Appl. Statist. (2002)
51, *Part* 4, *pp.* 493–506

Childhood malaria in the Gambia: a case-study in model-based geostatistics

Peter Diggle,

Lancaster University, UK

Rana Moyeed,

University of Plymouth, UK

Barry Rowlingson

Lancaster University, UK

and Madeleine Thomson

Liverpool University, UK

[Received January 2001. Final revision June 2002]

Summary. The paper develops a spatial generalized linear mixed model to describe the variation in the prevalence of malaria among a sample of village resident children in the Gambia. The response from each child is a binary indicator of the presence of malarial parasites in a blood sample. The model includes terms for the effects of child level covariates (age and bed net use), village level covariates (inclusion or exclusion from the primary health care system and greenness of surrounding vegetation as derived from satellite information) and separate components for residual spatial and non-spatial extrabinomial variation. The results confirm and quantify the progressive increase in prevalence with age, and the protective effects of bed nets. They also show that the extrabinomial variation is spatially structured, suggesting an environmental effect rather than variation in familial susceptibility. Neither inclusion in the primary health care system nor the greenness of the surrounding vegetation appeared to affect the prevalence of malaria. The method of inference was Bayesian using vague priors and a Markov chain Monte Carlo implementation.

Keywords: Epidemiology; Extrabinomial variation; Geostatistics; Insecticide-treated bed nets; Malaria; Satellite data; Spatial statistics

1. Introduction

Malaria is a major public health issue in much of the developing world. The mapping of variations in risk of malaria can help to improve the targeting of scarce resources for public health interventions. However, direct mapping of relevant environmental risk factors (which may vary considerably in both space and time) is impractical and this has led to investigations of whether environmental proxies obtained from routinely available satellite data can be used instead (Thomson *et al.*, 1996). Satellite-derived indices which can be used in this way include measures of rainfall, surface temperature and greenness of vegetation (Connor *et al.*, 1998).

Address for correspondence: Peter Diggle, Medical Statistics Unit, Department of Mathematics and Statistics, Faculty of Applied Sciences, Lancaster University, Lancaster, LA1 4YF, UK.
E-mail: P.Diggle@lancaster.ac.uk

 0035–9254/02/51493

The use of satellite data in the development of predictive models of the distribution of vector-borne diseases, such as malaria, has expanded rapidly in recent years (Thomson and Connor, 2000). However, few attempts have been made to address explicitly the problems of spatial autocorrelation in the interpretation of results (Robinson, 2000). Thomson *et al.* (1999) reported a regression analysis of the prevalence of malaria among children in 65 villages in the Gambia, including as child level covariates age and bed net use and, as village level covariates, whether or not the village belonged to the primary health care (PHC) structure of the Ministry of Health as well as a satellite-derived measure of seasonal greenness of the village environment. Their analysis showed strong extrabinomial variation which they accommodated in a marginal regression analysis using generalized estimating equations (Liang and Zeger, 1986) with a working covariance matrix derived from an assumed spatial correlation model. The marginal modelling approach essentially treats spatial variation in prevalence as a nuisance effect. It leads to asymptotically valid inference about marginal regression parameters under minimal assumptions but does not seek to explain any underlying spatial variation in prevalence; nor does it allow a smooth interpolation of residual spatial variation.

Our objective in this paper is to demonstrate how the model-based geostatistics framework of Diggle *et al.* (1998) can be adapted to extend the analysis of Thomson *et al.* (1999), to provide an explanation of the residual extrabinomial variation in the data and in particular to assess whether the extrabinomial variation is spatially structured. If so (meaning that adjusted prevalences are similar among neighbouring villages), then the explanation must be at least partly environmental. If not, a more likely explanation is that the extrabinomial variation is induced by variation in unmeasured, village-specific factors. For example, an unmeasured characteristic that is shared by members of an extended family within a village would induce non-spatial extrabinomial variation.

The basic ingredients of the modelling framework given by Diggle *et al.* (1998) are the following. Responses $Y_i, i = 1, \ldots, n$, are associated with covariate vectors z_i and spatial locations x_i. Conditionally on an unobserved spatial stochastic process $S(x)$, the responses Y_i are mutually independent and follow a generalized linear model (McCullagh and Nelder, 1989) with z_i as covariates and $S(x_i)$ as an offset. In the current application, we have multiple binary responses at each location (village), enabling the inclusion of an additional, non-spatially structured source of random variation between locations. We also use a different covariance model for $S(x)$, and the details of our Markov chain Monte Carlo (MCMC) implementation also differ from those in Diggle *et al.* (1998).

Section 2 describes the data set in more detail and formulates the particular model which we shall use for analysis. Section 3 describes the resulting analysis of the data. Section 4 compares our results with those obtained by Thomson *et al.* (1999) and discusses the wider potential of the model-based geostatistical approach for analysing data of this kind.

2. Model formulation

The data are obtained from samples of children in each of 65 villages in the Gambia, as indicated in Fig. 1.

These data were collected during the second year (1992) of a study designed to measure the effectiveness of the National Impregnated Bednet Programme in reducing child morbidity and mortality (D'Alessandro *et al.*, 1995). Five areas (two in the west, one in the central area and two in the east) were chosen as sentinel sites to reflect the varied cultural and ecological settings within the Gambia and all 42 PHC villages in these areas were included in the study. A further 23 non-PHC villages were randomly selected from areas 1, 3 and 5 to increase the sample size.

