Geospatial Health 1, 2006, pp. 127-139

Bayesian modelling of geostatistical malaria risk data

L. Gosoniu[1] , P. Vounatsou[1], N. Sogoba[2], T. Smith[1]

[1]Swiss Tropical Institute, Basel, Switzerland; [2]Malaria Research and Training Center, Universite du Mali, Bamako, Mali

Abstract. Bayesian geostatistical models applied to malaria risk data quantify the environment-disease relations, identify significant environmental predictors of malaria transmission and provide model-based predictions of malaria risk together with their precision. These models are often based on the stationarity assumption which implies that spatial correlation is a function of distance between locations and independent of location. We relax this assumption and analyse malaria survey data in Mali using a Bayesian non-stationary model. Model fit and predictions are based on Markov chain Monte Carlo simulation methods. Model validation compares the predictive ability of the non-stationary model with the stationary analogue. Results indicate that the stationarity assumption is important because it influences the significance of environmental factors and the corresponding malaria risk maps.

Keywords: remote sensing, epidemiology, disease control, arthropod-borne viruses.

Introduction

Malaria is the most prevalent human parasitic disease. Although reliable estimates are not available, rough calculations suggest that globally, 250 million new cases occur each year resulting in more than one million deaths (Bruce-Chwatt, 1952; Greenwood, 1990; WHO, 2004). Around 90% of these deaths happen in sub-saharan Africa, mostly in children less than 5 years old. The malaria parasite is transmitted from human to human via the bite of infected female *Anopheles* mosquitoes. Transmission depends on the distribution and abundance of the mosquitoes which are sensitive to environmental factors mainly temperature, rainfall and humidity. By determining the relations between the disease and the environment, the burden of malaria can be estimated at places where data on transmission are not available and high risk areas can be identified. Reliable maps of malaria transmission can guide intervention strategies and thus optimize the use of limited human and financial resources to areas of most need. In addition, early warning systems can be developed to predict epidemics from environmental changes.

Remote sensing is a useful source of satellite-derived environmental data. Geographic Information Systems (GIS) has emerged over the last 15 years as a powerful tool for linking and displaying information from many different sources such as environmental and disease data, in a spatial context. Integrated GIS and remote sensing have been applied to map malaria risk in Africa (Snow et al., 1996; Craig et al., 1999; Thomson et al., 1999; Hay et al., 2000; Kleinschmidt et al., 2001; Omumbo et al., 2002; Rogers et al., 2002). However, the mapping capabilities of existing GIS software are rather limited as they are unable to quantify the relation between environmental factors and malaria risk and to produce model-based predictions. GIS is also used in early warning systems for malaria epidemics (Abeku et al., 2004; Grover-Kopec et al., 2005; Thomson et al., 2006), however the thresholds for environmental factors have been based on expert opinion rather than observed data.

Statistical modelling gives mathematical descrip-

Corresponding author:
Penelope Vounatsou
Department of Public Health and Epidemiology
Swiss Tropical Institute, Socinstrasse 57, 4002-Basel, Switzwerland
Tel. +41 284 81 09; Fax +41 284 81 05
E-mail: penelope.vounatsou@unibas.ch

tions of the environment-disease relations, identifies significant environmental predictors of malaria transmission and provides predictions of malaria risk based on the above relations together with their precision. The standard statistical models assume independence of observations. However, malaria infectious cases cluster due to underlying common environments. When spatially correlated data are analysed this independence assumption leads to overestimation of the statistical significance of the covariates (Cressie, 1993). Spatial models incorporate the spatial correlation according to the way the geographical information is available. For areal data (typically counts or rates aggregated over a particular set of contiguous units) the spatial correlation is defined by a neighborhood structure. For geostatistical data (collected at fixed locations over a continuous study region) the spatial correlation is usually considered as a function of the distance between locations.

Linear regression is applied for modelling geostatistical continuous data which are normally distributed (Gaussian). The spatial correlation is introduced in the residuals (error terms) of the model. The parameters cannot be estimated simultaneously, thus iterative methods are used. The generalised least squares approach (GLS) estimates the regression coefficients conditional on the spatial correlation parameters. The correlation parameters can be estimated conditional on the regression coefficients empirically from the residuals or using maximum likelihood based approaches (Zimmerman and Zimmerman, 1991).

In this paper we present models for geostatistical prevalence data derived from malaria surveys carried out at a number of fixed locations. For this type of data and in general for non-Gaussian geographical data, spatial models introduce at each location an error term (random effect) and incorporate spatial correlation on these parameters. Estimation can use generalised linear mixed models (GLMM). However, this is difficult to apply for spatial problems with large number of locations (Gemperli and Vounatsou, 2004). In addition, estimation of standard errors depends on asymptotic results, which in the case of geostatistical models, do not give unique estimates (Tubilla, 1975).

Bayesian geostatistical models implemented via Monte Carlo methods avoid asymptotic inference and the computational problems encountered in likelihood-based fitting. They were introduced for the analysis of geostatistical data by Diggle et al. (1998) and have been employed in modelling the spatial distribution of parasitic diseases (Diggle et al., 2002; Gemperli et al., 2004; Raso et al., 2004, 2005, 2006; Abdulla et al., 2005; Gemperli et al. 2005, 2006; Clements et al., 2006). Most health applications of Bayesian geostatistical models have relied on an assumption of stationarity, which implies that the spatial correlation is a function of the distance between locations and independent of locations themselves. This assumption is questionable when malariological indices are modelled since local characteristics related to human activities, landuse, environment and vector ecology influence spatial correlation differently at the different locations.

In this paper we present and compare Bayesian stationary and non-stationary models for mapping malaria risk data in Mali. Using model validation we assess the assumption of stationarity and show the impact it can have on inference when non-stationary data are analysed. In Section 2 we describe the malaria data which motivated this work and the environmental predictors we extracted from remote sensing and GIS databases. Section 3 introduces the stationary and non-stationary Bayesian geostatistical models as well as the model validation approaches. The results are presented in Section 4 and the paper ends with final remarks and suggestions for future work given in Section 5.

