Tropical Medicine and International Health doi:10.1111/j.1365-3156.2006.01594.x

VOLUME 11 NO 4 PP 490–503 APRIL 2006

Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania

Archie C. A. Clements[1,2], Nicholas J. S. Lwambo[3], Lynsey Blair[1], Ursuline Nyandindi[4], Godfrey Kaatano[3], Safari Kinung'hi[3], Joanne P. Webster[1], Alan Fenwick[1] and Simon Brooker[2]

1 *Schistosomiasis Control Initiative, Imperial College London, London, UK*
2 *London School of Hygiene and Tropical Medicine, London, UK*
3 *National Institute for Medical Research, Mwanza, Tanzania*
4 *Ministry of Health, Dar es Salaam, Tanzania*

Summary

OBJECTIVE To predict the spatial distributions of *Schistosoma haematobium* and *S. mansoni* infections to assist planning the implementation of mass distribution of praziquantel as part of an on-going national control programme in Tanzania.

METHODS Bayesian geostatistical models were developed using parasitological data from 143 schools.

RESULTS In the *S. haematobium* models, although land surface temperature and rainfall were significant predictors of prevalence, they became non-significant when spatial correlation was taken into account. In the *S. mansoni* models, distance to water bodies and annual minimum temperature were significant predictors, even when adjusting for spatial correlation. Spatial correlation occurred over greater distances for *S. haematobium* than for *S. mansoni*. Uncertainties in predictions were examined to identify areas requiring further data collection before programme implementation.

CONCLUSION Bayesian geostatistical analysis is a powerful and statistically robust tool for identifying high prevalence areas in a heterogeneous and imperfectly known environment.

keywords spatial distribution, maps, Bayesian analysis, *Schistosoma haematobium*, *Schistosoma mansoni*, schistosomiasis, communicable disease control, Tanzania

Introduction

In implementing any control programme, understanding where at-risk populations live is fundamental for appropriate geographical targeting of resources and cost-effective control. Used appropriately, geographical information systems (GIS), remote sensing (RS) and spatial analysis have an as-yet unrealised potential as tools for standardized programme surveillance and implementation (Kamel Boulos 2004; Brooker *et al.* 2006). Numerous studies have been undertaken using satellite-derived environmental data to predict the distribution, abundance and prevalence of diseases and their vectors, including malaria (Hay *et al.* 2000; Rogers *et al.* 2002), leishmaniasis (Elnaiem *et al.* 2003), filariasis (Lindsay & Thomas 2000), trypanosomiasis (Rogers 2000) and schistosomiasis (Brooker *et al.* 2001, 2002a,b; Malone *et al.* 2001; Moodley *et al.* 2003; Kabatereine *et al.* 2004). Although a useful benchmark for future applications of disease mapping in Africa, using accessible and practical methods, these approaches could be improved upon by assessing uncertainty, which is inherent in all aspects of disease mapping, including the data, models, analyses and predictions (Elith *et al.* 2002) and by incorporating the spatial structure of the data. Maps that include estimates of uncertainty in model outputs can allow more informed and objective decision-making in relation to targeted disease control, as the control programme managers gain greater appreciation of decision risk.

Bayesian methods, which offer a flexible and robust approach, are increasingly being applied to spatial analysis, disease mapping and decision-making (Best *et al.* 2005). They provide convenient platforms for incorporating spatial correlation and by modelling both the observed data and any unknowns as random variables, they can incorporate uncertainty into the modelling process. Initially derived for use and now frequently employed in small-area analyses of chronic, non-infectious diseases

(Best *et al.* 2005), Bayesian approaches have recently been used to study the geographical distribution of tropical diseases, both at a large scale, including malaria (Gemperli *et al.* 2004), and at a smaller scale, including filariasis (Alexander *et al.* 2000), malaria (Diggle *et al.* 2002), onchocerciasis (Carabin *et al.* 2003) and *Schistosoma mansoni* infection (Raso *et al.* 2005). While the use of these spatial analytical methods is an attractive research objective, they have rarely been applied in the context of large-scale disease control and there remains a need to document the applicability of GIS, RS and spatial analysis as tools for enhanced planning and implementation of large-scale disease control programmes.

The aim of the present study was to produce accurate, validated prevalence surface maps for both *S. haematobium* and *S. mansoni* infections in northwest Tanzania in order to spatially define an implementation strategy based on mass treatment with praziquantel. An additional aim was to investigate the confidence of the prevalence predictions to inform decisions on whether sufficient data were collected to exclude areas from mass treatment. The applicability of the methods and the implications of the results are discussed in the general context of large-scale disease control programmes.

Methods

Context: control programme and study area

The study was conducted in the context of a national schistosomiasis and soil-transmitted helminths (STH) control programme in Tanzania, which was established in 2003 with support from the Schistosomiasis Control Initiative (SCI) and funded by the Bill and Melinda Gates Foundation. Following WHO guidelines, the programme classifies communities on the basis of prevalence of schistosomiasis in school-age children, according to three strategies: (1) schools where prevalence is <10%, praziquantel is to be made available in local health centres, (2) schools where prevalence is 10%–50%, mass treatment of all school age-children in the community is conducted and (3) schools where prevalence is >50%, mass treatment of all school age children plus other high-risk groups in the community is conducted. Because of the widespread distribution of STH in Tanzania, albendazole is co-administered with praziquantel.

In the first year of implementation (2005), five regions* in coastal Tanzania and six regions in northwest Tanzania will be targeted for mass drug administration. Our study was conducted as part of the intervention in northwest Tanzania. For logistical reasons, data collection was limited to a 670 × 530 km area incorporating Kagera, Mwanza, Shinyanga and Tabora regions (see inset, Figure 1), with the aim of making spatial predictions in all six regions (including Mara and Kigoma).

* Region is the first administrative division of Tanzania. Each region is divided into districts (the second administrative division) and each district into wards (the third administrative division).

Data collection

Parasitological data were collected from at least 60 children (range 60–65) in each of 143 primary schools in northwest Tanzania. Sample sizes (including numbers of schools and children within schools to be surveyed) were determined using simulation studies, based on existing data (Lwambo *et al.* 1999). Lists of schools from the regional basic education statistics reports were used as the sampling frame. Coordinates of the schools were not available and stratification was conducted according to districts. No two schools were selected from the same ward to ensure good geographical coverage of the district. Sampling started with grade 6 and proceeded down through the grades until 30 boys and 30 girls were selected. If surplus students were present, systematic random sampling was employed. Overall, students were aged 6–22 years, with a mean, median and mode age of 14 years and 95% of students were aged between 11 and 17 years. Urine and stool samples were collected from each student and processed. Egg counts were conducted according to standard parasitological procedures on a single slide from each urine sample and two Kato-Katz slides from each stool sample. Ethical approval was obtained from the Ethical Review Board of National Institute of Medical Research (NIMR), Tanzania and the National Health Service Local Research Ethics Committee (NHS-LREC) of St Mary's Hospital, London, UK (EC Number: 03 36).

Variable selection

Environmental data included satellite-derived mean land surface temperature (LST) and normalized difference vegetation index for 1982–1998, obtained from the National Oceanographic and Atmospheric Administration's (NOAA) Advanced Very High Radiometer (AVHRR), elevation, obtained from an interpolated digital elevation model from the Global Land Information System (GLIS) of the United States Geological Survey (http://edcwww.cr.usgs.gov/landdaac/gtopo30/), annual rainfall and the location of inland water-bodies. These data were imported into the GIS ArcMap Version 8.1 (ESRI, Redlands, CA, USA) and linked according to location to the parasitological data.