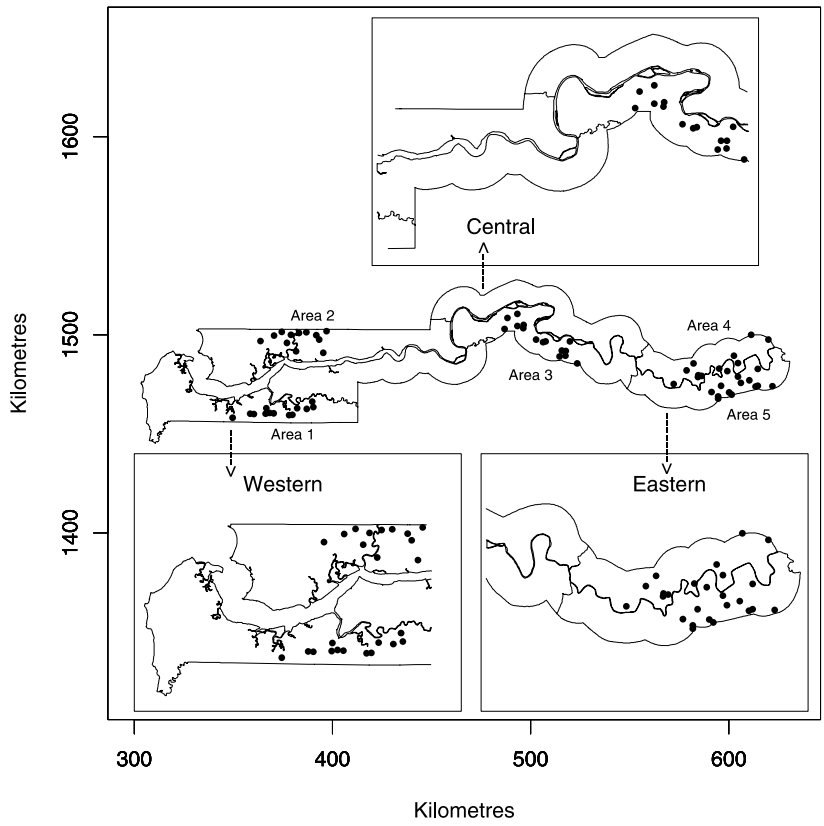

Fig. 1. Map of the Gambia with the 65 villages marked: the inset plots show in enlarged form the corresponding regions indicated by the arrows

For the jth child in the ith village, the response Y_{ij} is a binary indicator of the presence of malarial parasites in a blood sample. The associated covariate vector z_{ij} includes the age of the child in years, whether or not they regularly slept under a bed net, and if so whether the bed net was treated (with permethrin insecticide at a concentration of approximately 200 mg m^{-2}), a binary indicator of whether or not the village belongs to the PHC system, a five-level factor identifying the five sampling areas and a continuous measure of greenness derived from satellite data, namely the seasonal sum of the normalized difference vegetation index (Thomson *et al.*, 1999); note that the last two of these covariates are common to all children within a village.

To account for unexplained spatial variation, we postulate a stationary Gaussian process $S(x)$ with mean 0, variance σ^2 and correlation function $\rho(u) = \mathrm{corr}\{S(x), S(x')\}$, where u is the distance between locations x and x'. A general issue in models of this kind concerns the choice of parametric family for $\rho(u)$ (see, for example, Chiles and Delfiner (1999)). In principle, this should be chosen to give a good fit to the data. In most applications, it is reasonable to assume that $\rho(u)$ is monotone non-increasing in u, with a scale parameter ϕ which controls the rate at which the correlation approaches 0 with increasing u. Thus, $\rho(u) = \rho_0(u/\phi)$. As a general empirical model we favour the Matérn class of correlation functions, which have the form

$$\rho_0(u) = \frac{1}{2^{\delta-1}\,\Gamma(\delta)} u^{\delta}\, \mathcal{K}_{\delta}(u), \tag{1}$$

where $\mathcal{K}_{\delta}(\cdot)$ is the modified Bessel function of the second kind of order δ, and $\delta > 0$ is a smooth-

ness parameter. A process with this correlation function is ceiling(δ) $-$ 1 times mean squared differentiable. Thus the realizations will be mean squared continuous if $\delta \leqslant 1$, once differentiable if $1 < \delta \leqslant 2$, etc. For further details, see Matérn (1986), Handcock and Wallis (1994) and Stein (1999).

The Matérn model includes as special cases the exponential model, $\rho_0(u) = \exp(-u)$, when $\delta = 0.5$, and the Gaussian model, $\rho_0(u) = \exp(-u^2)$, as $\delta \to \infty$, corresponding respectively to mean-square continuous but non-differentiable $S(x)$ and to infinitely mean-square differentiable $S(x)$. The latter is probably unrealistically smooth for most applications and leads to very ill-conditioned covariance matrices for vectors of values $S = (S(x_1), \ldots, S(x_n))$. In a given application we can formally estimate δ from the data. However, there is some advantage to restricting its range to avoid the ill conditioning which results with large values of δ, and in a Bayesian context there is a computational saving, and essentially no loss of interpretive value, if we adopt a discrete prior for δ. For example, choosing between $\delta = 0.5, \delta = 1.5$ or $\delta = 2.5$ corresponds to choosing between mean-square continuous, once-differentiable and twice-differentiable processes $S(x)$.

Diggle *et al.* (1998) used a powered exponential family, $\rho_0(u) = \exp(-u^\delta)$ with $0 < \delta \leqslant 2$. This produces functions with a qualitatively similar shape to the Matérn family, but the parameter δ lacks a natural physical interpretation, since the underlying process $S(x)$ is mean square continuous but non-differentiable for all $0 < \delta < 2$, and infinitely mean square differentiable for $\delta = 2$.

To account for unexplained non-spatial variation, we postulate a set of village-specific random effects U_i, which are mutually independent and Gaussian, with mean 0 and variance ν^2.