Materials and Methods

Data

Malaria data

The malaria data were extracted from the "Mapping Malaria Risk in Africa" (MARA/ARMA,1998) data-

base. This is the most comprehensive database on malariological indices initiated to provide a malaria risk atlas by collecting published and unpublished data from over 10000 surveys across Africa. We analysed malaria prevalence data from surveys carried out in children between 1 and 10 years old at 89 sites in Mali (Fig. 1) between 1977 and 1995, including a total of 43, 492 children.

Climatic and environmental data

The environmental data and the databases from which they were extracted are given in Table 1. Preliminary non-spatial analysis indicated that the following factors and their transformation should be included in the analysis: Normalized Difference Vegetation Index (NDVI), NDVI squared, length of malaria season, amount of rainfall, maximum temperature, squared maximum temperature, minimum temperature, squared minimum temperature, distance to the nearest waterbody and squared distance to the nearest waterbody.

The length of malaria season was defined using the seasonality model of Gemperli et al. (2006). They defined a region and month as suitable for malaria transmission when rainfall, temperature and NDVI values were higher than pre-specified cut-offs.

The NDVI values were extracted from satellite information conducted by the NOAAA/NASA Pathfinder AVHRR Land Project (Agbu and James, 1994). NDVI is shown to be highly correlated with other measures of vegetation (Justice et al., 1985) and used as a proxy of vegetation and soil wetness. Index values can vary from -1 to 1 with higher values (0.3 - 0.6) indicating the presence of green vegetation, and negative values indicating water. The temperature and rainfall data were obtained from the "Topographic and Climate Data Base for Africa (1920-1980)" Version 1.1 by Hutchinson et al. (1996). We used the yearly averages over the months suitable for transmission according to the map of Gemperli et al. (2006). The distance to the nearest water source was calculated based on per-

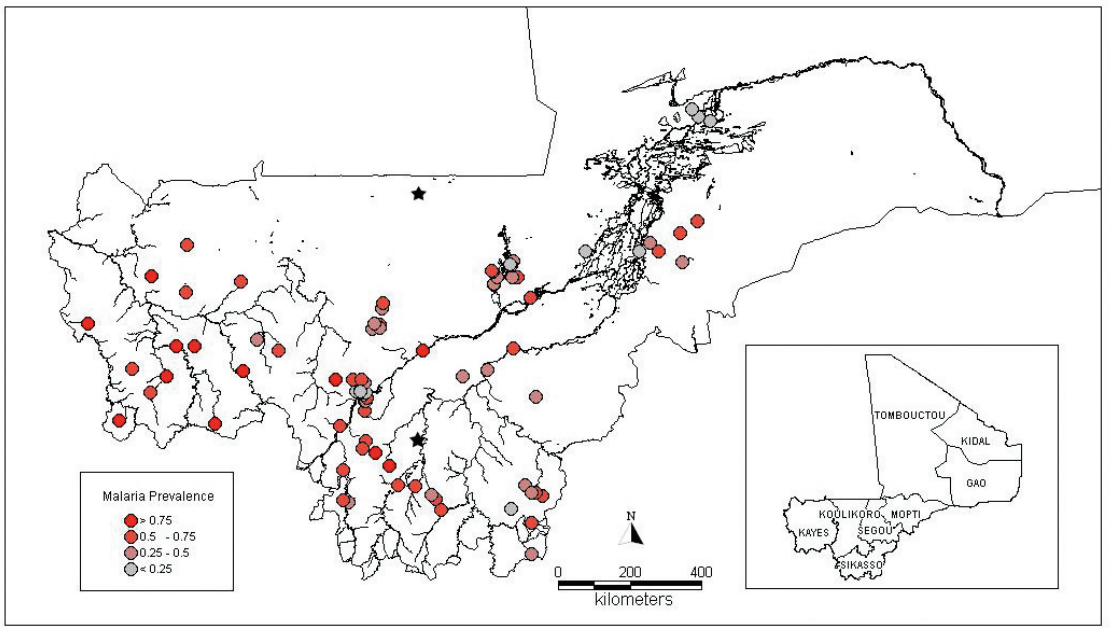

Fig. 1. Sampling locations with dot shading indicating the observed malaria prevalence. The stars indicate the centroids of two fixed tiles used to account for non-stationarity.

Table 1. Spatial databases used in the analysis.

Factor	Resolution	Source
Season length	5km^2	Gemperli et al., 2006
NDVI	8km^2	NASA AVHRR Land data sets
Temperature	5km^2	Hutchinson et al., 1996
Rainfall	5km^2	Hutchinson et al., 1996
Water bodies	1km^2	World Resources Institute, 1995

manent rivers and lakes extracted from "African Data Sampler" (WRI 1995). The covariates were standardized prior to the analysis.

Bayesian geostatistical models

Model formulation

The malaria data are derived from surveys carried out at the various locations. These are typical binomial data and modeled via logistic regression. Let N_i be the number of children tested at location s_i, i =1,...,n, Y_i be the number of those found with malaria parasites in a blood sample and $X_i = (X_{i1}, X_{i2}, ..., X_{ip})^T$ be the vector of p associated environmental predictors observed at location s_i. We assume that Y_i arised from a Binomial distribution, that is $Y_i \sim Bin(N_i, p_i)$ with parameter p_i measuring malaria risk at location si and model the relation between the malaria risk and environmental covariates X_i via the logistic regression $logit(p_i) = X_i^T \beta$, where $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ are the regression coefficients. This model assumes independence between the surveys. However, the geographical location introduces correlation since the malaria risk at nearby locations is influenced by similar environmental factors and therefore it is expected that the closer the locations the similar the way malaria risk varies. To account for spatial variation in the data we introduce an error term (random effect) ϕ_i at each location s_i, that is $logit(p_i) = X_i^T \beta + \phi_i$ and model the spatial correlation on the ϕ_i parameters, that is the ϕ_i's are not independent but they derive from a distribution which models the correlation or equivalently the covariance between every pair of random effects. We adopt the multivariate Normal distribution for the ϕ_i's since they represent error terms and therefore they are defined on a continuous scale, that is $\phi_i = (\phi_1, \phi_2, ..., \phi_n)^T \sim N(0, \Sigma)$. Σ is a matrix with elements Σ_{ij} quantifying the covariance $Cov(\phi_i, \phi_j)$ between every pair (ϕ_i, ϕ_j) at locations s_i and s_j respectively. The distribution of random effect Φ defines the so called Gaussian spatial process.

