Spatial parasite ecology and epidemiology: a review of methods and applications

RACHEL L. PULLAN[1]*, HUGH J. W. STURROCK[1], RICARDO J. SOARES MAGALHÃES[2], ARCHIE C. A. CLEMENTS[2] *and* SIMON J. BROOKER[1,3]

[1] *Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK*
[2] *School of Population Health, University of Queensland, Herston, Queensland, Australia*
[3] *Kenya Medical Research Institute-Wellcome Trust Research Programme, Nairobi, Kenya*

(*Received 13 January 2012; revised 11 March 2012; accepted 3 April 2012; first published online 19 July 2012*)

SUMMARY

The distributions of parasitic diseases are determined by complex factors, including many that are distributed in space. A variety of statistical methods are now readily accessible to researchers providing opportunities for describing and ultimately understanding and predicting spatial distributions. This review provides an overview of the spatial statistical methods available to parasitologists, ecologists and epidemiologists and discusses how such methods have yielded new insights into the ecology and epidemiology of infection and disease. The review is structured according to the three major branches of spatial statistics: continuous spatial variation; discrete spatial variation; and spatial point processes.

Key words: Spatial epidemiology, parasites, spatial statistics, geostatistics, mapping.

INTRODUCTION

Parasites are heterogeneously distributed within host populations (Anderson and May, 1991; Anderson, 1993). Usually, some of this heterogeneity will be spatially structured and explained by various ecological factors and species interactions that are themselves spatially structured. Improved understanding of the spatial patterns of infection and disease, and the processes behind them, can help predict spatial distributions in unsampled areas, assist in the geographical targeting of control interventions and improve our understanding of disease outbreaks.

A number of tools are now available to help us better quantify and understand spatial variation in the patterns of infection and disease. The recent application of global positioning systems (GPS), geographical information systems (GIS) combined with spatial statistical approaches, for example, has provided an improved understanding of spatial patterns and processes (Hay *et al*. 2000; Simoonga *et al*. 2009; Machault *et al*. 2011). This has, in turn, enabled us to predict spatial distributions using remotely sensed environmental data to assist the targeting of control and estimation of the burden of parasitic diseases (Brooker, 2007; Patil *et al*. 2011; Soares Magalhães *et al*. 2011*c*). In this review, we aim to provide an overview of available tools, methods and their applications for improving our understanding of the ecology and epidemiology of human parasitic diseases. As a framework, we consider separately the three major branches of spatial statistics: continuous spatial variation; discrete spatial variation; and spatial point processes (Cressie, 1991; Diggle, 1996).

* Corresponding author: Dr Rachel Pullan. E-mail: rachel.pullan@lshtm.ac.uk

Approaches to spatial analysis

Any statistical approach that accounts for either absolute location and/or relative position (spatial arrangement) of the data can be referred to as spatial. There are three main approaches, illustrated in Fig. 1. The feature that distinguishes between them is the basic underlying statistical model, and the assumptions that this makes regarding the spatial processes involved (Diggle, 2004). For instance, spatial statistics investigating continuous spatial dependency assume that the outcome occurs and is potentially measurable throughout space and, as such, spatial variation in the outcome can be modelled explicitly. In contrast, discrete spatial statistics investigate proximity and are used when data are only available at an aggregate area level. Here, spatial structure is modelled by considering dependency between neighbouring discrete units. Both of these approaches rely upon spatially sampled measurement data, and can be described as global in the sense that they model the overall degree of spatial autocorrelation for a dataset. Spatial point processes, on the other hand, concern the physical location of events distributed within a study region and are used to investigate either the general (i.e. global) propensity for points to cluster or the location of individual (i.e. local) spatial clusters of infection, disease or vector and