Conditionally on $S(x)$ and on the U_i, the Y_{ij} are then modelled as independent Bernoulli variates with success probabilities p_{ij} given by

$$\log\{p_{ij}/(1 - p_{ij})\} = \alpha + \beta' z_{ij} + U_i + S(x_i). \tag{2}$$

This is an example of a generalized linear mixed model, as reviewed by Breslow and Clayton (1993). Besag *et al.* (1991) used a similar model, except that in place of the continuous spatial process $S(x)$ they used a Markov random field defined on a finite set of locations x_i. We take the point of view that the prevalence of malaria is, in effect, a continuous spatial phenomenon, of which the villages in the survey (which are only a small fraction of the totality of settlements in the Gambia) provide an opportunistic sample, and that a continuous spatial process is therefore a more appropriate modelling framework in this application. More pragmatically, one of our objectives is to map the unexplained variation in the prevalence of malaria, and for this we need a continuous spatial model. Nevertheless, we acknowledge that the stationary Gaussian process is no more than a means to an end, and that a discrete spatial process such as a Markov random field could be used, for example, to construct choropleth maps on regions defined as neighbourhoods of each village in the sample. For further comments on the relative merits of continuous and discrete spatial variation models, see for example the discussions on Diggle *et al.* (1998), Besag *et al.* (1991) and Besag and Higdon (1999).

One particular issue which arises with model (2) is whether both the U_i- and the $S(x_i)$-terms are separately identifiable. The vector T, where $T_i = U_i + S(x_i)$, is multivariate Gaussian with covariance matrix $\nu^2 I + \sigma^2 R$, where R has ijth element $\rho(u_{ij}; \phi)$ and u_{ij} is the distance between x_i and x_j. This corresponds to a stochastic process $T(x)$ whose correlation function has a discontinuity at the origin. Given a parametric specification for $\rho(u)$ which is continuous at the origin, the parameters ν^2, σ^2 and ϕ are formally identifiable, but the extent to which they are well identified depends on the configuration of sampling locations x_i. Unless the sampling design

includes essentially coincident pairs of locations, parameter estimation involves extrapolation to the origin of smooth fitted behaviour away from the origin. The problem is exacerbated by the fact that we do not observe the T_i directly, only a set of binary outcomes (albeit well replicated at each location) conditional on the T_i.

3. Statistical analysis

3.1. Methods

We write $\mathbf{S} = (S(x_1), \ldots, S(x_n))$, $\mathbf{U} = (U_1, \ldots, U_n)$ and $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \nu^2, \sigma^2, \phi, \delta)$. In this application, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)$, where β_1 is the regression coefficient corresponding to the age of the child (in days), β_2 and β_3 measure respectively the effect of using untreated bed nets and the additional effect of using bed nets which are treated, β_4 is the regression coefficient corresponding to the greenness index, β_5 measures the effect of living in a PHC village and β_6, β_7, β_8 and β_9 are regression coefficients measuring the effects of areas 2, 3, 4 and 5 respectively in relation to area 1, the assumed base level.

We use Bayesian inference, implemented via an MCMC algorithm. We therefore need to specify prior distributions for the elements of $\boldsymbol{\theta}$. We chose independent improper uniform priors for α and the components of $\boldsymbol{\beta}$. For the parameters ν^2, σ^2 and ϕ we adopted the following vague priors: $f(\nu^2) \propto 1/\nu^2$; $f(\sigma^2) \propto 1/\sigma^2$; $f(\phi) \propto 1/\phi^2$. In the case of the variance components ν^2 and σ^2, these are the Jeffreys priors. As discussed earlier, it is reasonable to restrict the prior range of δ to avoid the problems of near singularity of the covariance matrix of $\mathbf{S}$ which arise when δ is large. We therefore used as prior for δ a uniform distribution on the interval (0, 3).

In the absence of any prior knowledge about the model parameters, the choice of non-informative improper priors is dictated by pragmatic considerations. We acknowledge that improper priors can sometimes give rise to improper posteriors. In our case, it is difficult to ascertain theoretically whether our priors yield a proper posterior. However, if the posterior is improper, then the MCMC algorithm would fail to converge. We carried out several very large MCMC runs with different starting values, but we did not encounter any convergence problems with the priors that we have used. As an example, our MCMC simulations failed to converge for the priors $f(\phi) \propto 1$ or $f(\phi) \propto 1/\phi$, whereas the choice $f(\phi) \propto 1/\phi^2$ caused no such problems.

Using Bayes's theorem, it is relatively straightforward to write down the posterior distributions of $\boldsymbol{\theta}, \mathbf{S}$ and $\mathbf{U}$ given the data. The steps for updating the signals and all the model parameters are described in detail in Diggle *et al.* (1998). The only major difference here stems from the fact that we have repeated observations, i.e. n_i measurements at each location (village) x_i. The updating of the regression parameters, whether relating to individual level or village level covariates, requires the use of the full likelihood.

We simulated realizations from the posterior distributions by means of a single-component Metropolis–Hastings algorithm. The parameters $\alpha, \boldsymbol{\beta}, \nu^2, \sigma^2$ and ϕ were updated by using a random-walk Metropolis algorithm with a Gaussian proposal density for α, $\boldsymbol{\beta}$, $\log(\nu^2)$, $\log(\sigma^2)$ and $\log(\phi)$, and a uniform proposal on (0, 3) for δ.

At each iteration, the updating of ϕ and δ requires the inversion of the $n \times n$ covariance matrix, where n is the number of villages. In our example, we have 2035 observations spread over $n = 65$ villages. The Markov chain was run for 60000 iterations, with convergence judged to have occurred after 10000 samples, on the basis of inspection of sample traces. The chain was thereafter sampled on every 10th iteration to yield a sample of 5000 values from the posterior for each of the elements of $\boldsymbol{\theta}, \mathbf{S}$ and $\mathbf{U}$.

