \varphi_m)$ using the Metropolis Algorithm.
3. Calculate β_{m+1} using (1). Choose $(\sigma^2_{m+1}, \rho_{m+1})$ to maximize $\Sigma_{j=1}^{B} \ln L(\sigma_m^2, \rho_m|W^{(j)})B$, where $W^{(j)} = (\theta^{(j)}, Y)$.
4. Set $m = 1$. Repeat steps 2 to 4. Repeat until convergence.

In our study, we used $B = 1000$, and the MCNR routine usually converged in less than 10 iterations.

2.3. *Selection of covariates*

The selection of covariates to be included in a model is an important issue in any regression approach. This is particularly relevant when GIS databases can provide a plethora of information on any number of environmental factors. We propose a new procedure for selecting covariates in the spatial context. The logic behind our proposal is as follows.

Geographical clustering in spatial data is often due to the presence of some common underlying ecological factors, such as land use patterns, degree of urbanization, etc. that are spatial in nature. In this sense, spatial correlation in the observed data exists because some underlying factor or factors have not been accounted for in the model. As a consequence, explicit inclusion of covariates in the model should reduce spatial dependence. This provides motivation for an intuitive method for selecting covariates for the analysis of spatial data. To identify the factors that are most influential on the observed response, we rank the importance of the covariates by the amount of reduction in spatial dependence it facilitates. Suppose all p available covariates are numbered $1, 2, \ldots, p$. Then we use the following algorithm to select covariates:

Step 1. Fit the Poisson-log normal model to the data without covariates (Model 0) and estimate spatial correlation ρ_{M0}. If this is very small, the data has negligible spatial correlation. If the spatial correlation is substantial, we go to Step 2.

Step 2. Add only covariate 1 to Model 0. Re-estimate the spatial correlation (say, ρ_1). Repeat Step 2, $(p-1)$ times, successively for covariates $2, 3, \ldots, p$, obtaining correlation estimates $\rho_2, \rho_3, \ldots, \rho_p$.

Step 3. Add to the model, covariate j for which the reduction in spatial dependence $\rho_{M0} - \rho_j$ $(j = 1, 2, \ldots, p)$ is the largest. Call this model, Model 1 and obtain an estimate for spatial correlation ρ_{M1} under this model. Number the remaining covariates $1, 2, \ldots, p-1$.

Step 4. Repeat steps 2 to 4 for Model 1, for the remaining $p-1$ covariates, adding that covariate j to the model, for which $\rho_{M1}-\rho_j$ $(j=1, 2, \ldots, p-1)$ is the highest.

Continue this process until the estimate of ρ_{Mj}, $j=0, 1, \ldots, p$, is negligibly small, or all covariates have been added to the model ($j=p$). If the spatial correlation is still significant even after considering all the available covariates, it implies that we should be searching for additional covariates, as well as looking for interactions among the existing ones. To identify the covariates that reduce over-dispersion, we can start with the final model obtained here and repeat all the above steps with σ^2 as the focus of attention.

The method for selecting environmental covariates described above identifies those factors that best explain the spatial structure, or small-scale variation, in the data. By contrast, traditional methods of covariate selection in statistics such as the likelihood ratio (LR), AIC or deviance function, rank the importance of each covariate by the extent to which it reduces overall variability, or large-scale variation in the data, regardless of any embedded spatial structure. Thus, in order to identify covariates that best explain an observed process, both in terms of spatial structure and overall, these two methods may have to be used in conjunction for covariate selection in spatial studies. So, we look upon our method as complementary to the traditional stepwise covariate selection methods. In addition, it provides a tool to distinguish between factors that explain the spatial structure in the data, versus factors that best explain the large-scale variation.