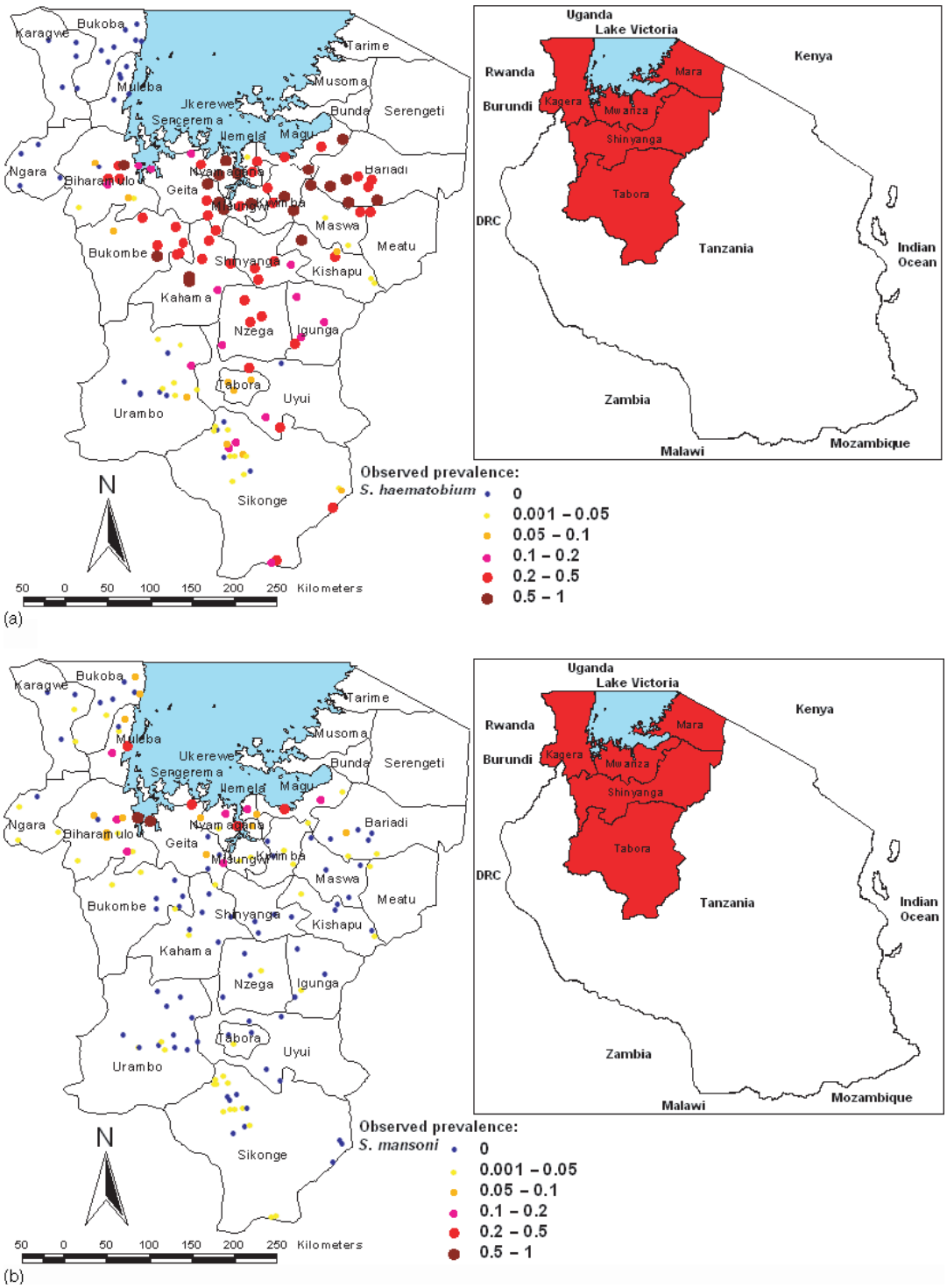

Figure 1 Geographical distribution of schools surveyed in northwest Tanzania and prevalence of (a) *S. haematobium* and (b) *S. mansoni* infections.

Although intensity of schistosome infection has greater relevance to transmission dynamics and morbidity (Anderson & May 1991), prevalence of infection, based on microscopic examination of schistosome eggs in stool or urine samples, remains the most widely used indicator of infection status and is employed by WHO in making recommendations for control. Binomial logistic regression models were constructed in Stata/SE Version 8.1 (Stata Corporation, College Station, TX, USA). Univariable logistic regression was conducted and variables with Wald's $P > 0.2$ were excluded from further analyses. Colinearity was investigated among all possible pairs of potential predictor variables and if any pair had a correlation coefficient $>|0.9|$, the member of the pair that was thought less likely to be biologically important was excluded. With the remaining variables, backwards-stepwise binary logistic regression was conducted using Wald's $P > 0.1$ as the exit criterion and $P \leq 0.05$ as the entry criterion. Non-linear relationships were examined using scatter-plots and entry of categorised predictor variables into the models. In the final model for *S. haematobium*, minimum LST (categorised into <35, 35–39 and >39 °C) and annual rainfall (categorised into <1050 and ≥1050 mm) were included as predictors. In the final model for *S. mansoni*, distance to perennial water bodies (categorised into <0.04, 0.04–0.1, 0.1–0.4 and >0.4 decimal degrees) and annual minimum temperature (as a continuous variable) were included.

Spatial and non-spatial Bayesian modelling

Spatially explicit binomial logistic regression was undertaken in WinBUGS Version 14 (MRC Biostatistics Unit, Cambridge, UK) with the predictor variables identified above entered as fixed-effects and the spatial structure of the residuals modelled as a Gaussian stochastic process using a geostatistical design (Diggle *et al.* 1998). Separate models were developed for *S. haematobium* and *S. mansoni*, including a Bayesian geostatistical model with covariates and a Bayesian geostatistical model without covariates for each of the two parasite species. The reason for including/excluding the covariates was to determine if this had an affect on the fit of the model and to determine whether the covariates accounted for part or all of the spatial correlation in the parasitological data. Additionally, binomial logistic regression models were constructed with the selected predictor variables included as fixed-effects, but without explicitly considering the spatial dependence structure of the data. Again, this was to determine the impact of including the spatial random effect on the fit of the model and also to determine the effect of accounting for spatial correlation on the credible intervals of the coefficients of the selected covariates. Detailed descriptions of the structure of the Bayesian geostatistical models and the process of model assessment are described in Appendix 1.

The deviance information criterion (DIC) statistic was calculated for the Bayesian geostatistical models with and without covariates and the fixed-effects models to determine statistically if addition of the geostatistical component and/or covariates improved model fit (models with a lower DIC statistic are considered to demonstrate a better fit). Maps of the summary statistics of the posterior distributions of predicted prevalence and the spatial random effects were then constructed using ArcMap Version 8.1. More details on how spatial predictions were made at non-sampled locations are provided in Appendix 2.