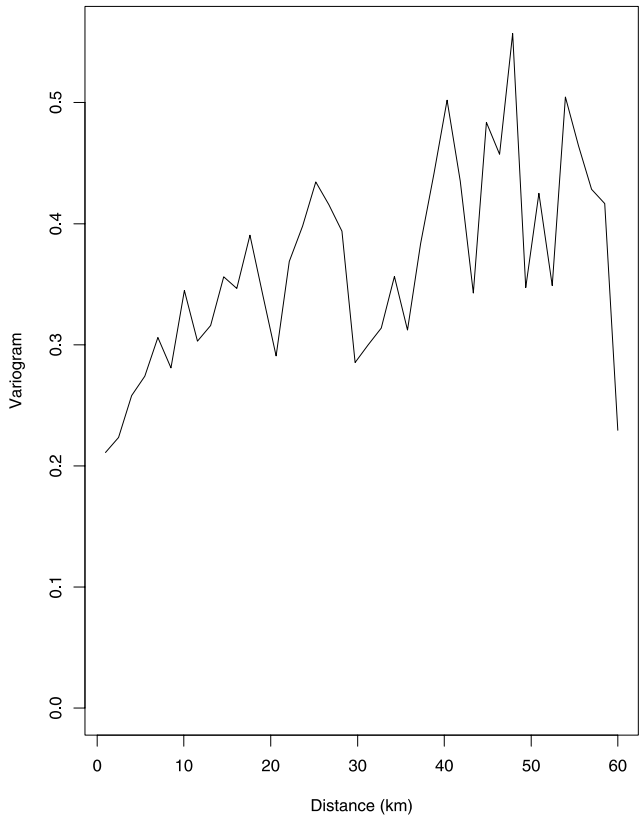

Fig. 2. Empirical variogram of the posterior mean estimates of village-specific random effects U_i

Diagnostic checking of the model is based on two different definitions of village level residuals, as follows. Firstly, under the assumed model (2) we define $\hat{p}_{ij}$ to be the expectation of Y_{ij}, with all parameter values set to their posterior means; following Zeger *et al.* (1988), the computation of each $E(Y_{ij})$ used the cumulative Gaussian approximation to the logistic function given in Johnson and Kotz (1970), page 6. We then defined $r_{ij} = (Y_{ij} - \hat{p}_{ij})/\sqrt{\{\hat{p}_{ij}(1 - \hat{p}_{ij})\}}$ and calculated the residual for the ith village as

$$r_i = \sum_{j=1}^{n_i} r_{ij}/\sqrt{n_i},$$

where n_i is the number of children sampled in the ith village. The corresponding fitted value was calculated as

$$f_i = \sum_{j=1}^{n_i} \hat{p}_{ij}/n_i.$$

Under a well fitting model for the expectations, a plot of residuals against fitted values should show no obvious relationship.

As a check on the fitted second-moment structure, we calculated a second set of village level residuals designed to eliminate both covariate effects and the fitted latent stochastic process. To

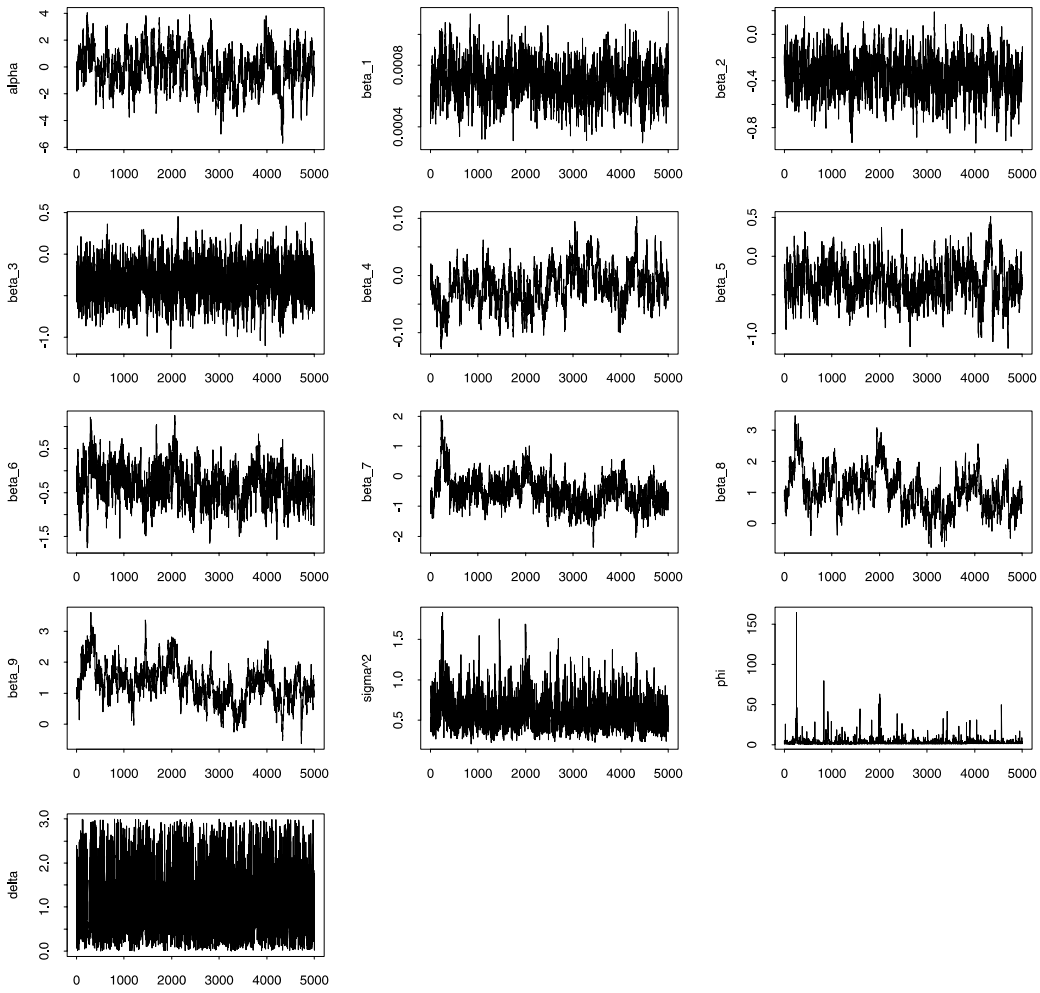

Fig. 3. Time series plots showing the MCMC output every 10th iteration after burn-in

achieve this, we replace the fitted expectations $\hat{p}_{ij}$ used above by quantities p_{ij}^* defined by

$$\text{logit}(p_{ij}^*) = \hat{\alpha} + z_{ij}'\hat{\beta} + \hat{U}_i + \hat{S}(x_i),$$

where, in each case, the 'hat' notation signifies the posterior mean value of the corresponding quantity. In a well fitting model, the resulting residuals, r_i^* say, should have approximately zero mean and unit variance and show no spatial structure.