The ultimate objective of our model-building exercise is to extrapolate disease risk/tick abundance at new locations. The principal barrier to such spatial extrapolation outside the study area is the presence of spatial correlation, which is essentially a local phenomenon. This motivates our covariate selection approach. We want to identify the environmental risk factors that best explain the spatial structure in the data, and can be used later for prediction of disease risk at a new location. Thus, our step-wise selection algorithm first concentrates on identifying covariates that reduce the spatial correlation ρ, and then addresses reduction of the over-dispersion parameter σ^2. These methods are illustrated in the next section in the analysis of *I. scapularis* data. A further refinement of our approach that is the subject of ongoing research by the authors would include, at each step of the selection process, a procedure for checking the utility of covariates previously added to the model, and removing those that do not explain spatial correlation or over-dispersion once newer covariates have been introduced. Such an improvement would make this algorithm a truly stepwise (as opposed to forward) selection process.

On a cautionary note, we observe that this method of covariate selection would not be appropriate if the underlying spatial process has intrinsic spatial correlation. For example, a spatial process that models the geographic or familial spread of infectious diseases or agents would have strong intrinsic correlation that is itself of scientific interest (Houwing-Duistermaat *et al.* 1998). However, this is not the case with abundance of *I. scapularis* ticks. The method is, thus, suitable for identifying covariates in this or similar situations.

2.4. *Prediction and model validation*

One of the goals of our study is to use the data to predict tick abundance θ at new locations and provide quantitative measures of the accuracy of such predictions,

i.e. prediction intervals. However, before using the model to predict the abundance θ at a new location, it should be validated for the observed data Y. We can do this by predicting tick counts Y for each of the sampled locations based on data from all the other locations. The procedures for calculating point predictors and prediction intervals for both tick abundance θ and tick count Y are presented in the following discussion.

A fuller treatment of the prediction techniques for obtaining point and interval predictions for tick abundance θ and tick count Y is presented in Das (1998). Here we first note that the prediction techniques for tick abundance θ are different, depending on whether one is interpolating in the study area (where tick counts Y have been observed), or spatially extrapolating in a new area for which tick counts are unavailable. First, we present prediction techniques for interpolation. In this situation, since θs are themselves unobservable, point and interval predictions for the abundance rate θ_0 at a new location s_0 within the study area should be based on the expected value of θ_0, conditional on the observed data Y, i.e. $E(\theta_0|Y)$. Further, to get 95% prediction intervals for θ_0, we need to find (θ_1, θ_u) such that the conditional probability $\Pr(\theta_1 \leqslant \theta_0 \leqslant \theta_u|Y) = 0.95$. In fact, a Monte Carlo approach can avoid tedious calculations for conditional expectations and probabilities. We use the Metropolis Algorithm (Hastings 1970) to generate a batch of realizations of θ_0 from its conditional probability distribution $(\theta_0|Y)$, fixing (β, σ^2, ρ) at their estimated values. If we denote this batch by $\theta_G = (\theta_0^{(1)}, \theta_0^{(2)}, \ldots, \theta_0^{(B)})$, for a sufficiently large B,

$$E(\theta_0|Y) \approx \frac{1}{B} \sum_{j=1}^{B} \theta_0^{(j)},$$

which is our point predictor for θ_0. Similarly, as θ_G are realizations from the distribution of $(\theta_0|Y)$, 95% prediction intervals (θ_1, θ_u) for θ_0 can be approximated by the 2.5 and 97.5 percentile values of the batch θ_G.

Spatial extrapolation of tick abundance outside the study area is feasible only when the inclusion of relevant covariates has reduced spatial correlation to negligible levels. (Note that, in such a situation, we can use the same method for interpolation as for extrapolation.) Since no tick counts are available, here we focus on the marginal distribution of θ. Then, the abundance rate θ_* at a new location s_* outside the study area would be predicted by

$$\hat{\theta}_* = x_*' \hat{\beta}$$

where x_* is the vector of known environmental covariates for location s_*. The associated prediction error would be estimated by

$$E(\theta_* - \hat{\theta}_*)^2 = \mathrm{var}(\theta_*) + x_*' \, \mathrm{var}(\hat{\beta}) x_*.$$

Note that $\mathrm{var}(\theta_*)$ can be estimated by $\hat{\sigma}^2$ and $\mathrm{var}(\hat{\beta})$ by $-E[(\partial^2 \ln L(\beta|Y, \theta)/\partial\beta^2)|Y]$, where $L(\beta|Y, \theta)$ is the likelihood function for the joint distribution of (Y, θ), Y denoting tick counts from the original study area used to fit the model. The latter quantity can be approximated using a Monte Carlo procedure.