Model validation

Validation was undertaken using independently collected parasitological data from 54 schools in Magu district (Lwambo *et al.* 1999). Predicted prevalence at validation locations was compared to observed prevalence (the gold standard) using receiver operating characteristics (ROC) analysis, a method widely applied to diagnostic test evaluation and recently applied to validation of regression models (Brooker *et al.* 2002b). The gold standard threshold used for the *S. haematobium* model was ≥0.5, which enabled an assessment of the accuracy of the selected *S. haematobium* model in delineating high priority areas where all at risk groups will receive mass treatment from lower priority areas where only school-age children will receive mass treatment. The gold standard threshold used for the *S. mansoni* model was ≥0.1, which enabled an assessment of the accuracy of the selected *S. mansoni* model in delineating areas receiving mass treatment from areas not receiving mass treatment. Neither the 0.1 threshold for the *S. haematobium* model nor the 0.5 threshold for the *S. mansoni* model could be tested because the validation dataset contained no schools with a *S. haematobium* prevalence less than 0.1 and only two schools with a *S. mansoni* prevalence greater than 0.5. The statistic used for the comparison was the area under the curve (AUC), which relates to the ROC curve, a plot of sensitivity *vs.* one minus specificity, where values greater than 0.9 indicate an extremely well-fitting model, values greater than 0.7 indicate a moderately well-fitting model and values approaching 0.5 indicating a model that is no improvement on random allocation of test status.

Application of spatial predictions to an intervention plan

The SCI treatment programme does not discriminate between *S. haematobium* and *S. mansoni* infections as praziquantel is effective against both species. The prevalence surfaces for *S. haematobium* and *S. mansoni* were, therefore, combined to produce a single intervention map, where contours equating to median predicted prevalence of 0.1 and 0.5, were placed in order to delineate areas with different intervention strategies. The uncertainty in the model predictions was also taken into consideration in planning the intervention. It was decided not to exclude areas where the 95% Bayesian credible intervals suggested that, despite a low median predicted prevalence, high prevalence of either infection may occur.

Results

The observed spatial distributions of infection prevalence based on the 143 schools surveyed are presented in Figure 1. The schools with the highest prevalence of *S. haematobium* infection were located south of Lake Victoria in Mwanza and Shinyanga regions whereas, for *S. mansoni*, they were located close to the shores of Lake Victoria, particularly in the south-western part of the lake.

Model selection and validation

The Bayesian spatial and non-spatial models for *S. haematobium* and *S. mansoni* are presented in Tables 1 and 2. Using the DIC, the *S. haematobium* model containing the geostatistical component but no covariates was the best-fitting, most parsimonious model. Although the covariates were significant when included in the model without the geostatistical component, they became non-significant when it was included. The value of the decay parameter for spatial correlation (ϕ) was 0.2, indicating that a high degree of spatial correlation of *S. haematobium* prevalence was evident between locations with relatively large distances separating them (Figure 2).

The *S. mansoni* model with both environmental covariates and the geostatistical component was the best-fitting, most parsimonious model according to the value of the DIC statistic. The odds ratios indicated that locations adjacent to perennial water-bodies had a much higher prevalence than locations at intermediate distances, which in turn had a much higher prevalence than the reference category (locations at greater than 0.4 decimal degrees from the nearest perennial water body). There was also a positive association between *S. mansoni* prevalence and annual minimum temperature. In contrast to the *S. haematobium* model, the value of the decay parameter

Table 1 Bayesian models for prevalence of *S. haematobium* in northwest Tanzania based on 143 schools, 2004

Model/variable	Coefficient, posterior median (95% Bayes credible interval)	Odds ratio, posterior median (95% Bayes credible interval)
(a) Bayesian logistic regression model, no geostatistical component		
α (intercept)	−1.8 (−1.9 to −1.7)	–
Mean LST 35.0–39.0 °C*	1.7 (1.6 to 1.8)	5.3 (4.8 to 6.0)
Mean LST >39.0 °C*	0.4 (0.1 to 0.6)	1.5 (1.2 to 1.8)
Annual rainfall >1050 mm†	−2.4 (−3.0 to −1.8)	0.09 (0.05 to 0.16)
DIC	2315	–
(b) Bayesian geostatistical model, no covariates		
α (intercept)	2.3 (−0.7 to 5.9)	–
κ (smoothing parameter)	0.9 (0.6 to 1.2)	–
ϕ (decay of spatial correlation)	0.2 (0.1 to 0.5)	–
DIC	662	–
(c) Bayesian geostatistical model with covariates		
α (intercept)	1.9 (−2.3 to 10.3)	–
LST 35.0–39.0 °C*	0.4 (−0.3 to 1.1)	1.5 (0.8 to 2.9)
LST >39.0 °C*	0.3 (−1.5 to 2.2)	1.4 (0.2 to 8.6)
Rainfall >1050 mm†	−1.1 (−3.4 to 1.1)	0.3 (3.3×10^{-2} – 3.1)
κ (smoothing parameter)	0.9 (0.6 to 1.3)	–
ϕ (decay of spatial correlation)	0.2 (0.1 to 1.0)	–
DIC	670	–

LST, land surface temperature; DIC, deviance information criterion.

* Reference category mean LST <35.0 °C.

† Reference category annual rainfall <1050 mm.

Table 2 Bayesian models for prevalence of *Schistosoma mansoni* in northwest Tanzania, based on 143 schools, 2004

Variable	Coefficient, posterior median (95% Bayes. credible interval)	Odds ratio, posterior median (95% Bayes. credible interval)
(a) Bayesian logistic regression model, no geostatistical component		
α (intercept)	−19.6 (−16.7 to −22.4)	–
Distance to water body <0.04 decimal degrees*	3.9 (3.5 to 4.4)	51.2 (34.4 to 78.0)
Distance to water body 0.04–0.1 decimal degrees*	2.3 (2.0 to 2.7)	10.5 (7.1 to 15.5)
Distance to water body 0.1–0.4 decimal degrees*	1.2 (0.9 to 1.6)	3.4 (2.4 to 5.0)
Annual minimum temperature°C	0.9 (0.7 to 1.1)	2.4 (2.0 to 2.9)
DIC	534	
(b) Bayesian geostatistical model, no covariates		
A (intercept)	−4.0 (−5.5 to −0.9)	–
κ (smoothing parameter)	0.8 (0.5 to 1.3)	–
ϕ (decay of spatial correlation)	1.4 (0.4 to 4.1)	–
DIC	399	
(c) Bayesian geostatistical model with covariates		
α (intercept)	−12.3 (−18.8 to −4.5)	–
Distance to water body <0.04 decimal degrees*	4.1 (2.8 to 5.4)	59.3 (15.8 to 230.0)
Distance to water body 0.04–0.1 decimal degrees*	2.3 (1.2 to 3.4)	10.3 (3.3 to 31.4)
Distance to water body 0.1–0.4 decimal degrees*	1.1 (0.1 to 2.0)	2.9 (1.1 to 7.1)
Annual minimum temperature°C	0.4 (0.0 to 0.8)	1.6 (1.0 to 2.3)
κ (smoothing parameter)	0.8 (0.5 to 1.3)	–
ϕ (decay of spatial correlation)	2.8 (1.0 to 5.7)	–
DIC	397	

DIC, deviance information criterion.
* Reference category: distance to perennial water body (>0.4 decimal degrees).

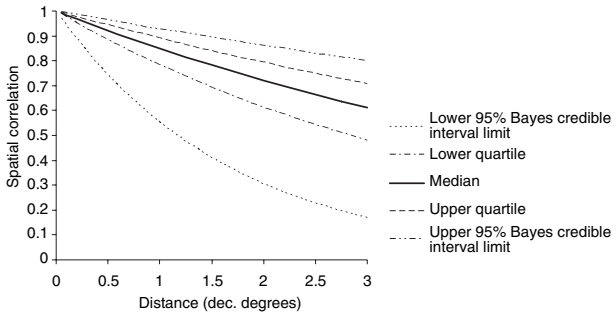

Figure 2 Distance decay of spatial correlation for *S. haematobium* in northwest Tanzania. Note: at the equator, one decimal degree equates to approximately 120 m.

for spatial correlation (ϕ) was 2.8, indicating that spatial correlation of *S. mansoni* prevalence occurred over much shorter distances (Figure 3).