Parasitology (2012), **139**, 1870–1887.
doi:10.1017/S0031182012000698

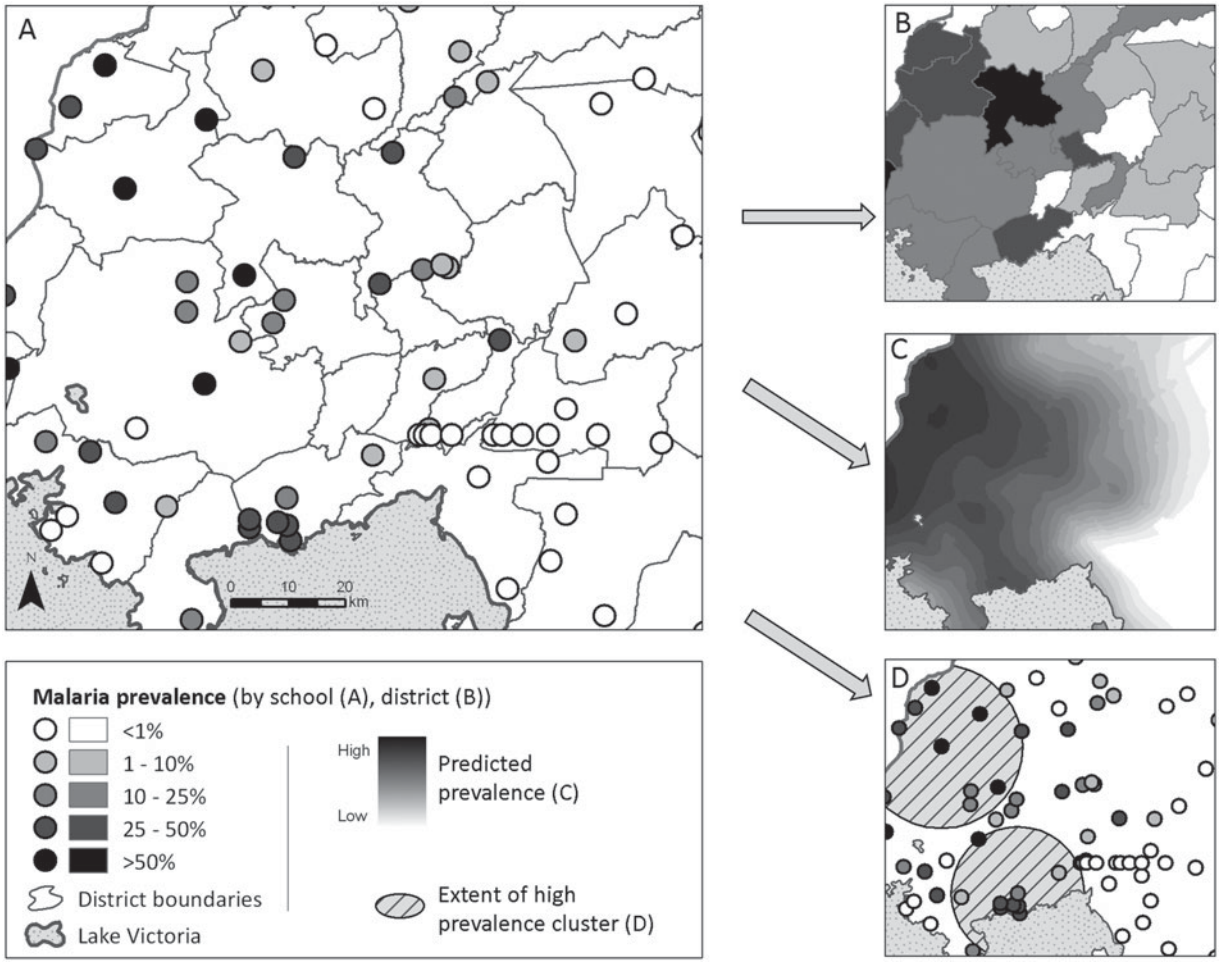

Fig. 1. An illustrated application of the three major branches of spatial statistics, using one dataset. (A) Data used for analysis: Point-level (school-level) malaria prevalence data for Western Kenya, collected during the National School Malaria Survey, 2010 (Gitonga *et al*. 2010). (B) Discrete spatial analysis: data are aggregated to the area level (in this case, mean district prevalence) for presentation and analysis. Discrete spatial statistics can be used to smooth between units, or investigate associations with covariates. (C) Continuous spatial analysis: characterizes spatial dependency (or autocorrelation) between points, and can be used to interpolate predicted outcomes across the entire study region (in this case, using Ordinary Kriging (Goovaerts, 1997)). (D) Spatial point processes: used to investigate the location of individual spatial clusters (indicated as hatched circles) in the outcome (in this case, Kulldorf's spatial scan statistic (Kulldorff and Nagarwalla, 1995)).

intermediate host populations, relative to the underlying population. Below, we describe each of these three approaches.

QUANTIFYING CONTINUOUS SPATIAL DEPENDENCE

Spatial dependence refers to the observation that infection indicators (e.g. prevalence of infection or quantitative egg counts) from samples taken in close proximity to each other are more likely to be related than would be expected by chance, either positively or negatively. This is commonly known as Tobler's first law of geography, whereby "everything is related to everything else, but nearby objects are more related than distant objects" (Tobler, 1970). When investigating continuous spatial dependence, we assume that the outcome can be characterized by a mean, a variance and a correlation structure that is a specified function of location. In such instances, assumptions of independence between observations do not hold true and thus any analysis that ignores spatial dependence risks making inaccurate or misleading inferences (Thomson *et al*. 