To get point predictors for tick counts Y, note that, under the Poisson-log normal model,

$$E(Y_i) = E[E(Y_i|\theta_i)] = E(D_i \theta_i) = D_i \exp(x_i \beta + \sigma^2/2). \tag{3}$$

Suppose Y_0 is the tick count at a new location s_0 with covariates x_0. Then, since

given θ, the Y_is are independent,

$$E(Y_0|Y) = E(E[(Y_0|Y)|\theta]) = E[E(Y_0|\theta)] = D_0 \exp(x_0\beta + \sigma^2/2). \quad (4)$$

We predict Y_0 by plugging in the estimated values of β and σ^2 in equation (4).

Calculation of prediction intervals for the tick counts is trickier, as the precise form of the marginal distribution of Y is difficult to evaluate. We use the following procedure to get prediction intervals for Y_0, given the observed Ys. First, note that, to get these prediction intervals we need to evaluate the conditional probability

$$\Pr(Y_0 \leqslant y|Y) = E[I(Y_0 \leqslant y)|Y]$$

(where $I(Y_0 \leqslant y)$ is an indicator function that has a value of 1 when $Y_0 \leqslant y$, and 0, otherwise)

$$= E[E(I(Y_0 \leqslant y)|Y|\theta_0)] = E[\Pr(Y_0 \leqslant y|Y, \theta_0)] = E[\Pr(Y_0 \leqslant y|\theta_0)]$$

since, given θ_0, Y_0 is distributed as a Poisson $(D_0\theta_0)$ random variable, independently of Y. So,

$$\mathrm{PR}(Y_0 \leqslant y|\theta_0) = \sum_{k=0}^{y} \exp(-D_0\theta_0)\frac{(D_0\theta_0)^k}{k!}. \quad (5)$$

We can again approximate the mean value of this probability through a Monte Carlo procedure. Thus

$$E[\Pr(Y_0 \leqslant y|\theta_0)] = \frac{1}{B}\sum_{j=1}^{B} \Pr(Y_0 \leqslant y|\theta_0^{(j)}),$$

where the form of $\Pr(Y_0 \leqslant y|\theta_0^{(j)})$ is given by equation (5), and $\theta_0^{(1)}, \theta_0^{(2)}, \ldots, \theta_0^{(B)}$ are as before. Once the probabilities $\Pr(Y_0 \leqslant y|Y)$ have been estimated in this way, the approximate lower and upper 95% prediction intervals of Y_0 are those integer values of y, for which this probability exceeds 0.025 and 0.975, respectively.

We can use the spatial interpolation procedures described here for cross-validation, by obtaining 95% prediction intervals of tick counts for each location, based on data from every other location, and finding the proportion of these prediction intervals that actually cover the observed counts. If our model is sensible, this should be close to 95%, or whatever the nominal level of coverage is chosen to be.

3. Tick abundance on deer—a case study

3.1. *The study data*

The purpose of this example is to demonstrate application of the proposed modelling approach to predict adult female *I. scapularis* abundance. The study area covered approximately 6400 km^2, in five counties (Harford, Baltimore, Carroll, Howard and Anne Arundel) along the north-western shore of the Chesapeake Bay in Maryland. Deer killed by hunters during the firearms season of 1991 were examined at hunting check stations for the presence and abundance of ticks by visual examination. Collection methods are described in Glass *et al.* (1994). Briefly, the head, neck, chest and legs of deer were examined using a standardized protocol and ticks from each deer were removed and stored separately. Collected ticks were brought to the laboratory, and identified to species and developmental stage. Hunters were asked to identify the geographic location where each deer was collected. Locations were identified on Maryland State Department of Transportation maps

(1:24 000) and the coordinates determined using the Maryland State Plane Coordinate System (North America Datum 1983).