Validation of the selected *S. haematobium* model using an observed prevalence threshold of ≥0.5 gave an AUC of 0.76 indicating a moderately good predictive performance of the selected model in the district from which the validation data were obtained. The AUC for the selected *S. mansoni* model using an observed prevalence threshold of ≥0.1 was 0.95, indicating an extremely good predictive performance.

Spatial predictions: *Schistosoma haematobium*

The outputs of the model are distributions indicating the likely range of values of prevalence at each prediction location and as such, give an indication of the uncertainty surrounding the predictions. Maps of the median and 95% Bayesian credible intervals for the posterior distributions of predicted *S. haematobium* prevalence are presented in Figure 4. An area of high median predicted prevalence (>50%) was apparent directly south of Lake Victoria (but not adjacent to it). Most of the western part of the study area was predicted to have low median prevalence (<10%). On examination of the 95% Bayesian credible intervals, it was clear that *S. haematobium* is not likely to occur at levels above 10% in the northwest of the study area (Kagera region) and Urambo district. However, with the wide credible intervals in the south-western part of the intervention area (Kigoma region), predictions could not be made with a sufficient degree of certainty in that area.

Spatial predictions: *Schistosoma mansoni*

Maps of the median and 95% Bayesian credible intervals for the posterior distributions of predicted *S. mansoni* prevalence are presented in Figure 5. According to the

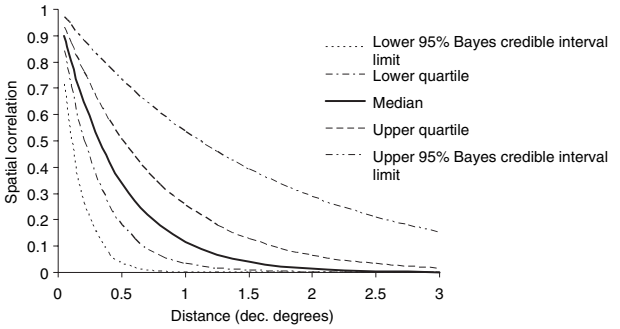

Figure 3 Distance decay of spatial correlation for *S. mansoni* in northwest Tanzania. Note: at the equator, one decimal degree equates to approximately 120 m.

median prediction, all high-prevalence areas (>50%) were located adjacent to the shores of Lake Victoria, particularly along the southern shore and the islands in the southern part of the lake. Intermediate prevalence was predicted along the shores of other perennial water bodies (e.g. Lake Tanganyika) and the eastern and western shores of Lake Victoria. Examination of the 95% Bayesian credible intervals demonstrated a high level of certainty that *S. mansoni* occurs at a high prevalence along the southwest shore of Lake Victoria and the islands in the southwest part of the lake, but that high prevalence may also occur adjacent to other parts of the lake and other water-bodies, despite a low to intermediate median prevalence prediction. It also demonstrated with a high level of certainty that a low prevalence of *S. mansoni* occurs at large distances from perennial water bodies (e.g. in Tabora and Shinyanga regions). Examination of an interpolated surface of the median posterior values of θ_i, the geostatistical component of the model, may give an indication of the factors that have an important role in determining the distribution of *S. mansoni* that were not included in the model (Figure 6). Interestingly, the area of low residual *S. mansoni*

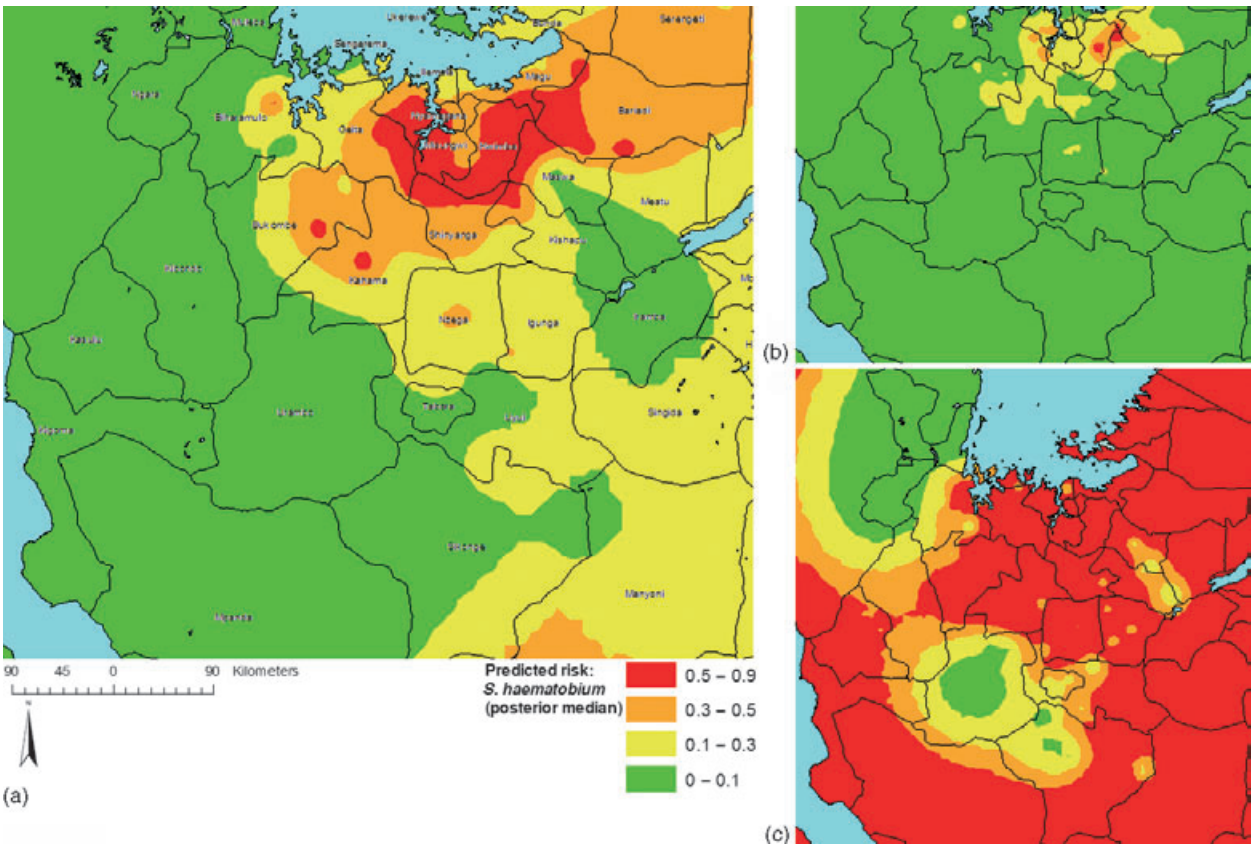

Figure 4 Prevalence predictions for *S. haematobium* in northwest Tanzania: (a) posterior median; (b) lower 95% Bayesian credible interval limit; (c) upper 95% Bayesian credible interval limit.