1999). Quantifying continuous spatial dependence can also provide additional insight into spatial determinants of infection and disease, and thus indicate interesting avenues for investigation. For example, spatial dependence over large spatial scales may suggest the influence of major climatic correlates of infection, whilst spatial dependence existing only between near locations (typical of highly focal infections) might suggest the involvement of local, micro-environmental factors. An understanding of the distance at which spatial dependence occurs can also inform spatial interpolation and prediction and spatial sampling (see below).

When investigating continuous spatial dependence, it is important to distinguish between *first order* (i.e. generally large-scale, deterministic spatial trends) and *second order* (i.e. small-scale, stochastic) effects (Pfeiffer *et al*. 2008). First order trends, for example a north-south gradient in the prevalence of infection, can be readily modelled and accounted for by standard regression techniques. Second order effects arise from spatial dependence and represent the tendency for neighbouring values to be similar in their deviation from the global mean. It is therefore the presence of second order effects that violates assumptions of independence between observations, and thus should be the main focus of any spatial analysis. The categorisation of first order and second order effects of course will change according to the scale of the analysis – for example, variation that appears as a trend at small spatial scales may be seen as second order variation on a larger scale (Legendre and Fortin, 1989; Weins, 1989; Levin, 1992). Similarly, clear deterministic (first order) relationships between infection prevalence and climatic factors evident at country scales may disappear at a community level, overridden by local environmental and socio-demographic characteristics. Most spatial analyses will first begin with identifying any trends in the global mean, and will then focus on investigating underlying spatial dependency in the residuals (Pfeiffer *et al*. 2008). Second order effects are generally assumed to be *stationary* and *isotropic*, meaning that correlation between neighbouring observations is independent from absolute location and does not depend on direction. If dependency between observations is defined by either the physical location of the observations, or by direction, the process is respectively known as *non-stationary* or *geometrically anisotropic*, which can be considerably harder to analyse and model.