Stationary model

Assuming stationarity, spatial correlation is considered to be a function of distance only and irrespective of location. Under this assumption, we take $\Sigma_{ij} = \sigma^2 corr(d_{ij}; \rho)$, where *corr* is a parametric correlation function of the distance d_{ij} between locations s_i and s_j. Several correlation functions have been suggested by Ecker and Gelfand (1997). In this application, we choose an exponential correlation function $corr(d_{ij}; \rho) = exp(-d_{ij}\rho)$, where $\rho > 0$ measures the rate of decrease of correlation with distance and it is known as the range parameter of the spatial process. For the correlation function chosen, the minimum distance for which the correlation becomes less than 5% is $3/\rho$. σ^2 measures within location variation and it is known as the sill of spatial process. The above specification of spatial correlation is isotropic, assuming that correlation is the same in all directions.

Non-stationary model

The assumption of stationarity is not always justified, especially over large geographical areas. Differences in agro-ecological zones, health systems and socio-economic indicators may change geographical correlation differently at various locations. In recent years, non-stationary specifications are based on piecewise Gaussian processes (Kim et al., 2002; Gemperli et al., 2003) kernel convolution methods (Higdon et al., 1999; Fuentes et al., 2002) and normalized distance-weighted sums of stationary processes (Banerjee et al., 2004). In Raso et al. (2005) we extended the

Banerjee et al. (2004) model for non-Gaussian prevalence data to map hookworm risk in the region of Man in Cote d'Ivoire. In this paper, we use the same approach to analyse the Mali malaria prevalence data.

The study area is partitioned into K subregions, a stationary spatial process ω_k is assumed in each subregion $k = 1,..., K$ that is $\omega_k = (\omega_{k1},...,\omega_{kn})^T \sim N(0,\Sigma_k)$ and the spatial random effect ϕ_i at each location s_i is modeled as a weighted sum of the subregion-specific stationary processes, that is $\phi_i = \Sigma^K_{k=1} a_{ik}\omega_{ki}$, where a_{ik} are decreasing functions of the distance between location s_i and the centroid of the subregion k. This is equivalent to say that $\phi = (\phi_1,\phi_2,...,\phi_n)^T \sim N(0, \Sigma^K_{k=1} A_k\Sigma_k A_k)$, where $A_k = diag\{a_{1k}, a_{2k},..., a_{nk}\}$ is a matrix which has the elements $a_{1k}, a_{2k},..., a_{nk}$ on the main diagonal and 0 outside the main diagonal. The Σ_k are specified using exponential correlation functions as in the case of the stationary model, that is $(\Sigma_k)_{ij} = \sigma^2_k \, exp(-d_{ij}\rho_k)$. Note that the spatial parameters σ^2_k and ρ_k are specific for each subregion k.

Three non-stationary models were fitted with $K = 2,3,4$. Due to relatively small number of locations included in our data we have not investigated models with larger number of tiles to avoid estimating spatial parameters from tiles with few locations and thus over-parametrising the models. The sub-regions were obtained by overlaying a rectangular grid over the study area. We first divide the rectangle in half north-to-south and then, to obtain four sub-regions, each of these rectangles is partitioned in half west-to-east. For $K = 3$ we divide the north part of our study area in two rectangles and consider the south area as one sub-region. We have chosen the North-South configuration because most environmental and socio-economic differences in Mali are between North and South rather than East and West part of the country. To ensure that there are enough data points in each tile to estimate the model parameters we did not allow tiles with number of points less than a pre-specified minimum of 10. The centroids of two fixed tiles are shown in Fig. 1.

Bayesian specification and implementation

The Bayesian approach to inference allows parameter estimation using information coming from the data via the likelihood function as well as information coming from other sources prior seen the data (i.e. previous studies, subjective judgments) which is formalised via prior distributions. Bayes theorem combines the likelihood function and the prior distribution defining a new quantity, known as posterior distribution which forms the basis of Bayesian inference. Parameters are considered as random and their estimation results not only in a single value, but in the probabilities of their possible values which are given by their probability distribution, known as marginal posterior distribution.

To complete the Bayesian model formulation of the geostatistical models mentioned above we need to specify prior distributions for their parameters. For the regression coefficients we adopt a non-informative uniform prior distribution with bounds $-\infty$ and ∞ which reflects lack of prior knowledge other than that the regression coefficients can take any positive or negative value. For the spatial parameters σ^2, σ^2_k, ρ, and ρ_k we adopt inverse gamma and gamma prior distributions respectively with parameters chosen to have mean equal to 1 and variance equal to 100.

We estimate the parameters of the model using Markov chain Monte Carlo simulation and in particular Gibbs sampling (Gelfand and Smith, 1990). Starting with some initial values about the parameters, the algorithm iteratively updates the parameters by simulating from their full conditional distributions, that is the posterior distribution of each parameter conditional on the remaining parameters. The full conditional distributions of σ^2 and σ^2_k, $k = 1,..., K$ are inverse gamma distributions and simulation from them is straightforward. The rest of the parameters do not have full conditional distributions of known forms. We simulate from the non-standard distributions by employing a random walk Metropolis algorithm (Tierney, 1994) having a Normal proposal density with mean equal

to the estimate of the corresponding parameter from the previous Gibbs iteration and variance equal to a fixed number, iteratively adapted to optimize the acceptance rates. We run five chains with a burn-in of 5000 iterations. Convergence was assessed by inspection of ergodic averages of selected model parameters.

The analysis was implemented in Fortan 95 (Compaq Visual Fortran Professional 6.6.0) using standard numerical libraries (NAG, The Numerical Algorithms Group Ltd.).

Prediction model

Bayesian kriging (Diggle et al., 1998) is used to predict the malaria risk at locations where malaria data are not available. This approach treats the malaria risk at a new location as random and calculates its predictive posterior distribution, which provides not only a single estimate of the risk but a whole range of likely values together with their probabilities to be the true values at a specific location. This makes it possible to estimate the prediction error, a substantial advantage over the classical kriging methods. We estimated the predictive posterior distributions at new locations via simulation. Predictions were made for 28,000 pixels, covering the whole area of south Mali. Further details are given in the Appendix.