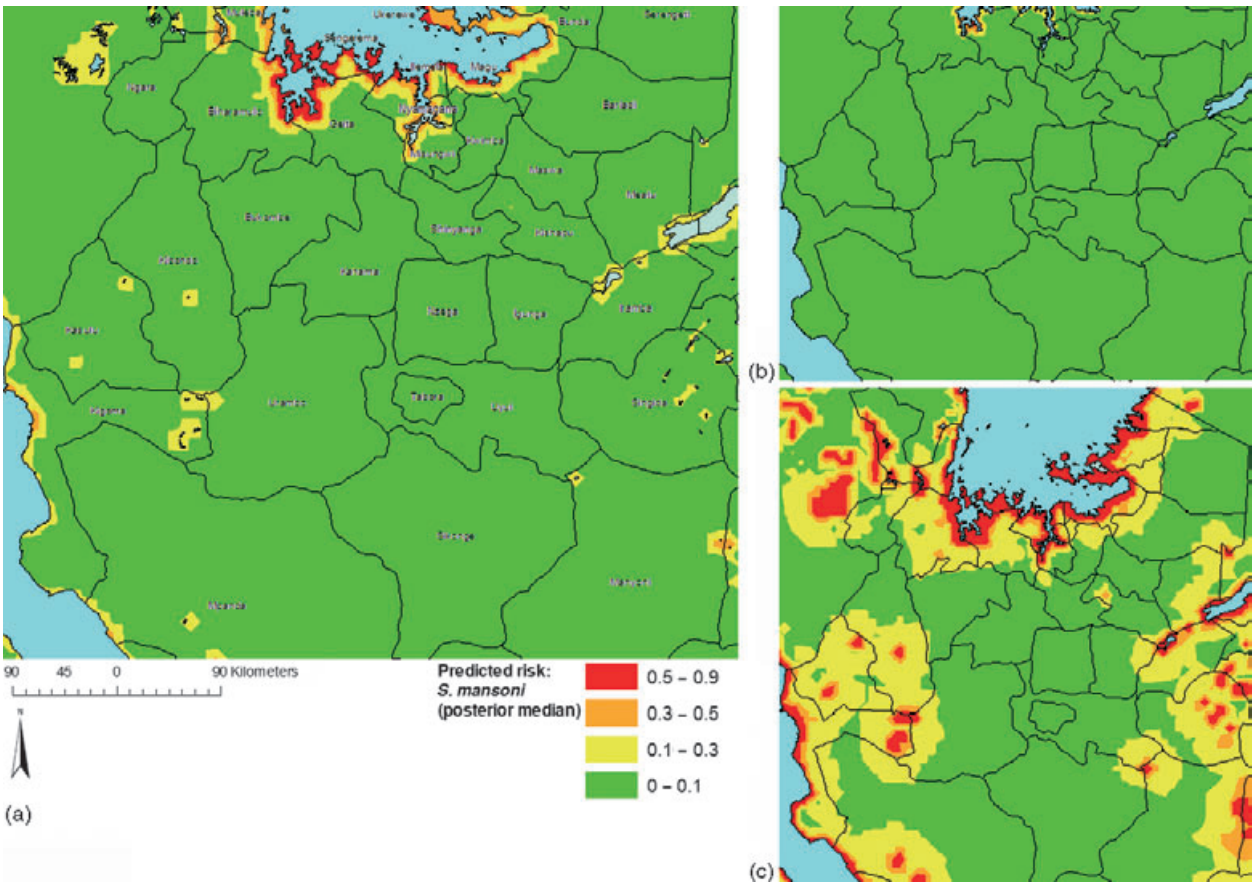

Figure 5 Prevalence predictions for *S. mansoni* in northwest Tanzania: (a) posterior median; (b) lower 95% Bayesian credible interval limit; (c) upper 95% Bayesian credible interval limit.

prevalence approximates the area of high *S. haematobium* prevalence.

Intervention plan

The intervention map, combining both predicted *S. haematobium* and *S. mansoni* risk, is presented in Figure 7. Wide credible intervals in the spatial predictions were mainly apparent in the districts in Mara and Kigoma regions from which no schools were sampled. As all of Mara region was placed in the intervention zone, no further evidence was considered necessary, but as all of Kigoma region fell outside the intervention zone, it was decided not to exclude Kigoma without conducting further data collection and model validation in this region. Additionally, it is believed that no transmission of *S. mansoni* occurs in Lake Tanganyika and model validation in Kigoma region will include an assessment of *S. mansoni* prevalence along the lake.

Discussion

As new initiatives for the control of helminths and other parasites are underway there is a need to geographically stratify areas by risk to guide treatment interventions. Fulfilment of this requirement has been facilitated by developments in GIS/RS and spatial analytical methods. Our study employed statistically robust Bayesian geostatistical models to predict the spatial distributions of *S. haematobium* and *S. mansoni* infections in northwest Tanzania, enabling the targeted implementation of an on-going national schistosomiasis control programme. To the authors' knowledge, this is the first report where spatially explicit analytical methods have been applied to every stage of the planning and implementation of a large-scale infectious disease control programme.

The application of Bayesian modelling in the current study conferred considerable advantages over traditional, frequentist modelling approaches, in that the spatial

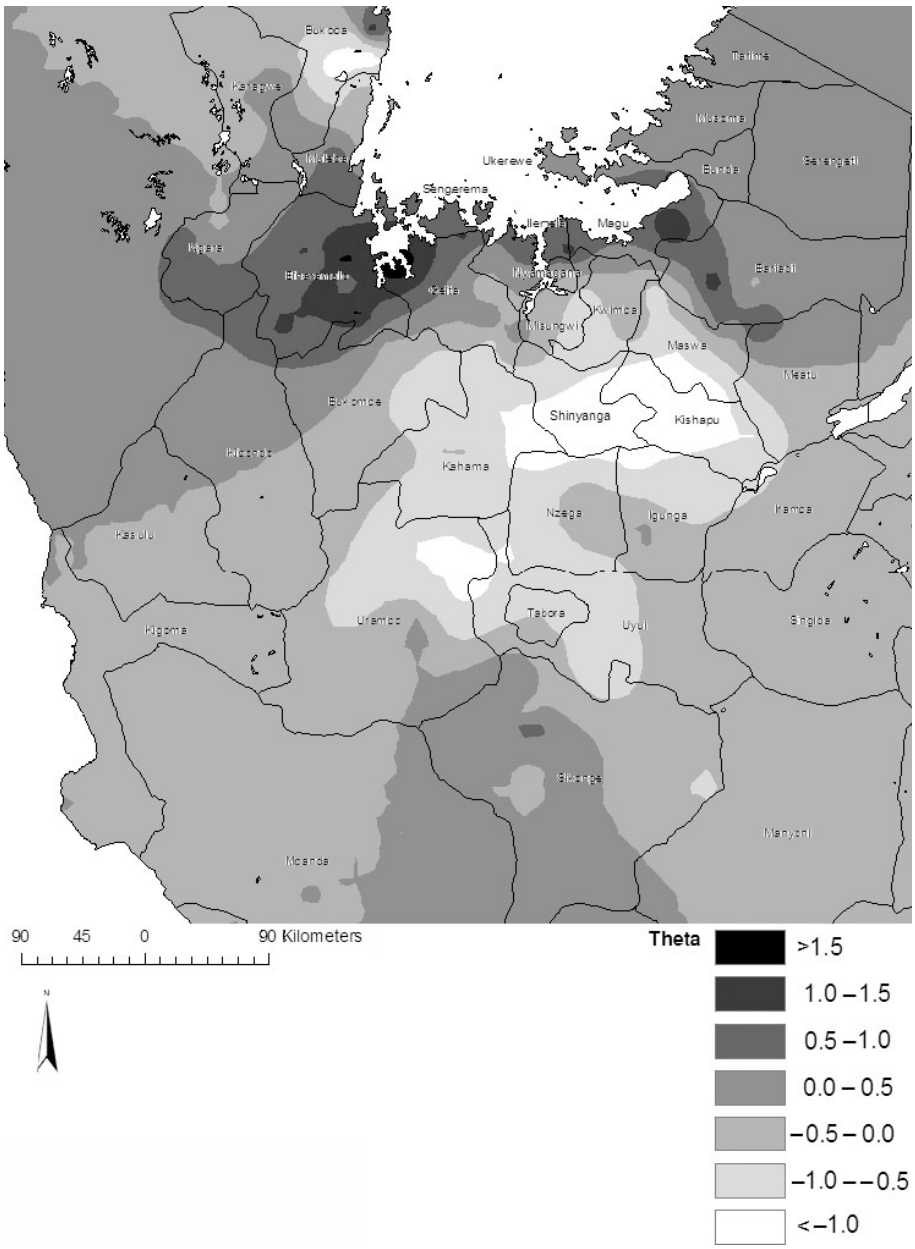

Figure 6 Geostatistical random-effect component (ϕ_i) of the spatially explicit logistic regression model for *S. mansoni* in northwest Tanzania.