3.2. Results

When we attempted to fit the full model (2), we found that there was not enough information in the data to estimate both ν^2 and σ^2 simultaneously. Various reparameterizations were tried, but these made no difference.

We therefore considered removing the spatial component $S(x)$, thereby reducing the model to a standard, non-spatial generalized linear mixed model. However, the estimates of the presumed spatially uncorrelated village level random effects in fact showed strong evidence of spatial dependence. This is indicated in Fig. 2, which plots the empirical variogram of the posterior

Table 1. Estimates and 95% intervals for the parameters of model (3)

Parameter	*2.5% quantile*	*97.5% quantile*	*Mean*	*Median*
α	−2.966473	2.624348	−0.131214	−0.077961
β_1 (age)	0.000455	0.000933	0.000689	0.000685
β_2 (untreated)	−0.673143	−0.042011	−0.357825	−0.359426
β_3 (treated)	−0.753803	0.088418	−0.32954	−0.325853
β_4 (greenness)	−0.085675	0.047924	−0.020068	−0.020834
β_5 (PHC)	−0.787913	0.129883	−0.344846	−0.349915
β_6 (area 2)	−1.14419	0.51023	−0.324665	−0.331634
β_7 (area 3)	−1.40862	0.558616	−0.5321	−0.559229
β_8 (area 4)	−0.109472	2.425342	1.049441	1.016969
β_9 (area 5)	0.164828	2.606357	1.309553	1.325129
σ^2	0.311756	1.050227	0.585592	0.553477
ϕ	0.915789	10.20069	2.522294	1.422975
δ	0.079522	2.784646	1.084108	0.937436

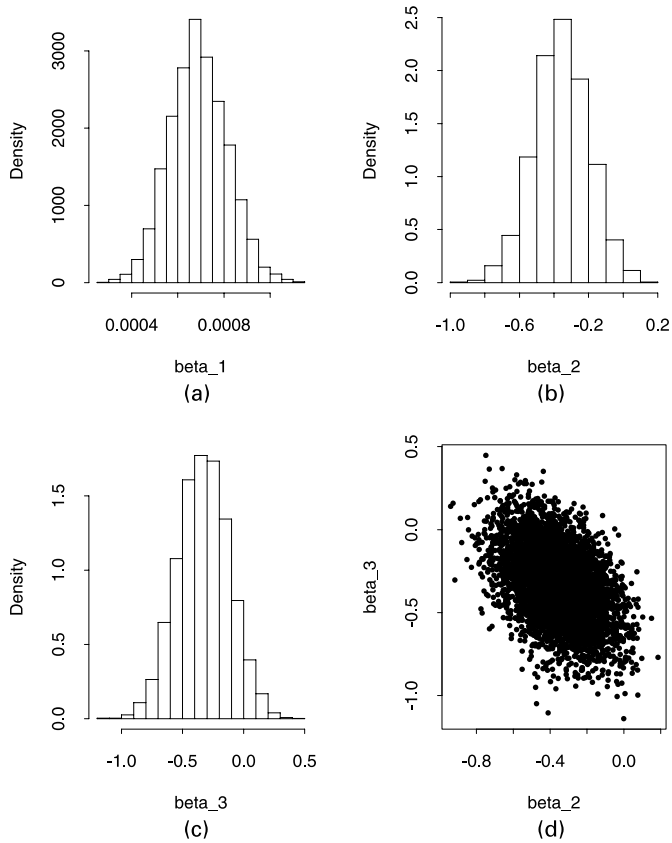

Fig. 4. Empirical posterior distributions for the regression parameters β_1 (age effect), β_2 and β_3 (bed net effects)