3.2. *Environmental covariates*

Environmental data related to land use/land cover, soils, watersheds, elevation, slope, aspect and forest distributions were used as predictor variables in the Poisson-log normal model. These data were imported as separate layers in a raster-based geographical information system (GIS) (Eastman 1993). Information on soils was derived from the US Geological Survey's State Soil Geographic database with a reported minimal mapping unit of 625 ha. Land-use and land-cover information was derived from 1985 data from the Maryland Office of Planning as an Anderson level II classification, providing 13 categories of land use (Glass *et al.* 1994). Watersheds were categorized to the sub-basin level. Elevation, slope and aspect were extracted from 7.5-min digital elevation maps produced by the US Geological Survey. Forest distributions were determined from the land-use/land-cover database, modified with Landsat Thematic Mapper images obtained in 1991, using a supervised classification procedure.

We characterized each environmental covariate by creating a 10 km buffer around the collection site of each deer. Although the buffer size was somewhat arbitrary it corresponds to the reported home range of white-tailed deer during this time of year (Swihart *et al.* 1994) and using smaller buffers (5 km) had little impact on the resulting analyses. The GIS was used to determine the amount of the area within the 10 km buffer that corresponded to each class of the recorded environmental covariates. These data were then exported from the GIS to an Splus (MathSoft Inc. 1996) database for statistical analyses.

The major environmental factors (land-use patterns, soil types, etc.) were categorized as percentages (e.g. percentage of land within the buffer of soil type 1, 2, etc.). For example, a particular buffer could have three different land-use patterns, the percentage in each category adding up to 100. Then, as the percentage of land under one category wholly depends on that under the other two (and vice versa), including all three in the model is superfluous, and could cause problems in fitting the model. In this study we have used simple exploratory analyses to avoid such collinearity problems. For each environmental feature, we examined all the different categories that it was divided into, and excluded those that were totally determined by the others. For instance, if we had three land-use categories, we would drop the third since the proportion of land falling under that category can be determined by the information on the other two categories.

4. Results

A total of 574 adult female *I. scapularis* ticks, ranging from 0 to 11 per deer, were collected from 210 deer killed at 119 locations within the 6400 km^2 study area (figure 1). Typically only a single deer was sampled from each site (range 1–5). Generally, most deer were collected in the west and north of the study area, the regions with the lowest human population densities. Exploratory analysis revealed that an unusually large number of deer had been killed in a military base at Fort Meade in Anne Arundel County. Though Fort Meade covers approximately 100 km^2, no details were available as to the precise location of the kills within the Fort area. For this reason, data from this military area were excluded from the analysis.

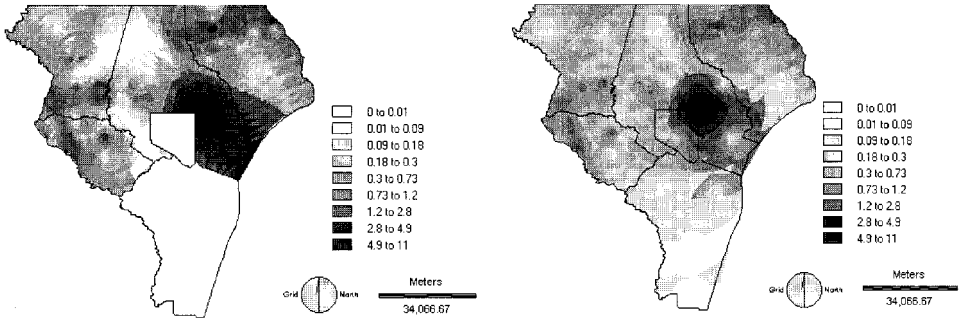

Figure 1. Contour plots showing the geographic distribution of observed number of adult female *Ixodes scapularis* ticks per deer in Howard, Carroll, Baltimore and Harford counties of Maryland in 1991 (on the left), compared to predicted tick abundances for the same areas (on the right). Extraction of relevant covariate information from a GIS makes predictions for neighboring Ann Arundel county and Baltimore City possible, even though no reliable tick counts are available from these places.