dependence structure of the observed data was able to be incorporated in the modelling process in an intuitive and interpretable way and the uncertainties surrounding the predicted prevalence estimates could be fully appreciated, enabling objective decisions to be taken as to the need for further data collection. Failure to account for spatial correlation in the data can lead to inflated significance of regression coefficients and overestimation of the confidence of predictions (Thomson *et al.* 1999), whereas uncertainty in geographical data may be associated with adverse consequences arising from erroneous decisions taken using the data. Agumya and Hunter (2002) discussed a number of ways in which the decision risks associated with geographical data uncertainty can be managed, including

Figure 7 Intervention map for northwest Tanzania with prevalence contours defining areas to be excluded from mass treatment (prevalence of *S. mansoni* or *S. haematobium* <0.1), areas to receive mass treatment of school-age children (prevalence of *S. mansoni* or *S. haematobium* >0.1) and areas to receive priority for mass treatment (prevalence of *S. mansoni* or *S. haematobium* >0.5), including possible targeting of other high-prevalence groups in addition to school-age children.

avoidance, reduction, retention and transfer of risk. We took a risk-avoidance approach by deciding not to rely on model predictions in areas where the Bayesian credible intervals suggested that schistosomiasis could be a problem despite low posterior median estimates of prevalence.

In the non-spatial model for *S. haematobium* and the spatial and non-spatial models for *S. mansoni*, temperature, rainfall and distance to perennial water bodies were important predictors of either or both parasites. These statistical relationships are consistent and interpretable with the known biology of freshwater snails, the intermediate hosts for schistosomes: *S. haematobium* is transmitted by snails from the genus *Bulinus* and *S. mansoni* is transmitted by the genus *Biomphalaria* (Sturrock 1993; Brown 1994). *B. choanomphala* in Lake Victoria and *B. sudanica*, which is found adjacent to the lake, are the main intermediate snail hosts of *S. mansoni* in the study area (Lwambo *et al.* 1999). This distribution of *Biomphalaria* spp. hosts explains, at least in part, the high prevalence of *S. mansoni* in schools at the lakeshore and the low prevalence away from the lakeshore. In contrast, *S. haematobium* has three snail intermediate hosts in the area: *B. nasutus*, *B. africanus* and *B. globosus*, which occupy a mosaic of habitats throughout the area such that transmission of the parasite is widespread (Webbe 1962). Malacological studies have demonstrated that the population dynamics of both *Bulinus* spp. and *Biomphalaria* spp. are sensitive to temperature and rainfall (Appleton 1978; Sturrock 1993; Brooker & Michael 2000).

It is well recognised that schistosomiasis, like many parasitic diseases, has a focal distribution that is associated with spatial correlation in observed parasitological data. The importance of accounting for spatial correlation was highlighted by the fact that the coefficients for the predictors in the *S. haematobium* model became non-significant and although remaining significant, the credible intervals surrounding the coefficients in the *S. mansoni* model were much wider after accounting for spatial correlation. Incorporation of the spatial dependence structure of the data made it apparent that, despite the known biological importance of the environmental covariates, the statistical relationships in the non-spatial models were not supported by the data and spurious significant relationships between the covariates and prevalence would

have been accepted had spatial correlation not been considered. The different range of spatial correlation for the two species of schistosome at least partly reflects differing ecologies of the two parasites, most probably due to the different environmental requirements of their respective intermediate hosts.

The statistical technique used in the current study required that a number of assumptions be made – the most significant of which is that the spatial correlation structure does not vary across the study area (the assumption of stationarity), although methods have been developed recently to deal with non-stationarity in a Bayesian geostatistical framework (Gemperli 2003). Additionally, the method used has not been extended to incorporate potential anisotropy (i.e. where spatial correlation is stronger in one or more directions). The main practical drawback of the approach used in the current study in comparison to previous (non-Bayesian) efforts remains the relative complexity and computational difficulty of Bayesian modelling, despite advances such as Gibbs sampling, improved computer hardware with ever-faster processors and the development of the freely available WinBUGS programme.

An important issue is data quality and the relationship it has with the accuracy of predictions and, ultimately, the population-level impact of the intervention on the outcomes of interest (in our case, reducing morbidity, mortality and the socioeconomic consequences of schistosomiasis). Often, spatial analyses are conducted retrospectively using data from a range of different sources, collected during surveys that have divergent goals [e.g. the MARA/ARMA database (MARA/ARMA 1998)] and the data are therefore not optimal for the purpose of disease mapping. In the current study, spatial analysis and disease mapping were considered as integral components of the planning process and the study and sampling designs were specifically aimed at obtaining statistically efficient, high-quality data for the production of accurate prevalence maps. One potential outcome of demonstrating the practical benefits of high-quality maps to a large-scale control programme such as SCI is that future programmes will pay more attention to the need for collecting optimal data for subsequent spatial analysis.

As well as employing a more robust analytical approach to predict disease prevalence, an important aim of disease mapping in the current study was to define target populations for a national schistosomiasis control programme. However, disease mapping is one of a number of alternative approaches that may be employed to achieve this aim. These alternatives include targeting all individuals without any attempt to define high and low prevalence areas, lot quality assurance sampling (where the minimum number of individuals required to give an assessment of whether prevalence is above a pre-determined 'trigger' threshold, with a given confidence, is sampled in every location) (Brooker *et al.* 2005), targeting all communities within a specified distance of a known infection source (e.g. Lake Victoria for *S. mansoni*) (Brooker 2002) and the use of morbidity questionnaires (for self-reported urinary schistosomiasis) (Lengeler *et al.* 2002). An important area of further research to support (or refute) the credibility of spatial analysis and disease-mapping as tools for planning large-scale interventions is to compare them to the alternative methods described above using economic criteria, such as those used in cost-effectiveness analyses. With each alternative, the relative costs (e.g. monetary costs, human resources and time) and the relative efficacy (in terms of the proportion of locations correctly classified, a function of the sensitivity and specificity of each method) need to be balanced in the context of the wider logistic and operational issues. It will also be important to determine how different targeting approaches alter the allocation of financial resources at local levels and whether this substantially affects the health impact of the programme.

In conclusion, geographic targeting has tremendous potential to enhance the cost-effectiveness of national disease control programmes. Bayesian geostatistical analysis has proved a powerful and statistically robust tool for identifying high prevalence areas in a heterogeneous and imperfectly known environment. The maps we derived using these methods are now being employed at the district level in Tanzania to deliver the intervention where it is needed most.