Model validation

In total we fitted 4 models (a stationary and three non-stationary). Model fit was carried out on a randomly selected subset of our data (training set) including 69 locations. The remaining dataset of 20 locations was used for validation (testing set). These subsets were selected by assigning a Uniform distribution on the locations.

The goodness-of-fit of each model was assessed using the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). This quantity considers the fit of the data but penalises models that are very complex.

The predictive ability of the models was assessed using a Bayesian "p-value" analogue calculated from the predictive posterior distribution. In particular, for each one of the test locations we calculated the area of the predictive posterior distribution which is more extreme than the observed data. The model predicts the observed data well for a specific location when the observed data is close to the median of the predictive posterior distribution and therefore the "p-value" close to 0.5. A boxplot is used to summarise the "p-values" calculated from the 20 test locations under a particular model. The boxplot displays the minimum, the 25th, 50th, 75th quantile as well as the maximum of the distribution of the 20 "p-values". We consider as best the model with median "p-value" closer to 0.5. The "p-value" is calculated using simulation-based inference by $1/1000\Sigma^{1000}_{j=1} \min(I(p^{rep(j)}_i > p_i^{obs}), I(p^{rep(j)}_i < p_i^{obs}))$, where $I(\cdot)$ denotes the number of points fulfilling the specific condition in the argument, p_i^{obs} is the observed prevalence at test site s_i and $p_i^{rep} = p_i^{rep(1)}, \ldots, p_i^{rep(1000)}$ are 1000 replicated data from the predictive distribution at test location s_i.

A χ^2-based measure was also calculated as an alternative way of comparing the predictive ability of the models. For every test location s_i, we calculated the statistic $\chi^2_i = ((Y_i^{obs} - \hat{Y}_i)^2/\hat{Y}_{ii}$, where Y_i^{obs} are the observed count at test location s_i and $\hat{Y}_{ii}$ is the median of the predictive posterior distribution at s_i. For each model, we obtained the distribution as well as the sum $T\chi^2$ of the χ^2_i values over the 20 test points. The best model was the one with the lowest median of the χ^2_i values and the lowest $T\chi^2$, estimating predicted counts which are closer to the observed ones.

In addition to the above approaches, for each model we calculated 5 credible intervals (the equivalent of confidence intervals in the Bayesian framework) with probability coverage equal to 5%, 25%, 50%, 75% and 95% respectively of the posterior predictive distribution at the test locations. The model which gave better predictions was the one with the highest percentage of locations within the interval of smallest coverage.

Results

The pooled data have shown an overall malaria prevalence of 44.0% (19, 156 children). The median malaria prevalence estimated at village level was 51.3%, ranging from 5.3% to 95.5%.

The univariate non-spatial analysis showed that the following environmental indicators and their transformations were associated with malaria prevalence: NDVI, length of malaria season, rainfall, maximum temperature, minimum temperature and distance to the nearest water body. The relation between malaria risk and rainfall was linear. The logarithmic transformation of NDVI described best its relation with the malaria risk. Polynomial terms of order 2 for minimum temperature, maximum temperature and distance to water gave the best association with malaria prevalence. The results of the bivariate non-spatial logistic regression are summarized in Table 2. All covariates significant at a 15% significance level were included in the spatial analysis.

Fig. 2 compares the predictive ability of the stationary and 3 non-stationary (with 2, 3 and 4 tiles respectively) multiple logistic regression models using the Bayesian "p-value" approach. Each boxplot summarise the distribution of the 20 "p-values" calculated from the predictive posterior distribution of the 20 test locations. The median of this distribution for the non-stationary model with two tiles is the closest to 0.5, suggesting that this is the best model. The same conclusion was drawn by comparing the models using the chi-squared measure. Fig. 3 shows that the non-stationary models with two and three tiles have similar medians of the distribution of χ^2-values over the 20 test locations, but the non-stationary model with two tiles had the lowest $T\chi^2$ value, indicating the smallest deviations between the observations and model predictions.

In Table 3 are presented the percentages of test locations with malaria prevalence which falls in each of the 5 credible intervals of the posterior predictive distribution. We observe that the non-stationary model with two fixed tiles includes 10% of the test locations in the narrowest interval of 5% probability content. This is the highest percentage in comparison to the remaining fitted models. Also in the 95% credible interval the non-stationary model with two fixed tiles has the highest percentage of observed prevalences at test locations, namely 80% in comparison with 75% reported by the other three models.

Table 2 depicts the results of the stationary and the best fitting non-stationary model with two tiles. The stationary model suggested that the following environmental factors are associated with malaria risk: NDVI (in logarithmic scale), maximum temperature, minimum temperature and distance to the nearest water body (in polynomial forms of order 2) and rainfall. In the non-stationary model the rainfall as well as the second order polynomial of the distance to water were not any more related with the

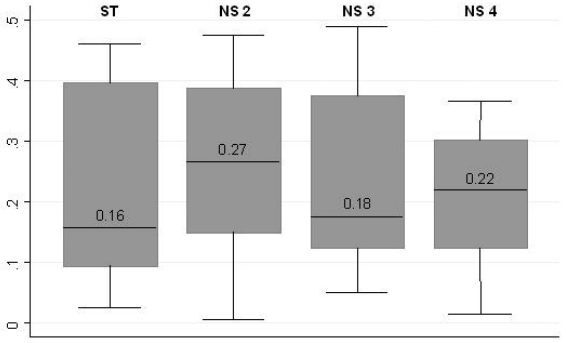

Fig. 2. The distribution of Bayesian p-values for the stationary model (ST), and the non-stationary with 2 (NS 2), 3 (NS 3), and 4 (NS 4) tiles.