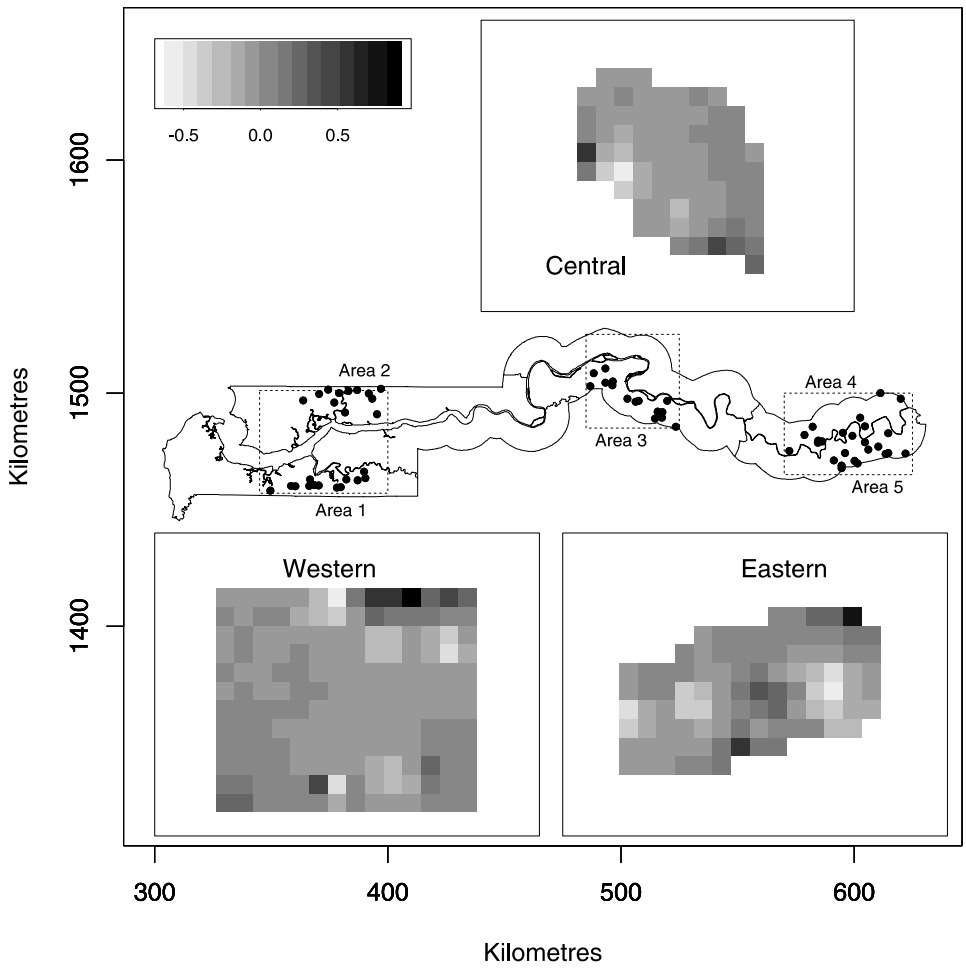

Fig. 5. Maps of the predicted (posterior mean) surface of $\hat{S}(x)$ shown in enlarged form in the inset plots for the regions indicated by the dotted boxes

mean estimates $\hat{U}_i$; the rising trend up to a distance of around 30 km demonstrates that the five-level area factor does not adequately explain the short-range spatial structure. As there are many pairs of villages within 30 km of each other, and since spatial interpolation is one of our goals, the non-spatial model is clearly inadequate.

We therefore removed the non-spatial component U_i from model (2), to give the revised model

$$\log\{p_{ij}/(1-p_{ij})\} = \alpha + \beta' z_{ij} + S(x_i). \tag{3}$$

Fig. 3 displays the MCMC traces of $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \sigma^2, \phi, \delta)$ that were obtained during the fitting of model (3). Each trace consists of 5000 values sampled after burn-in. The traces differ in the extent to which they show good mixing, but each shows a reasonable degree of convergence to a stationary distribution.

Table 1 shows the posterior mean, median and 95% credible interval for each of the parameters in model (3). With regard to the regression parameters, Table 1 shows that the prevalence of malaria increases with age, and that bed net use reduces the prevalence.

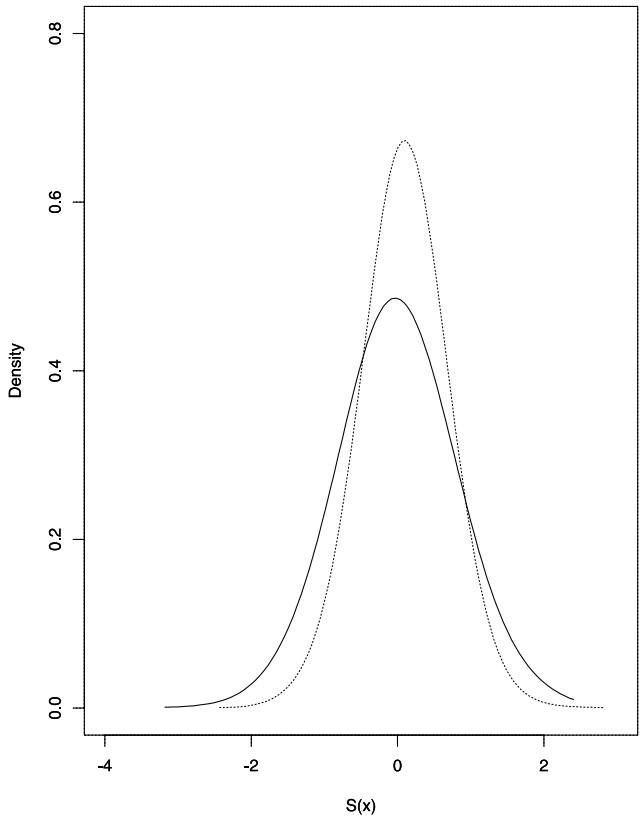

Fig. 6. Posterior density estimates for $S(x)$ at two selected locations: ———, x at the remote location (452, 1493); · · · · · · · , x at (520, 1497), which is close to the observed sites in the central region

Posterior distributions for each of the corresponding regression parameters are approximately Gaussian. Fig. 4 focuses on the regression parameters corresponding to the age and bed net effects. Each part of Fig. 4 shows the empirical distribution of 5000 sampled values from the posterior. Fig. 4(a) confirms that zero is in the extreme lower tail of the posterior for the age-effect parameter β_1. Figs 4(b)–4(d) confirm firstly that zero is in the upper tail of the marginal posterior for each of the two bed net parameters β_2 and β_3 and secondly that because of the negative posterior correlation between β_2 and β_3 the point $\beta_2 = \beta_3 = 0$, which would correspond to no effect of bed net use, is completely infeasible.

Neither inclusion in the PHC system nor the greenness of the surrounding vegetation appears markedly to affect the prevalence of malaria, as in each case the 95% posterior interval comfortably straddles zero. The regression coefficients β_6, β_7, β_8 and β_9 in Table 1 indicate that areas 2 and 3 are very similar to area 1, whereas area 5 and to some extent area 4 have somewhat higher prevalence than area 1. This would suggest that the eastern part is ecologically different from the central and western regions.