4.1. *Selection of environmental factors*

A total of 54 environmental covariates were extracted from the GIS. An ordinary Poisson model was first fit to the data. The relative magnitude of deviance for the model, which is a measure of the excess variability in the data not explained by Poisson regression (McCullagh and Nelder 1989), was about 1.5, indicating the presence of over-dispersion in the data. This failure of the Poisson model assumption is compounded by the likely presence of spatial correlation in the data. As discussed earlier, confidence intervals and prediction intervals based on Poisson regression are thus incorrect. Hence we decided to use the Poisson-log normal model approach to model this data.

We used the algorithm presented in the previous section for selection of covariates. Without any environmental covariates, there was significant spatial correlation (estimated $\rho = 0.78$) in tick abundance. The greatest reduction in spatial correlation of tick abundance (estimated $\rho = 0.47$) was achieved by incorporating the land-use variable measuring the proportion of agricultural land in the 10 km buffer (table 1). The other factor that further reduced the correlation to 0.31 was the land-use variable that measured the proportion of low-density residential housing in the buffer.

Table 1. Stepwise selection of covariates for the Lyme disease tick data: at each step, inclusion of the given covariates caused the maximum reduction in estimated values of, first, the spatial correlation ρ and then, the over-dispersion σ^2.

Covariates[1]	Estimated ρ	Estimated σ^2
None	0.78	
Agricultural land$^{(A)}$	0.47	
(A)+ low density residential housing$^{(B)}$	0.31	1.9
(B)+ flat land$^{(C)}$	0.31	1.89
(C)+ inside forest$^{(D)}$	0.31	1.85
(D)+ slope (0–4°)$^{(E)}$	0.31	1.84
(E)+ forest edge	0	1.71

[1] Percentage of land area having said features (agriculture, low density residential housing, etc.) in a 10 km^2 circle around tick collection site.

The remaining environmental variables failed to reduce the spatial correlation further; so, we concentrated on identifying variables associated with over-dispersion in the data. The environmental variables relating to over-dispersion were land use (low-density housing), slope of the land (flat, and 0–4°), and forest edge (amount of forest edge, and area inside the forest). Inclusion of the proportion of flat land at each location was most effective in reducing the over-dispersion. Finally, when all the above covariates were included in the model, addition of the proportion of land that represented the edge of forested areas reduced the spatial correlation to zero (table 1). No other environmental factor had any appreciable effect on reducing the over-dispersion. Thus, the final model had no remaining spatial correlation (as all the important covariates that could induce spatial dependence had been included) but extra-Poisson variability (estimated $\sigma^2 = 1.71$) remained in the data (table 1).

4.2. *Interpretation of regression coefficients*

Estimates of the rate ratios for the environmental covariates in our model (table 2) indicated that tick abundance was positively related to the slope of the landscape. Abundance nearly doubled for a one-point increase in the percentage of moderately sloping land within the buffer (rate ratio (RR) = 1.95; 95% confidence interval (CI) 0.91 to 4.2). The amount of forest edge within the buffer also was associated with increased abundance of female ticks (RR = 1.3; 95% CI 0.98 to 1.76). Both factors were marginally statistically significant at the 5% level, and the latter accounted for a large degree of the spatial correlation in the data (table 1).

Conversely, the proportion of land that was flat was negatively associated with tick density; for every point decrease in the percentage of flat land, tick abundance rates increased by almost 50% (RR = 0.67, 95% CI 0.49 to 0.92). Similarly, agricultural land had an inverse relationship with tick abundance. A unit decrease in the percentage of agricultural land in the buffer was associated with a 7% increase in female tick abundance (RR = 0.93; 95% CI 0.875 to 0.98). Increasing amounts of contiguous forest within the buffer (or, area inside forest, as opposed to forest edge), also was associated with a somewhat lower tick abundance (RR = 0.88; 95% CI 0.75 to 1.03), as were more areas with low density residential housing (RR = 0.80; 95% CI 0.66 to 0.98).

The results obtained here are biologically plausible and intuitively appealing.

Table 2. Effect of environmental factors on tick abundance: estimates of regression coefficients and their precision for the Lyme disease tick data.