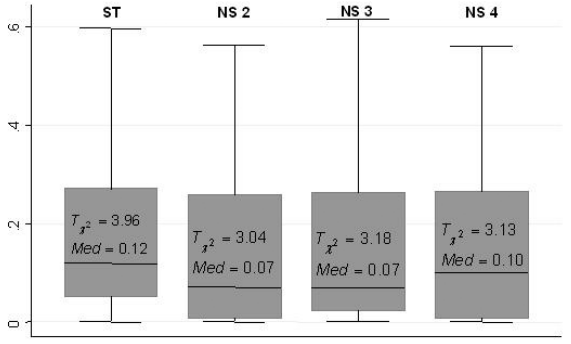

Fig. 3. The distribution and the sum $T\chi^2$ of the χ^2-values over the 20 test points.

Table 2. Posterior estimates for model parameters.

Variable	Bivariate non-spatial model		Stationary spatial model		Non-stationary (2 tiles) spatial model	
	Median	95% CI[a]	Median	95% CI[a]	Median	95% CI[a]
Intercept			0.13	(-0.25, 0.50)	0.21	(-0.20, 0.63)
Log(NDVI)	0.26	(0.24, 0.28)	0.97	(0.43, 1.49)	0.85	(0.28, 1.40)
Log(NDVI)^2	-0.11	(-0.12, -0.10)	0.20	(-0.15, 0.56)	0.13	(-0.25, 0.47)
Seson Length	0.24	(0.22, 0.26)	-0.37	(-0.90, 0.15)	-0.27	(-0.85, 0.30)
Rainfall	0.23	(0.21,0.25)	-0.78	(-1.24, -0.30)	-0.60	(-1.13, 0.01)
Maximum Temperature	-0.40	(-0.42, -0.37)	-1.26	(-1.90, -0.62)	-1.02	(-1.73, -0.18)
Maximum Temperature^2	-0.13	(-0.14, -0.12)	0.07	-0.21, 0.32)	0.05	(-0.21, 0.32)
Minimum Temperature	-0.05	(-0.07, -0.03)	0.94	(0.37, 1.52)	0.90	(0.28, 1.48)
Minimum Temperature^2	-0.22	(-0.23, -0.21)	-0.36	(-0.72, 0.01)	-0.30	(-0.69, 0.09)
Distance to water^2	0.4	(0.38, 0.42)	0.48	(0.11, 0.87)	0.42	(0.03, 0.81)
Distance to water^2	0.10	(0.08, 0.12)	-0.17	(-0.33, -0.002)	-0.15	(-0.31, 0.01)
σ_1^2 (tile 1)			0.81	(0.58, 1.17)	0.88	(0.56, 1.45)
σ_2^{2b}(tile 2)					0.65	(0.31, 1.44)
ρ_1 (tile 1)			2.63	(1.11, 6.09)	0.34	(0.11, 1.68)
ρ_2^b (tile 2)					3.49	(1.37, 7.51)
DIC			507.47		507.50	

[a] Credible intervals (or posterior intervals).

[b] In the case of non-stationary spatial model with 2 fixed tiles we get a set of spatial parameters for each tile.

malaria risk. As we were expected, the higher the value of the NDVI (indicating the presence of green vegetation) the higher the malaria risk. A negative relation with maximum temperature showed that the lower the maximum temperature the higher the malaria risk. Also, malaria risk increases with an increase in the minimum temperature. Surprisingly, the models estimated a positive relation with the distance to water, implying that the risk increases with the distance from permanent water bodies.

The stationary model calculates a posterior median for ρ equal to 2.63 (95 % credible interval: 1.11, 6.09) which, in our exponential setting indicates that the minimum distance for which the spatial correlation is smaller than 5% is equal to $3/\rho$ =1.14 km (95% credible interval: 0.49, 2.71). The best fitting 2-tile non-stationary model confirms that spatial correlation changes as we move from the North to the South part of the country. In particular the minimum distance with negligible correlation is 0.86 km (95% credible interval: 0.40, 2.19) in the North and 8.90 km (95% credible interval: 1.79, 26.88) in the South part. It is interesting to see that although the models differ in their predictive ability (Figs. 2 and 3), the goodness of fit DIC measure does not favor any of the models, showing that it is not able to assess which model has the best predictions.

The smooth maps of malaria prevalence in subsaharan Mali obtained from the stationary and non-stationary spatial model with two tiles are shown in Figs. 4 and 5.

Both maps predicted high malaria prevalence in the region of Kayes (South-West Mali), with the exception of the district of Kayes and in the region of Segou (East-Center Mali). Low prevalence was predicted in the regions of Gao, Tombactou and Kidal (North Mali) and in the district of Kati (Center Mali). Differences between the stationary and non-stationary models appear in the districts of Ansongo, Gourma Rharous, Douentza and western district of Tombactou region (Goundam). Figs. 6 and 7 depict the prediction error from the stationary and non-stationary models respectively. The error is higher in the North Mali where the observed data

Table 3. Percentage of test locations with malaria prevalence falling in the 5%, 25%, 50%,75% and 95% credible intervals of the posterior predictive distribution.

Credible Interval	Bayesian geostatistical model			
	Stationary	NS-2 tiles	NS-3 tiles	NS-4 tiles
5%	5%	10%	0%	5%
25%	15%	25%	15%	25%
50%	30%	55%	50%	35%
75%	55%	60%	65%	55%
95%	75%	80%	75%	75%

were very sparse. The prediction error obtained from the non-stationary model was lower, ranging from 0.36 to 5.7 in comparison to that obtained from the stationary one which varied from 0.70 to 8.11.

Discussion

Accurate maps of malaria risk are important tools in malaria control as they can guide interventions and assess their effectiveness. These maps rely on predictions of risk at locations without observed prevalence data. Malaria is an environmental disease and environmental factors are good predictors of transmission, but the relation between environmental factors, mosquito abundance and malaria prevalence is not linear. This relation can be established only by means of adequate spatial statistical models which can be used for improving predictions of malaria transmission not only in space (for risk mapping) but also in time (for developing early warning systems for malaria epidemics). In this study we present Bayesian geostatistical approaches to assess the malaria-environmental relation for the purpose of malaria risk mapping.