To interpret the results for the distance scale parameter ϕ, note that the minimum distance between the sampling locations is 0.95 km and the maximum is 273.3 km. Note in particular that the inclusion of the five-level area factor in the model still leaves substantial spatial variation to be explained: the 95% posterior interval for σ^2 is (0.31, 1.05) on the logit scale. The posterior interval for ϕ of (0.9, 10.2) km indicates that this residual spatial variation operates on a rela-

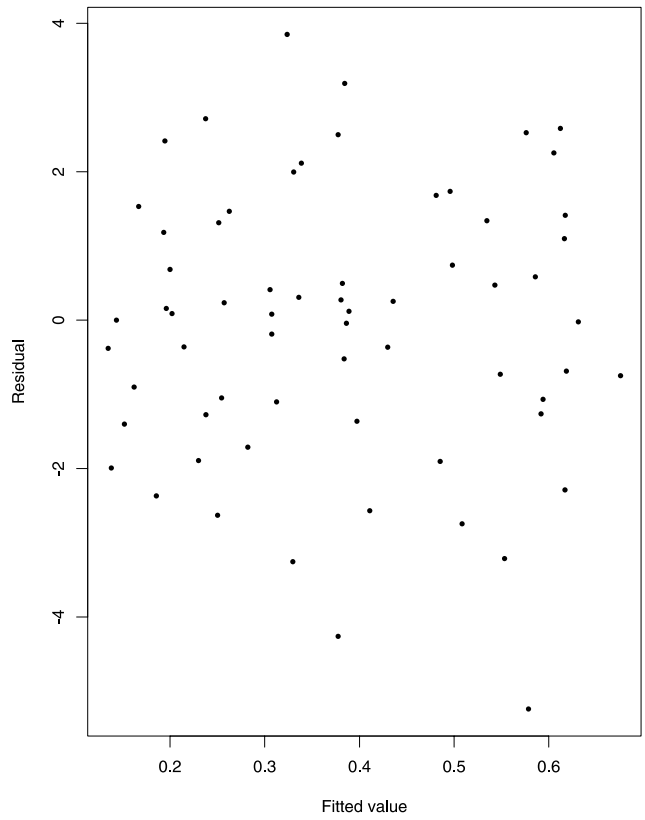

Fig. 7. Village level residuals against fitted values for model (3)

tively small scale. The posterior interval for δ is only slightly narrower than the corresponding central interval from the uniform prior, but the shape of the posterior (not shown) is markedly non-uniform with a single mode around $\delta = 0.5$.

Turning now to spatial prediction, recall from Fig. 1 that the 65 villages formed three well-separated spatial clusters in the western, central and eastern parts of the country. Fig. 5 shows the maps of the predicted (posterior mean) surface $\hat{S}(x)$ within each of these three concentrations of villages. These were computed on three fine grids of 168, 77 and 96 locations covering parts of the western, central and eastern regions of the Gambia respectively. At each prediction location, 5000 values of $S(x)$ were simulated from the chain and the sample mean used to estimate the posterior mean $\hat{S}(x)$.

A general feature of the predictions from any geostatistical model for which the spatial correlation decays to 0 is that for locations x that are sufficiently remote from all sampled locations the data convey essentially no information about the corresponding values of $S(x)$ other than that provided indirectly by estimation of unknown model parameters. It follows that, at such locations, the predictive (posterior) distribution for $S(x)$ is approximately a continuous mixture of zero-mean Gaussian variates, with probability distribution function

$$\int f(s; \sigma^2)\, p(\sigma^2)\, \mathrm{d}\sigma^2,$$

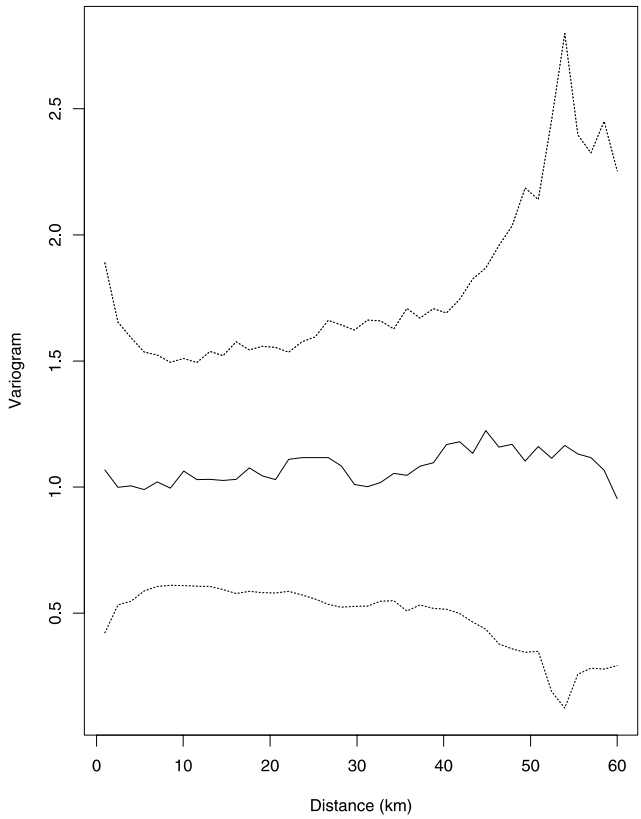

Fig. 8. Village level residual empirical variogram plot: ———, posterior mean; ·····, pointwise 95% posterior intervals constructed from simulated realizations of the fitted spatial model

where $f(\cdot)$ is Gaussian and $p(\cdot)$ is the posterior for σ^2. All such distributions have mean 0, and this explains why Fig. 5 includes uniformly grey shading, corresponding to zero predicted values $\hat{S}(x)$, over areas remote from all the sampled villages. At locations that are sufficiently close to one or more villages, the observed prevalences in those villages directly affect the predictive distribution of $S(x)$; specifically, the predictive distribution shows a reduced variance around a predicted value which is a compromise between the prior mean (on the logit scale) of 0 and the observed prevalence. This effect is illustrated in Fig. 6, which compares the posterior distributions for two locations close to and remote from the central cluster of villages; each posterior density is estimated by kernel smoothing of the corresponding MCMC-generated empirical sample.