Acknowledgements

We sincerely extend thanks to the children and school teachers who took part in the parasitological surveys, to staff at the National Institute for Medical Research (Petro Mnyeshi, Silinus Nyanda, John Igogote, Philbert Kashangaki and Sahani Kaboya) for conducting the parasitological examinations, to District and Regional Health and Education Officers in northwest Tanzania for assisting survey teams, to staff of the school health programme located in the central offices of the Tanzanian Ministry of Health (Mohamed Mmole, Ali Ngomelo, Isaak Njau, Edith Ufwinki and Augusta Assey) and Ministry of Education (Mrs Urassa), for their administrative assistance, to Dr Ali Mzige, head of Preventive Services, Tanzanian Ministry of Health, for his support of the SCI programme, to Prof. Charles Kihamia of Muhimbili University College of Health Sciences for providing scientific and administrative support and to Dr John Changalucha, director of the National Institute for Medical

Research, Mwanza for supporting the surveys. We also thank Dr Nicky Best of Imperial College London for giving advice on Bayesian statistical modelling, Sarah Whawell, Professors Christl Donnelly and Roy Anderson of Imperial College London and Professors Mike Kenward and Simon Cousens of the London School of Hygiene and Tropical Medicine for statistical and epidemiological advice on planning the data collection and Dr Simon Hay of Oxford University for providing access to the satellite data. Funding for the activities of the Schistosomiasis Control Initiative (SCI) in Tanzania is generously provided by the Bill and Melinda Gates Foundation. Simon Brooker is supported by a Wellcome Trust Advanced Training Fellowship (073656).

References

Agumya A & Hunter GJ (2002) Responding to the consequences of uncertainty in geographical data. *International Journal of Geographical Information Science* **16**, 405–417.

Alexander N, Moyeed R & Stander J (2000) Spatial modelling of individual-level parasite counts using the negative binomial distribution. *Biostatistics* **1**, 453–463.

Anderson RM & May RM (1991) *Infectious diseases of humans: Dynamics and control*. Oxford University Press, Oxford.

Appleton CC (1978) Review of literature on abiotic factors influencing the distribution and life-cycles of Bilharziasis intermediate host snails. *Malacological Review* **11**, 1–25.

Best N, Richardson S & Thomson A (2005) A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* **14**, 35–59.

Brooker S & Michael E (2000) The potential of geographical information systems and remote sensing in the epidemiology and control of human helminth infections. *Advances in Parasitology* **47**, 246–288.

Brooker S, Hay SI, Issae W *et al.* (2001) Predicting the distribution of urinary schistosomiasis in Tanzania using satellite sensor data. *Tropical Medicine & International Health* **6**, 998–1007.

Brooker S (2002) Schistosomes, snails and satellites. *Acta Tropica* **82**, 209–216.

Brooker S, Beasley M, Ndinaromtan M *et al.* (2002a) Use of remote sensing and a geographical information system in a national helminth control programme in Chad. *Bulletin of the World Health Organization* **80**, 783–789.

Brooker S, Hay SI & Bundy DAP (2002b) Tools from ecology: useful for evaluating infection risk models. *Trends in Parasitology* **18**, 70–74.

Brooker S, Clements ACA & Bundy DAP (2006) Global epidemiology, ecology and control of soil-transmitted helminth infections. *Advances in Parasitology* **62**, 223–265.

Brooker S, Kabatereine NB, Myatt M, Stothard JR & Fenwick A (2005) Rapid assessment of *Schistosoma mansoni*: the validity, applicability and cost-effectiveness of the lot quality assurance sampling method in Uganda. *Tropical Medicine & International Health* **10**, 647–658.

Brown DS (1994) *Freshwater Snails of Africa and their Importance*. Taylor and Francis, London.

Carabin H, Escalona M, Marshall C *et al.* (2003) Prediction of community prevalence of human onchocerciasis in the Amazonian onchocerciasis focus: Bayesian approach. *Bulletin of the World Health Organization* **81**, 482–490.

Diggle PJ, Tawn JA & Moyeed RA (1998) Model-based geostatistics. *Applied Statistics* **47**, 299–350.

Diggle P, Moyeed R, Rowlingson B & Thompson M (2002) Childhood malaria in the Gambia: a case-study in model-based geostatistics. *Applied Statistics* **51**, 493–506.

Elith J, Burgman MA & Regan HM (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distributions. *Ecological Modelling* **157**, 313–329.

Elnaiem DE, Schorscher J, Bendall A *et al.* (2003) Risk mapping of visceral leishmaniasis: the role of local variation in rainfall and altitude on the presence and incidence of kala-azar in eastern Sudan. *American Journal of Tropical Medicine and Hygiene* **68**, 10–17.

Gemperli A (2003) Development of spatial statistical methods for modelling point-referenced spatial data in malaria epidemiology. PhD thesis, University of Basel, pp. 111–134.

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. *American Journal of Epidemiology* **159**, 64–72.

Hay SI, Omumbo JA, Craig MH & Snow RW (2000) Earth observation, geographical information systems and *Plasmodium falciparum* malaria in sub-Saharan Africa. *Advances in Parasitology* **47**, 174–215.

Kabatereine NB, Brooker S, Tukahebwa EM, Kazibwe F & Onapa A (2004) Epidemiology and geography of *Schistosoma mansoni* in Uganda: implications for planning control. *Tropical Medicine & International Health* **9**, 372–380.

Kamel Boulos MN (2004) Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. *International Journal of Health Geographics* **3:1**.

Lengeler C, Utzinger J & Tanner M (2002) Questionnaires for rapid screening of schistosomiasis in sub-Saharan Africa. *Bulletin of the World Health Organization* **80**, 235–242.

Lindsay SW & Thomas CJ (2000) Mapping and estimating the population at risk from lymphatic filariasis in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **94**, 37–45.

Lwambo NJS, Siza JE, Brooker S, Bundy DAP & Guyatt H (1999) Patterns of concurrent infection with hookworm and schistosomiasis in school children in Tanzania. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **93**, 497–502.

Malone JB, Yilma JM, McCarroll JC, Erko B, Mukaratirwa S & Xinyu Zhou (2001) Satellite climatology and the environmental risk of *Schistosoma mansoni* in Ethiopia and East Africa. *Acta Tropica* **79**, 59–72.

MARA/ARMA (1998) *Towards and atlas of malaria risk in Africa, first technical report of the MARA/ARMA collaboration*. Available at http://www.mara.org.za.

Moodley I, Kleinschmidt I, Sharp B, Craig M & Appleton C (2003) Temperature-suitability maps for schistosomiasis in South Africa. *Annals of Tropical Medicine and Parasitology* **97**, 617–627.

Raso G, Matthys B, N'Goran EK, Tanner M, Vounatsou P & Utzinger J (2005) Spatial risk prediction and mapping of *Schistosoma mansoni* infections among schoolchildren living in western Côte d'Ivoire. *Parasitology* **131**, 1–12.

Rogers DJ (2000) Satellites, space, time and the African Trypanosomiases. *Advances in Parasitology* **47**, 129–171.

Rogers DJ, Randolph SE, Snow RW & Hay SI (2002) Satellite imagery in the study and forecast of malaria. *Nature* **415**, 710–715.

Sturrock RF (1993) The intermediate hosts and host- parasite relationships. In: *Human Schistosomiasis* (eds P Jordan, G Webbe & RF Sturrock) CAB International, Wallingford, pp. 33–85.

Thomson MC, Connor SJ, D'Alessandro U *et al.* (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. *American Journal of Tropical Medicine and Hygiene* **61**, 2–8.