The Bayesian stationary and non-stationary models we presented for analysing the malaria survey data in Mali showed that the statistical modeling approach plays an important role in inference. It influences not only the estimation of parameters related with the spatial structure of the data but also the significance of the malaria risk predictors, the resulting malaria risk maps and the associated predicted errors. Model validation should routinely accompany any model fitting exercise. For the purpose of validation, we recommend to carry out the model fitting on the 80% of the data locations and compare the predictive ability of the models on the remaining locations. For the purpose of mapping, we suggest, once the best model is selected, to apply it to the whole dataset so that the final maps are based on as much data as possible.

Non-stationarity is an important feature of malaria data which is often ignored. Gemperli (2003) developed a non-stationary model for analysing malaria risk data which divides the study region in random tiles, assuming a separate correlation structure within region but independence between tiles. The independence assumption is not justifiable. The number and configuration of tiles are random parameters estimated by the data. The non-stationary modelling approach we adopt here addresses the between-tile independence problem by assuming not only a separate correlation structure within tile but also between-tile correlation. We demonstrate this modelling approach for a fixed space partitioning. This modelling approach is more appropriate when modelling malaria data over large areas covering different ecological zones which define the fixed partition. An extension of the model will allow different covariate effects in each zone. We are currently working on such an approach and implementing it in analysing MARA malaria risk data from West and Central Africa. A further extension of the methodology presented here is to assume random rather than fixed partition of region in tiles. This methodology could be applied in mapping malaria data over large areas with no clear way of finding a fixed partition (i.e. no clearly defined ecological zones).

The main advantage of the Bayesian model formulation is the computational ease in model fit and prediction compared to classical geostatistical methods. Both the stationary and especially the non-stationary models have a large number of parameters. Bayesian computation implemented via MCMC enables simultaneously estimation of all model parameters together with their standard errors. In addition, Bayesian kriging allows model-based predictions (together

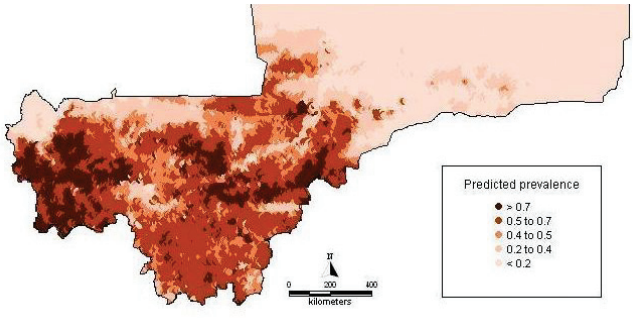

Fig. 4. Map of predicted malaria risk for south Mali using the stationary model.

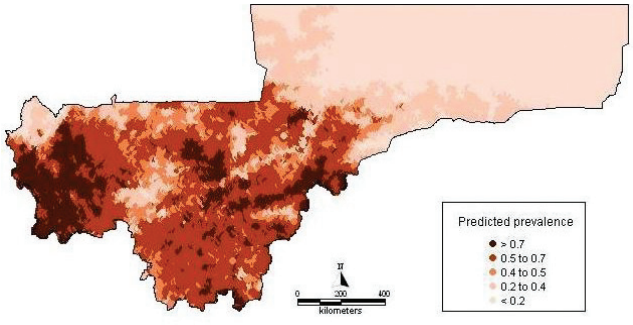

Fig. 5. Map of predicted malaria risk for south Mali using the non-stationary model with 2 fixed tiles.

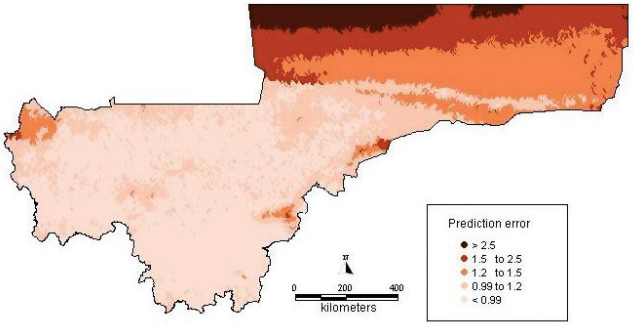

Fig. 6. Map of prediction error for south Mali using the stationary model.

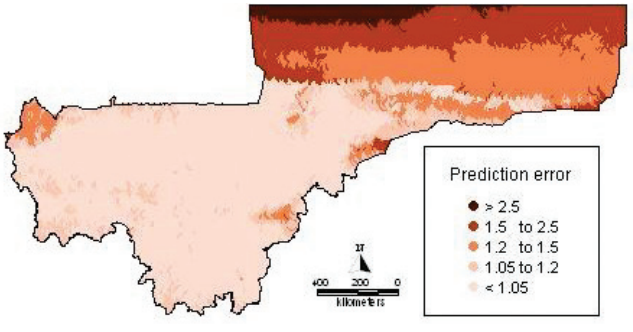

Fig. 7. Map of prediction error for south Mali using the non-stationary model with 2 fixed tiles.

with the prediction error) taking into account the non-stationary feature of the data. This is not possible in a maximum likelihood based framework.

The significant positive association between our data and the distance to water was unexpected. Possible explanation could be because the majority of the main cities (most populated areas) in Mali are located along the river Niger. During the dry season the receding of the river create numerous water pools which serve as vector breeding habitats. The time lag between the rainfall and vector abundance and between vector abundance and the occurrence of the disease may have also played an important role.

Earlier analyses of the MARA data in Mali (Kleinschmidt et al., 2000; Gemperli, 2003) differ in the way the spatial structure is incorporated in the model as well as in the way the covariate effects were modelled. Kleinschmidt et al. (2000) determined the relation between malaria prevalence and environmental predictors by fitting an ordinary logistic regression by maximum likelihood method without taking into account spatial correlation. The prediction map was improved by kriging the residuals and adding them to the map on a logit scale. The main weaknesses of this analysis are firstly that estimation of environmental effects did not take into account the spatial correlation and thus the significance of the covariates may have been underestimated; and secondly the kriging assumes normality, which usually does not hold for the residuals of the logistic regression. Gemperli (2003) re-analysed the data using the Bayesian non-stationary model with random tiles mentioned above. Both previous analyses found a negative relation between malaria risk and distance to water, while Gemperli (2003) suggested also a positive relation with rainfall. Neither analyses assessed non-linear covariate effects. The different analyses reported different covariate effects and produced different maps of prevalence from essentially the same database. Neither performed model validation on test data.