Covariates[1]	Rate ratio[2]	Standard error	95% confidence intervals
Agricultural land	0.93	0.03	(0.875, 0.98)
Low residential housing	0.8	0.1	(0.66, 0.98)
Flat land	0.67	0.16	(0.49, 0.92)
Inside forest	0.88	0.08	(0.75, 1.03)
Slope (0–4°)	1.95	0.39	(0.91, 4.2)
Forest edge	1.31	0.15	(0.98, 1.76)

[1] Percentage of land area having said features (agriculture, low density residential housing, etc.) in a 10 km^2 circle around tick collection site.

[2] The Rate Ratio is given by $\exp(\beta)$, where β is the regression coefficient. Being the *factor* by which abundance of female ticks increases, per unit increase in the value of a covariate, it is free of units.

Published literature on the epidemiology of Lyme disease and the biology of *I. scapularis* is consistent with the patterns identified by the model. Maupin *et al.* (1991) showed that tick abundance decreased significantly within the suburban landscape along transects as one moves from the forest edge to ornamental vegetation and lawns. Similarly, agricultural land has been found to be unsuitable for *I. scapularis* populations (Kitron *et al.* 1992, Glass *et al.* 1994). The basis for this may be related to environmental effects, which increase the rate of mortality through desiccation (Bertrand and Wilson 1996) and physical disturbance of the habitat in these areas. The marked difference in the effect of forest edge (ecotone) and interior forest area on adult tick abundance is striking. The biological mechanism for this is unclear; however, ecotonal habitat is well recognized as a critical factor in the abundance of white-tailed deer (Nixon *et al.* 1991) which influences tick abundance (Wilson *et al.* 1985). Additionally, Goddard (1992) found that *I.scapularis* was absent from habitat in which there was either '… no shade or in totally shaded areas' (p. 503), while ticks tended to reside in areas with 30–80% mixed shade, which occurs along the forest edge. Thus, the amount of completely forested area, as opposed to forest ecotone, may be associated with lower numbers of adult *I. scapularis*. The effect of the steepness of land on adult tick abundance is not evident but has been previously noted by other authors (Glass *et al.* 1995).

4.3. *Cross-Validation and Prediction*

To cross-validate our model, we predicted tick counts for each of the 118 sampled locations, based on data from the remaining 117 locations. The 95% prediction intervals included the observed values at these locations in over 94% (112/118) of the cases, thus satisfactorily validating the model in this region. We note that this method of cross-validation may be very computer intensive for large data. In our case, for a sample size of 118, it took about 10 hours on a UNIX machine. For large data sets, an alternative would be to set aside a portion of the data for prediction. The model fitted on the reduced data can then be validated on the prediction subset by examining the proportion of data points in this subset that were included in the calculated 95% prediction intervals.

An important advantage of our modelling approach is that we can use the fitted model for extrapolating tick abundance in a different area. Since spatial correlation is absent in the fitted model, this does not require tick counts from such an area, only information on the environmental features that were previously identified as crucial in explaining the spatial structure of the data. In figure 1 we present contour plots of the geographic distribution of observed tick counts per deer, and predicted tick abundance in the whole study area, as well as Ann Arundel county and Baltimore City. Extraction of relevant covariate information from a GIS made predictions for Ann Arundel county and Baltimore City possible, even though no reliable tick counts were available from these places. A comparison of the observed and predicted tick abundance, for the regions where tick counts were actually observed, (figure 1) shows remarkable consistency between the two. Predictions for Ann Arundel county and Baltimore City are generally low because both have a high preponderance of flat, non-sloping land and the latter is an urban area with no forests in the vicinity. Previously, we saw that both these factors are inversely related to tick abundance.