Webbe G (1962) The transmission of *Schistosoma haematobium* in an area of Lake Province, Tanganyika. *Bulletin of the World Health Organization* **27**, 59–85.

Corresponding Author **Simon Brooker**, Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. Tel.: +44 (0) 207 927 2614; Fax: +44(0) 207 927 2918; E-mail: Simon.Brooker@lshtm.ac.uk

Analyse spatiale Bayesienne et cartographie de la maladie: outils d'aide à la planification et à l'implémentation du programme de contrôle de la schistosomiase en Tanzanie

OBJECTIF Prédire la distribution spatiale des infections par *Schistosoma haematobium* et *S. mansoni* pour aider à la planification et à l'implémentation de la distribution en masse de praziquantel faisant partie du programme de contrôle national en cours en Tanzanie.
MÉTHODES Les modèles géostatistiques Bayesiens ont été développés en utilisant des données parasitologiques provenant de 143 écoles.
RÉSULTATS Dans les modèles de *S. haematobium*, bien que la température de la surface terrestre et la pluviométrie étaient significativement prédictives de prévalence, elles devenaient moins prédictives lorsque les corrélations spatiales étaient prises en compte. Dans les modèles de *S. mansoni*, la distance aux points d'eau et la température minimale annuelle étaient significativement prédictives même après ajustement avec les corrélations spatiales. La corrélation spatiale intervenait plus fortement pour *S. haematobium* que pour *S. mansoni*. Les incertitudes dans les prédictions ont été examinées pour identifier des régions nécessitant une collection supplémentaire de donnée avant l'implémentation du programme.
CONCLUSION L'analyse géo-spatiale Bayesienne est un robuste et puissant outil statistique pour l'identification de régions à forte prévalence dans un environnement hétérogène et non bien connu.

mots clés distribution spatiale, cartes, analyse bayesienne, *Schistosoma haematobium*, *Schistosoma mansoni*, schistosomiase, contrôle de maladie infectieuse, Tanzanie

Análisis espacial Bayesiano y mapeo de la enfermedad: herramientas para mejorar la planeación e implementación del programa de control de la esquistosomiasis en Tanzania

OBJETIVO Predecir las distribuciones espaciales de las infecciones por *Schistosoma haematobium* y *S. mansoni* para ayudar en la planeación de la implementación de una distribución masiva de praziquantel, como parte del programa nacional de control que se lleva a cabo en Tanzania.
MÉTODOS Se desarrollaron modelos geoestadísticos Bayesianos utilizando los datos parasitológicos de 143 escuelas.
RESULTADOS En los modelos de *S. haematobiu*, aunque la temperatura de la superficie terrestre y la lluvia fueron predictores significativos de prevalencia, se convirtieron en no-significativos cuando se tenía en cuenta la correlación espacial. En los modelos de *S. mansoni*, la distancia a los cuerpos de agua y la temperatura anual mínima fueron predictores significativos, incluso cuando se ajustaba para la correlación espacial. La correlación espacial ocurrió en mayores distancias para *S. haematobium* que para *S. mansoni*. Se examinaron las incertidumbres en las predicciones para identificar áreas que requirieran una mayor recolección de datos antes de implementar el programa.
CONCLUSIÓN El análisis geoestadístico Bayesiano es una herramienta estadística robusta y potente para la identificación de áreas de alta prevalencia en medios heterogéneos y poco conocidos.

palabras clave distribución espacial, mapas, análisis bayesiano, *Schistosoma haematobium*, *Schistosoma mansoni*, esquistosomiasis, control de enfermedades infecciosas, Tanzania

Appendix 1: Model structure and iterative model assessment

The Bayesian geostatistical models were of the form:

$$Y_i \sim Binomial(n_i, p_i),$$

$$logit(p_i) = \alpha + \sum_{\mathrm{N}} \beta_{\mathrm{N}} \times x_{\mathrm{N},i} + \theta_i$$

where Y_i is the observed number positive at location i, n_i is the number tested at location i, p_i is predicted prevalence at location i, α is the intercept, $\sum_{\mathrm{N}} \beta_{\mathrm{N}} \times x_{N,i}$ is a vector of N predictor variables measured at each location i multiplied by their coefficients and ϕ_i, the residual spatial component, is defined by a powered exponential spatial correlation function:

$$f(d_{ij}; \phi, \kappa) = \exp[-(\phi d_{ij})^{\kappa}]$$

where d_{ij} are the distances between pairs of points i and j, ϕ is the rate of decline of spatial correlation with distance and κ is the degree of spatial smoothing (1). Non-informative priors were specified for the intercept (uniform prior with bounds $-\infty$ and ∞) and the coefficients (normal prior with mean = 0 and precision = 1×10^{-4}). The prior distribution of κ was uniform with upper and lower bounds set at 0.5 and 1.5, respectively. The prior distribution of ϕ was also uniform with upper and lower bounds set at 0.1 and 6.0, which gave possible values for spatial correlation of 0.99–0.55 with a separating distance of 0.1 decimal degrees (the minimum distance between observed data points) and possible values for spatial correlation of 0.00–0.55 with a separating distance of 6.0 decimal degrees (the maximum distance between observed data points), assuming $\kappa = 1.0$.

Three chains of the models were run consecutively. A burn-in of 1000 iterations was allowed, followed by 10 000 iterations where values for the intercept and coefficients were stored. Diagnostic tests for convergence of the stored variables were undertaken, including visual examination of history and density plots of the three chains and visual analysis of the Brooks, Gelman and Rubin statistic (2). Convergence was successfully achieved after 10 000 iterations for the *S. mansoni* models and 40 000 iterations for the *S. haematobium* models. The chains were also examined for autocorrelation by visual assessment of the in-built autocorrelation function of WinBUGS. Considerable autocorrelation was apparent for the intercept and some coefficients and it was decided to reduce autocorrelation by thinning subsequent sampling by only storing every third iteration. A total of 10 000 iterations were then run, giving 10 000 stored values from the posterior distribution of each variable (3 chains × 10 000 iterations/3). Descriptive statistics for the posterior distributions of each variable were calculated and analysed. In order to determine whether sufficient iterations had been conducted to fully describe the posterior distributions, Monte Carlo error (MCE)/SD was calculated for each variable and if MCE/SD was less than 0.05, it was decided that sufficient iterations had been conducted. With all stored variables, MCE/SD was less than 0.05 after 10 000 values were generated and stored.

1. Spiegelhalter D, Thomas A, Best N & Lunn D (2003) WinBUGS user manual. Available at http://www.mrc-bsu.cam.ac.uk/bugs.
2. Brooks SP & Gelman A (1998) Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational Graphical Statistics* 7, 434–455.

Appendix 2: Spatial predictions

A set of prediction locations was specified according to a 0.05 × 0.05 decimal degree square grid where the raw data suggested that most of the spatial variation of each infection occurred and a 0.1 × 0.1 decimal degree square grid covering the rest of the study area, where little or no infection was apparent. Grid dimensions were calculated according to computational limits and to give a meaningful prediction density for the intervention. The values of the predictor variables were then determined for each of the prediction locations using ArcMap Version 8.1.

Predicted prevalence was calculated for each prediction location using the Bayesian model that gave the lowest DIC statistic and the 'Spatial.unipred' function of WinBUGS (1), which solves the model equation at each prediction location given the values of each covariate at the prediction location and the distance between prediction locations and observed data locations.

1. Thomas A, Best N, Lunn D, Arnold R & Spiegelhalter D (2004) Geobugs user manual. Available from http://www.mrc-bsu.cam.ac.uk/bugs