A number of statistics have been developed to better describe second order spatial dependency, including Moran's *I* and the inversely related Geary's C (Bailey and Gatrell, 1995). These indicators of global spatial association evaluate whether outcome values are clustered, randomly distributed or evenly dispersed in space, and may form a starting point for more detailed spatial analyses. A more widely used descriptive approach however is the semi-variogram – a cornerstone of classical geostatistics (Goovaerts, 1997). Semi-variograms define semi-variance (a measure of expected dissimilarity between a given pair of observations) as a function of the distance separating those observations, providing information about the range and rate of decay of spatial autocorrelation, as well as the relative contribution of spatial factors to total variation in the outcome. An empirical semi-variogram can be estimated from survey data by calculating the squared difference between all pairs of observations. For ease of interpretation, semi-variance values are grouped and averaged according to separation distance, termed lags. If spatial autocorrelation is present in the data, semi-variance typically increases to a maximum value, termed the sill, before plateauing (Fig. 2). In some instances, semi-variance may continue to rise (known as an 'unbounded' variogram), indicative of first order effects such as directional trends which must be removed from the data, for example by using regression methods.

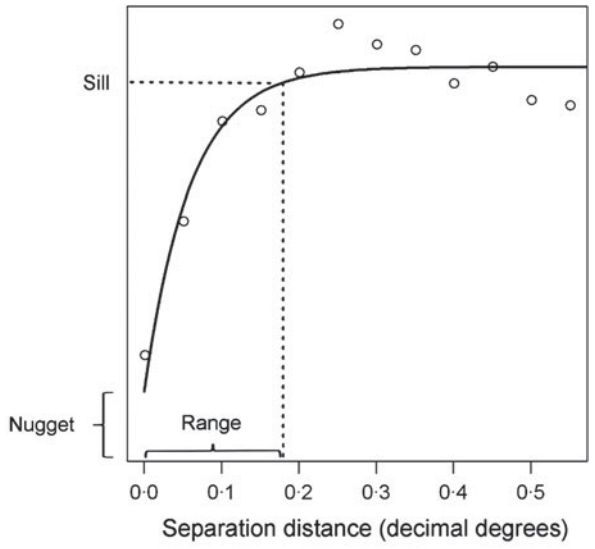

Fig. 2. An example of a semi-variogram, showing its major components. The range represents the separation distance, at which 95% of sill variance is reached, and here is approximately 20 km. The nugget represents the stochastic variation between points, measurement error or spatial autocorrelation over distances smaller than those represented in the data. Data are from a school-based survey of blood in urine indicative of genitourinary schistosomiasis from Coast province, Kenya (Kihara *et al*. 2011) and were de-trended (i.e. first order spatial structure was removed) using a quadratic trend surface.

Once the empirical semi-variogram is estimated, a model semi-variogram can then be fitted as a line through the plotted semi-variance values for each lag. There are a number of permissible model functions that can be used to fit valid semi-variograms, although the most common are the exponential, spherical and Gaussian functions (Cressie, 1991). The modelled value of semi-variance at the intercept (i.e. where points are separated by negligibly small distances) is termed the nugget and represents the stochastic variation between points, measurement error or spatial autocorrelation over distances smaller than those represented in the data. The distance at which the sill is reached is termed the range, and represents the distance over which spatial autocorrelation exists. Points separated by distances larger than the range are therefore equally as dissimilar irrespective of the distance between them. Semi-variograms are a commonly used descriptive tool, and have been used for example to explore spatial heterogeneity of

parasite populations within and between communities, and to quantify the spatial scale at which variation occurs (Srividya *et al*. 2002; Brooker *et al*. 2004*b*; Sturrock *et al*. 2010).

A spatial tool that builds upon semi-variogram analysis is kriging, a weighted moving average technique that interpolates or smooths estimates (depending on whether a zero nugget is assumed), based on values at neighbouring locations and parameters from the semi-variogram. It also provides a relative estimate of prediction error (also known as kriging variance) at each prediction location. In parasite epidemiology, kriging has been widely used for predicting spatial patterns (e.g. the prevalence of infection) at unsampled locations. Taking a recent example, Zouré *et al*. (2011) used this method to produce spatially smoothed contour maps of the interpolated prevalence of eye worm (an indicator for *Loa loa* infection), based on rapid mapping questionnaire data from a sample of 4,798 villages covering 11 potentially endemic country (Zouré *et al*. 2011). The resulting maps were used to identify zones of hyperendemicity, including several previously unknown foci, and provide critical information for large-scale ivermectin treatment programmes. An extension to ordinary kriging is universal kriging, which includes variation due to both covariates and spatial autocorrelation (Goovaerts, 1997), and has for example been used to map malaria risk across Mali (Kleinschmidt *et al*. 2000). This process considers the variable of interest as a first order large-scale trend determined by covariates and a second order spatially auto-correlated residual. An important feature of universal kriging variance therefore is that it incorporates both the error associated with the trend estimation as well as the error of the spatial interpolation.