The predicted prevalence map from the non-stationary model with 2 tiles is in a better agreement with the eco-geographical descriptive epidemiology of malaria in Mali (Doumbo et al., 1989) than the maps obtained from the other models. The two maps of predicted malaria prevalence obtained from the stationary and the non-stationary model with two tiles were shown to different malaria epidemilogists in Mali. They all agreed that that the non-stationary model predicts better the epidemilogical situation of malaria in Mali. However, they found that the prevalence in the western part of the country (Kayes region) is over-estimated in comparison with the southern region of Mali (Sikasso). Also previous mapping approaches (Kleinschmidt, 2000; Gemperli, 2003) suggested high malaria prevalence in the western region of Mali. The relatively high predictive standard deviation observed in the North-Western (region of Kayes) and the desert fringes (Tombouctou, Gao and Kidal regions) of the country is probably because of the very few number of data points in these areas rather than the statistical approach. Only one survey has been carried out in the northern regions since 1988.

Further analyses which include recent data, particularly in areas where very few number of data points were observed such as in the north part are needed because environmental changes in the last decades are likely to have influenced malaria transmission dynamics in Mali.

Acknowledgments

The authors would like to acknowledge the MARA / ARMA collaboration for making the malaria prevalence data available. We are thankful to Dr. Sekou F. Traour, Dr. Seydou Doumbia, Dr. Madama Bouar and Dr. Mahamoudou Tour for their useful comments on the predicted risk maps. This work was supported by the Swiss National Foundation grant Nr.3252B0-102136/1.

Appendix

Once the spatial parameters are estimated and the environmental covariates X_0 at unsampled locations are known, we can predict the malaria risk at new sites $s_0 = (s_{01}, s_{02}, \ldots, s_{0l})$

from the predictive distribution
$P(Y_0|Y,N)=\int P(Y_0|\beta,\phi_0)P(\phi_0|\phi,\sigma^2,\rho)P(\beta,\phi,\sigma^2,\rho|Y,N)d\beta d\phi_0 \; d\phi d\sigma^2 d\rho$, where $Y_0 =(Y_{01},Y_{02},...,Y_{0l})$ are the predicted number of cases at locations s_0, $P(\beta,\phi,\sigma^2,\rho|Y,N)$ is the posterior distribution and ϕ_0 is the vector of random effects at new site s_0. The distribution of ϕ_0 at unsampled locations given ϕ at observed locations is normal $P(\phi_0|\phi,\sigma^2,\rho)= N(\Sigma_{01}\Sigma^{-1}{}_{11}\phi,\Sigma_{00} - \Sigma_{01}\Sigma^{-1}{}_{11}\Sigma^T{}_{01})$ with $\Sigma_{11} = E(\phi\phi^T)$ the covariance matrix built by including only the sampled locations $s_1,s_2,...,s_n$, $\Sigma_{00} = E(\phi_0\phi_0^T)$ the covariance matrix formed by taking only the new locations $s_{01},s_{02},...,s_{0l}$ and $\Sigma_{01} = E(\phi_0\phi^T)$ describing covariances between unsampled and sampled locations. For the non-stationary models, $\phi_0 = \Sigma^K{}_{k=1}\; a_{0k}\omega_{k0}$,where a_{0k} are decreasing functions of the distance between new location s_0 and the centroid of the subregion k.

Conditional on ϕ_{0i} and β, Y_{0i} are independent Bernoulli variates $Y_{0i} \sim Ber(p_{0i})$ with malaria prevalence at unsampled site s_{0i} given by $logit(p_{0i})= X^t{}_{0i}\beta + \phi_{0i}$.For the test locations the predicted number of cases Y_{ti} arise from a Binomial distribution $Y_{ti} \sim Bin(N_{ti},p_{ti})$, where N_{ti} is the number of tested children and p_{ti} is the predicted prevalence at test site s_{ti}. The predictive distribution is numerically approximated by the average

$$1/r\Sigma^r{}_{q=1}[\Pi^l{}_{i=1}P(Y_{0i}{}^{(q)}|\beta^{(q)},\phi_{0i}{}^{(q)}]P(\phi_0{}^{(q)},\sigma^{2(q)},\rho^{(q)})$$

where $(\beta^{(q)},\phi^{(q)},\sigma^{2(q)},\rho^{(q)})$ are samples drawn from the posterior $P(\beta,\phi,\sigma^2,\rho|Y,N)$.

References

Abdulla S, Gemperli A, Mukasa O, Armstrong Schellenberg Jr, Lengeler C, Vounatsou P, Smith T, 2005. Spatial effects of the social marketing of insecticide-treated nets on malaria morbidity. Trop Med Int Health 10, 11-18.

Abeku TA, Hay SI, Ochola S, Langi P, Beard B, DeVlas SJ, CoxJ, 2004. Malaria epidemic early warning and detection in African highlands. Trends Parasitol 20, 400-405.

Agbu PA, James ME, 1994. NOAA/NASA Pathfinder AVHRR Land Data Set User's Manual, Goddard Distributed Active Archive Center. NASA Goddard Space Flight Center, Greenbelt.

Banerjee S, Gelfand AE, Knight JR, Sirmans CF, 2004. Spatial modeling of house prices using normalized distance-weighted sum of stationary processes. J Bus Econ Statist 22, 206-213.

Bruce-Chwatt LJ, 1952. Malaria in African infants and children in Southern Nigeria. Ann Trop Med Parasitol 46, 173-200.

Clements ACA, Lwambo NJS, Blair L, Nyandindi U, Kaatano G, Kinung'hi S, Webster JP, Fenwick A, Brooker S, 2006. Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. Trop Med Int Health 11, 490-503.

Craig MH, Snow RW, Le Sueur D, 1999. A climate-based distribution model of malaria transmission in sub-Saharan Africa. Parasitol Today 15, 105-111.

Cressie NAC, 1993. Statistics for spatial data, New York: Wiley.

Diggle P, Moyeed R, Rowlingston B, Thomson M, 2002. Childhood Malaria in the Gambia : a Case-study in Model-based Geostatistics. Appl Stat 51, 493-506.