3.3. Diagnostic checks

Fig. 7 plots village level residuals r_i against fitted values and shows the desired absence of any obvious relationship, indicating an adequate first-order fit. Fig. 8 plots the empirical variogram of the village level residuals r_i^* together with pointwise 95% posterior intervals derived by simulating independent replicated spatial samples from the fitted model at each MCMC iteration. The empirical variogram is essentially flat, approximately equal to 1 and lies well within the 95% posterior limits throughout the region plotted, suggesting an adequate fit to the second-order structure.

4. Discussion

The present analysis extends that of Thomson *et al.* (1999) by explicitly modelling a spatial component of variation in the prevalence of malaria. Thomson *et al.* (1999) were concerned only with inference about the regression parameters $\boldsymbol{\beta}$. As noted earlier, they used generalized estimating equations in association with a spatially structured working covariance matrix. Very similar approaches were used in Gotway and Stroup (1997) and in Gumpertz *et al.* (2000).

Inferences from generalized estimating equations are not directly comparable with ours for two reasons. Firstly, they use non-Bayesian inference; secondly, their regression parameters have marginal intepretations whereas ours are to be interpreted conditional on the unobserved spatial process $S(x)$. Note also that for the Gambia malaria data Thomson *et al.* (1999) used a quadratic regression model for the effect of greenness. With these provisos, there is a reasonable qualitative agreement between the approximate 95% confidence intervals for the β_i reported by Thomson *et al.* (1999) and the corresponding 95% posterior intervals given in Table 1.

All the above suggests that if interest focuses on the regression parameters $\boldsymbol{\beta}$ there is little to be gained from an elaborate spatial modelling exercise for these data. However, in many applications including this, the practical interest extends to constructing predictive maps for the risk of malaria throughout the country, as an aid to the targeting of scarce public health resources. Our results suggest that this requires smooth spatial interpolation of estimated village effects. Point interpolations could be obtained by simple smoothing of the $\hat{U}_i$ from the non-spatial model, but interpolations derived from a fitted spatial model have the advantage that they lead to a complete predictive distribution, rather than just a point estimate, at any given location. A failure to take account of the posterior uncertainty in $S(x)$ would overestimate the precision of malarial prevalence predictions in unsampled villages. Note also that a general interpretation of our model is that the spatial term $S(x)$ represents the cumulative effect of unidentified covariates which, if they had been available, would have been included as additional terms on the right-hand side of model (3). Residual spatial maps of the kind shown in Fig. 5 might therefore be useful in helping to identify additional covariates, whose origin could be environmental, social or both.

Acknowledgements

This work was supported by the European Union training and mobility of researchers network in 'Computational and statistical methods for the analysis of spatial data' (ERB-FMRX-CT960095), and by the Engineering and Physical Sciences Research Council (GR/L56206).

We also thank Umberto D'Alessandro and colleagues, who collected the morbidity data on which these analyses are based, and Stephen Connor for the provision of the satellite data. Financial support for M. C. Thomson from the Department for International Development, UK, is gratefully acknowledged; however, the Department are not responsible for views expressed here.

References

Besag, J. and Higdon, D. (1999) Bayesian analysis of agricultural field experiments (with discussion). *J. R. Statist. Soc.* B, **61**, 691–746.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.*, **43**, 1–59.

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

Chiles, J.-P. and Delfiner, P. (1999) *Geostatistics*. New York: Wiley.

Connor, S. J., Flasse, S., Perryman, A. and Thomson, M. C. (1998) Environmental information systems; can they help improve malaria risk mapping and forecasting of epidemics? *Disasters*, **22**, 39–56.

D'Alessandro, U., Olaleye, B., McGuire, W., Bennet, S., Langerock, P., Aikins, M. K., Thomson, M. C., Cham, M. K., Cham, B. A. and Greenwood, B. M. (1995) Reduction in mortality and in morbidity from malaria in Gambian children following the introduction of a National Insecticide Impregnated Bednet Programme. *Lancet*, **345**, 479–483.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299–350.

Gotway, C. and Stroup, W. (1997) A generalized linear model approach to spatial data analysis. *J. Agric. Biol. Environ. Statist.*, **2**, 157–178.

Gumpertz, M. L., Wu, C. and Pye, J. M. (2000) Logistic regression for Southern Pine Beetle outbreaks with spatial and temporal autocorrelation. *For. Sci.*, **46**, 95–107.

Handcock, M. S. and Wallis, J. R. (1994) An approach to statistical spatial-temporal modeling of meteorological fields. *J. Am. Statist. Ass.*, **89**, 368–378.

Johnson, N. L. and Kotz, S. (1970) *Distributions in Statistics*, *Continuous Univariate Distributions*, vol. 2. Boston: Houghton-Mifflin.

Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Matérn, B. (1986) Spatial variation. *Lect. Notes Statist.*, **36**.

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

Robinson, T. (2000) Spatial statistics and geographical information systems in epidemiology and public health. In *Remote Sensing and Geographical Information Systems in Epidemiology* (eds S. I. Hay, S. E. Randolph and D. J. Rogers). London: Academic Press.

Stein, M. L. (1999) *Interpolation of Spatial Data*. Berlin: Springer.

Thomson, M. C. and Connor, S. J. (2000) Environmental information systems for the control of arthropod vectors of disease. *Med. Vet. Entomol.*, **14**, 227–244.

Thomson, M. C., Connor, S. J., D'Alessandro, U., Rowlingson, B., Diggle, P., Creswell, M. and Greenwood, B. (1999) Predicting malaria infection in Gambian children from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results. *Am. J. Trop. Med. Hyg.*, **61**, 2–8.

Thomson, M. C., Connor, S. J., Milligan, P. J. W. and Flasse, S. (1996) The ecology of malaria as seen from earth observation satellites. *Ann. Trop. Med. Parasit.*, **90**, 243–264.

Zeger, S. L., Liang, K.-Y. and Albert, P. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.