5. Conclusion

The geographical distribution of vector-borne diseases is determined by vector abundance, which in turn is affected by environmental factors. Combining an

environmental GIS with spatial analysis can therefore aid in predicting vector abundance, which in turn can help in predicting disease risk. Public education remains one essential means of preventing diseases; knowledge of the environmental factors that help or hinder the distribution of vectors could enable public health professionals to anticipate, take preventive steps and react quickly to new outbreaks. In this context, it is important that we have the statistical tools to identify the risk factors for vector abundance and use this knowledge for extrapolating disease risk. To this end we have presented a regression-based approach for spatial modelling of count data. There are some key advantages to this approach:

(*a*) Our method provides an alternative to the standard kriging and neighbourhood-based formulations of spatial statistics. This regression approach facilitates reduction of spatial correlations by inclusion of proper environmental covariates, and, in contrast to the standard methods, allows for extrapolation of vector abundance to new areas.

(*b*) Our mixed models formulation avoids restrictive assumptions and provides a flexible way of incorporating spatial correlations and over-dispersion.

(*c*) Though GIS can be a useful tool (Baker 1992, Pope *et al.* 1994, Dunning *et al.* 1995), such databases contain a huge number of environmental factors. It is thus essential to develop a procedure that identifies factors that explain the spatial structure in the data. We propose a stepwise procedure, where at each step we include that covariate that best explains the spatial dependence and variability in the data.

(*d*) The principles underlying the modelling framework and inference procedures used in this paper are applicable in more general settings. For example, if the responses were presence/absence of species instead of tick counts, we could have used a Binomial normal mixed model in place of our Poisson log-normal model. Given the prevalence of spatial data in ecological studies that are of the presence/absence variety, or counts, we believe the methods proposed in this paper are of significant importance to ecological research in general.

Acknowledgements

This work was part of the PhD dissertation of the first author. Partial funding support was provided by EPA cooperative agreement CR823143 to Jonathan Patz and CDC Research Contract 200-94-0818 to Gregory E. Glass. The authors also thank Professor Peter J. Diggle for helpful comments on an earlier version of the manuscript.

References

Apperson, C. S., Levine, J. F., and Nicholson, W. L., 1990, Geograhic occurrence of *Ixodes scapularis* and *Amblyomma americanum* (Acari: Ixodidae) infesting white-tailed deer in North Carolina. *Journal of Wildlife Diseases*, **26**, 550–553.

Baker, W. L., 1992, Effects of settlement and fire suppression on landscape structure. *Ecology*, **73**, 1879–1887.

Bertrand, M. R., and Wilson, M. L., 1996, Microclimate-dependent survival of unfed adult *Ixodes scapularis* (Acari: Ixodidae) in nature: life cycle and study design implications. *Journal of Medical Entomology*, **33**, 619–627.

Besag, J. E., 1974, Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, Series B, **35**, 192–236.

Breslow, N. E., 1984, Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

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

Breslow, N. E., and Day, N. E., 1980, *The Design and Analysis of Cohort Studies* (Lyon: International Agency for Research on Cancer).

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

Cressie, N., 1993, *Statistics for Spatial Data* (New York: John Wiley).

Cressie, N., and Chan, N. H., 1989, Spatial modeling of regional variables. *Journal of the American Statistical Association*, **84**, 393–401.

Das, A., 1998, Topics in Spatial Statistics. Ph.D. Dissertation, Department of Biostatistics, Johns Hopkins University School of Hygiene & Public Health, Baltimore.

Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977, Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, **44**, 1–38.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A., 1998, Model-based Geostatistics (with discussion). *Journal of the Royal Statistical Society*, Series C, **47**, 299–350.

Dunning, J. B., Stewart, D. J., Danielson, B. J., Noon, B. R., Root, T. L., Lamberson, R. H., and Stevens, E. E., 1995, Spatially explicit population models: current forms and future uses (in spatially explicit population models). *Ecological Applications*, **5**, 3–11.

Eastman, J. R., 1993, *IDRISI. Version 4.1* (Worcester, MA: Clark University).

Glass, G. E., Amerasinghe, F. P., Morgan, J. M., and Scott, T. W., 1994, Predicting *Ixodes scapularis* abundance on white tailed deer using geographic information systems. *American Journal of Tropical Medicine and Hygiene*, **51**, 538–544.