Diggle PJ, Tawn JA, 1998. Model-based geostatistics. Appl Stat 47, 299-350.

Doumbo O, Outtara NI, Koita O, Maharaux A, Toure YT, Traoure SF, Quilici M, 1989. Approche eco-geographique du paludisme en milieu urbain: ville de Bamako au Mali. Ecol Hum 8, 3-15.

Ecker MD, Gelfand AE, 1997. Bayesian Variogram Modeling for an Isotropic Spatial Process. JABES 4, 347-368.

Fuentes M, Smith RL, 2002. A new class of nonstationary spatial models. Technical Report, Statistics Department, North Carolina State University.

Gelfand AE, Smith AFM, 1990. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85, 398–409.

Gemperli A, 2003. Development of Spatial Statistical Methods for Modelling Point-Referecnced Spatial Data in Malaria Epidemiology. Doctoral Dissertation, Swiss Tropical Institute, University of Basel.

Gemperli A, Vounatsou P, 2004. Fitting generalized linear mixed models for pointreferenced data. JMASM 2, 497-511.

Gemperli A, Vounatsou P, Kleinschmidt I, Bagayoko M, Lengeler C, Smith T, 2004. Spatial patterns of infant mortality in Mali; the effect of malaria endemicity. Am J Epidemiol 159, 64-72.

Gemperli A, Vounatsou P, Sogoba N, Smith T, 2005. Malaria mapping using transmission models: application to survey

data from Mali. Am J Epidemiol 163, 289-297.

Gemperli A, Sogoba N, Fondjo E, Mabaso M, Bagayoko M, Brïet O, Anderegg D, Liebe J, Smith T, Vounatsou P, 2006. Mapping Malaria Transmission in West-and Central Africa. Trop Med Int Health, (to appear).

Greenwood BM, 1990. Populations at risk. Parasitol Today 6, 188.

Grover-Kopec E, Kawano M, Klaver RW, Blumenthal B, Ceccato P, Connor SJ, 2005. An online operational rainfall-monitoring resource for epidemic malaria early warning systems in Africa. Malaria J 21, 4-6.

Hay SI, Omumbo JA, Craig MH, Snow RW, 2000. Earth observation,geographic infortmation system and *Plasmodiu*m *falcipyru*m malaria in sub-Saharan Africa. Adv Parasitol 47, 173-215.

Higdon D, Swall J, Kern J, 1999. Nonstationary spatial modeling. Bayesian Stat 6, 761-768.

Hutchinson MF, Nix HA, McMahon JP, Ord KD, 1996. Africa-A topographic and climate database (CD-ROM). The Australian National University Canberra, ACT 0200, Australia.

Justice CO, Townshend JRG, Holben BN, Tucker CJ, 1985. Analysis of the phenology of global vegetation using meteorological satellite data. Int J Rem Sens 6, 1271–1318.

Kim H-M, Mallik BK, Holmes CC, 2002. Analyzing non-stationary spatial data using piecewise Gaussian process. Technical Report, Texas A & M University, Corpus Christi, TX 78412.

Kleinschmidt I, Bagayoko M, Clarke GPY, Craig M, Le Sueur D, 2000. A spatial statistical approach to malaria mapping. Int J Epidemiol 29, 355-361.

Kleinschmidt I, Omumbo JA, Briet O, van de Giesen N, Sogoba N, Mensah NK, Windmeijer P, Moussa M, Teuscher T, 2001. An empirical malaria distribution map for West Africa. Trop Med Int Health 6, 779-786.

Omumbo JA, Hay SI, Goetz SJ, Snow RW, Rogers DJ, 2002. Satellite imagery in the study and forecast of malaria. Nature 415, 710-715.

Raso G, N'Goran EK, Toty A, Luginbuhl A, Adjoua CA, Tian-Bi NT, Bogoch II, Vounatsou P, Tanner M, Utzinger J, 2004. Efficacy and side effects of praziquantel against *Schistosoma mansoni* in a community of western Cote d'Ivoire. Trans R Soc Trop Med Hyg 98, 18-27.

Raso G, Vounatsou P, Gosoniu L, Tanner M, N'Goran EK, Utzinger J, 2005. Risk factors and spatial patterns of hookworm infection among school children in a rural area of Western Cote d'Ivoire. Int J Parasitol 36, 201-210.

Raso G, Vounatsou P, Singer BH, N'goran EK, Tanner M, Utzinger J, 2006. An integrated approach for risk profiling and spatial prediction of *Schistosoma mansoni* hookworm coinfection. Proc Natl Acad Sci USA 103, 6934-6939.

Rogers DJ, Randolph SE, Snow RW, Hay SI, 2002. Updating historical maps of malaria transmission intensity in East Africa using remote sensing. Photogrammetric Engineering and Remote Sensing 68, 161-166.

Snow RW, Marsh K, Le Sueur D, 1996. The need for maps of transmission intensity to guide malaria control in Africa. Parasitol Today 12, 455-456.

Spiegelhalter DJ, Best N, Carlin BP, van der Linde A, 2002. Bayesian measures of model complexity and fit. J R Stat Soc Ser B, 64, 583 639.

Thomson MC, Connor SJ, D'Alessandro U, Rowlingson B, Diggle P, Cresswell M, Greenwood B, 1999. Predicting malaria infection in Gambia 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 MC, Doblas-Reyes FJ , Mason SJ , Hagedorn R, Connor SJ, Phindela T, Morse AP, Palmer TN, 2006. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439, 576-579.

Tierney L, 1994. Markov chains for exploring posterior distributions (with discussion). Ann Stat 22, 1701-1762.

Tubilla A, 1975. Error convergence rates for estimates of multidimensional integrals of random functions. Technical Report No.72, Department of Statistics, Stanford University, Stanford, CA.

World Health Organization, 2004. Making health research work for people-Progress 2003-2004. http://www.who.int/tdr/publications/publications/pdf/pr17/malaria.pdf

World Resources Institute, 1995. African Data Sampler (CD-ROM) Edition I.

Zimmerman DL, Zimmerman MB, 1991. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. Technometrics 33, 77-91.