Glass, G. E., Schwartz, B. S., Morgan, J. M. III, Johnson, D. T., Noy, P. M., and Israel, E., 1995, Environmental risk factors for Lyme disease identified with geographic information systems. *American Journal of Public Health,* **85**, 944–948.

Goddard, J., 1992, Ecological studies of adult *Ixodes scapularis* in central Mississippi: questing activity in relation to time of year, vegetation type, and meterologic conditions. *Journal of Medical Entomology*, **29**, 501–506.

Hastings, W. K., 1970, Monte Carlo sampling methods using Marlov chains and their applications. *Biometrika*, **57**, 97–109.

Houwing-Duistermaat, J. J., Van Houwelingen, H. C., and Terhell, A., 1998, Modelling the cause of dependency with application to filaria infection. *Statistics in Medicine*, **17**, 2939–2954.

Kitron, U., Jones, C. J., Bouseman, J. K., and Baumgartner, D. L., 1992, Spatial analysis of the distribution of *Ixodes dammini* (Acari: Ixodidae) on white-tailed deer in Ogle county, Illinois. *Journal of Medical Entomology*, **29**, 259–266.

Lastavica, C., Wilson, M. L., Berardi, V. P., Spielman, A., and Deblinger, R. D., 1989, Rapid emergence of focal epidemic of Lyme disease in coastal Massachusetts. *New England Journal of Medicine*, **320**, 133–137.

MathSoft Inc., 1996, Splus Version 3.3.

Maupin, G. O., Fish, D., Zultowsky, J., Campos, E. G., and Piesman, J., 1991, Landscape ecology of Lyme disease in a residential area of Westchester County, New York. *American Journal of Epidemiology*, **133**, 1105–1113.

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

McCulloch, C. E., 1997, Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162–170.

Nixon, C. M., Hansen, L. P., Brewer, P. A., and Chelsvig, J. E., 1991, Ecology of the white-tailed deer in an intensively farmed region of Illinois. *Wildlife Monographs*, **118**.

Patz, J. A., Epstein, P. R., Burke, T. A., and Balbus, J. M., 1996, Global climate change and emerging infectious diseases. *Journal of the American Medical Association*, **275**, 217–223.

Pope, K. O., Rejmankova, E., Savage, H. M., Arredondo-Jimenez, J. I., Rodriguez, M. H., and Roberts, D. R., 1994, Remote sensing of tropical wetlands for malaria control in Chiapas, Mexico. *Ecological Applications*, **4**, 81–90.

Ribeiro, J. M., Seulu, F., Abose, T., Kidane, G., and Teklehaimanot, A., 1996, Temporal and spatial distribution of anopheline mosquitoes in an Ethiopian village: implications for malaria control strategies. *Bulletin of the World Health Organization*, **74**, 299–305.

Solberg, V. B., Olson, J. G., Boobar, L. R., Burge, J. R., and Lawyer, P. G., 1995, Prevalence of *Ehrlichia chaffeensis*, spotted fever group rickettsia, and *Borrelia* spp. infections in ticks and rodents at Fort Bragg, North Carolina. *Journal of Vector Ecology*, **21**, 81–84.

Steere, A. C., 1989, Lyme Disease. *New England Journal of Medicine*, **321**, 586–596.

Swihart, R. K., Picone, P. M., DeNicola, A. J., and Cornicelli, L., 1994, Ecology of urban and suburban white tailed deer. In *Urban Deer A Manageable Resource?*, edited by J. McAninch and L. P. Hansen (Berlin: Springer Verlag), pp. 35–44.

Wallis, R. C., Brown, S. E., Kloter, K. O., and Main, A. J., 1978, Erythema chronicum migrans and Lyme arthritis: field study of ticks. *American Journal of Epidemiology*, **108**, 322–327.

Wilson, M. L., Adler, G. H., and Spielman, A., 1985, Correlation between abundance of deer and that of the deer tick *Ixodes dammini* (Acari: Ixodidae). *Annals of the Entomological Society of America*, **78**, 172–176.

Yasui, Y., and Lele, S., 1997, A regression method for spatial disease rates: an estimating function approach. *Journal of the American Statistical Association*, **92**, 21–32.