Major limitations of a classical geostatistical approach include the inability to account fully for inherent uncertainties, such as those arising from the constraints of finite sampling, imperfect survey measurement, uneven data distribution, or of the fitted semi-variogram parameters themselves. It is also less appropriate when considering non-Gaussian outcomes (e.g. proportions and parasite counts). All of these factors can have considerable implications for risk mapping approaches. This has led spatial epidemiologists to turn towards model-based geostatistics (MBG), in which classical geostatistics is embedded in the (usually Bayesian) framework of a generalised linear model (Diggle *et al*. 1998). This offers a more explicit technical and conceptual framework for capturing the relationship between infection outcomes and covariates, providing a more realistic account of uncertainty in both covariance and mean functions (Diggle *et al*. 1998; Cressie *et al*. 2009). Importantly, the model can then be used to generate a distribution of possible values (i.e. a posterior probability distribution) for infection indicators at unsampled locations (interpolation). A rapid expansion of increasingly sophisticated mapping initiatives based upon this method is now being driven by increased computing capacity and the availability of spatially referenced epidemiological data (Soares Magalhães *et al*. 2011*c*; Patil *et al*. 2011). Below we discuss some of the more recent advances using the MBG approach, focusing on two key areas of direct relevance to the effective targeting and evaluation of control programmes: predicting spatial distributions and designing sampling strategies.

Quantifying spatial dependence in order to predict spatial distributions

One application for MBG is predicting the spatial distribution of infection and disease, especially in situations where outcome data are geographically sparse. For example, MBG approaches have been used to model prevalence of infection at regional, national and sub-national levels for a variety of human parasitic diseases, including malaria (Clements *et al*. 2009*c*; Hay *et al*. 2009; Gosoniu *et al*. 2010; Reid *et al*. 2010*a*, *b*), soil-transmitted helminths (STHs) (Raso *et al*. 2006*a*; Pullan *et al*. 2011*a*), schistosomiasis (Clements *et al*. 2006*a*, *b*, 2008*a*, 2009*a*, *b*; Vounatsou *et al*. 2009; Schur *et al*. 2011*a*), lymphatic filariasis (Stensgaard *et al*. 2011) and trypanosomiasis (Wardrop *et al*. 2010). Such maps can provide detailed information on the distribution of infection and disease risk, maximising the usefulness of the data that are available whilst best capturing inherent uncertainties, and can be helpful for the monitoring and evaluation of interventions. Overlaying prevalence of infection maps with human population surfaces can present a novel means for burden estimation, as has been done at regional and global scales for schistosomiasis (Clements *et al*. 2008*a*; Schur *et al*. 2011*a*) and malaria (Hay *et al*. 2010).

Nevertheless, despite being statistically appealing, predictive prevalence surfaces (together with their associated uncertainty) still require some degree of interpretation before being useful for practical disease control guidance. One advantage of the Bayesian approach is the ability to produce maps demonstrating the strength of evidence (i.e. the probability) that intervention prevalence thresholds have been exceeded. For example, studies have identified those areas where there is strong evidence that STH, schistosomiasis or *Loa loa* infection prevalence exceed policy implementation thresholds for mass drug administration, and where high uncertainty warrants further surveys (Diggle *et al*. 2007; Clements *et al*. 2008*a*; Pullan *et al*. 2011*a*). An example of such an approach applied to STH infections, taken from Pullan *et al*. (2011*a*), is shown in Fig. 3.

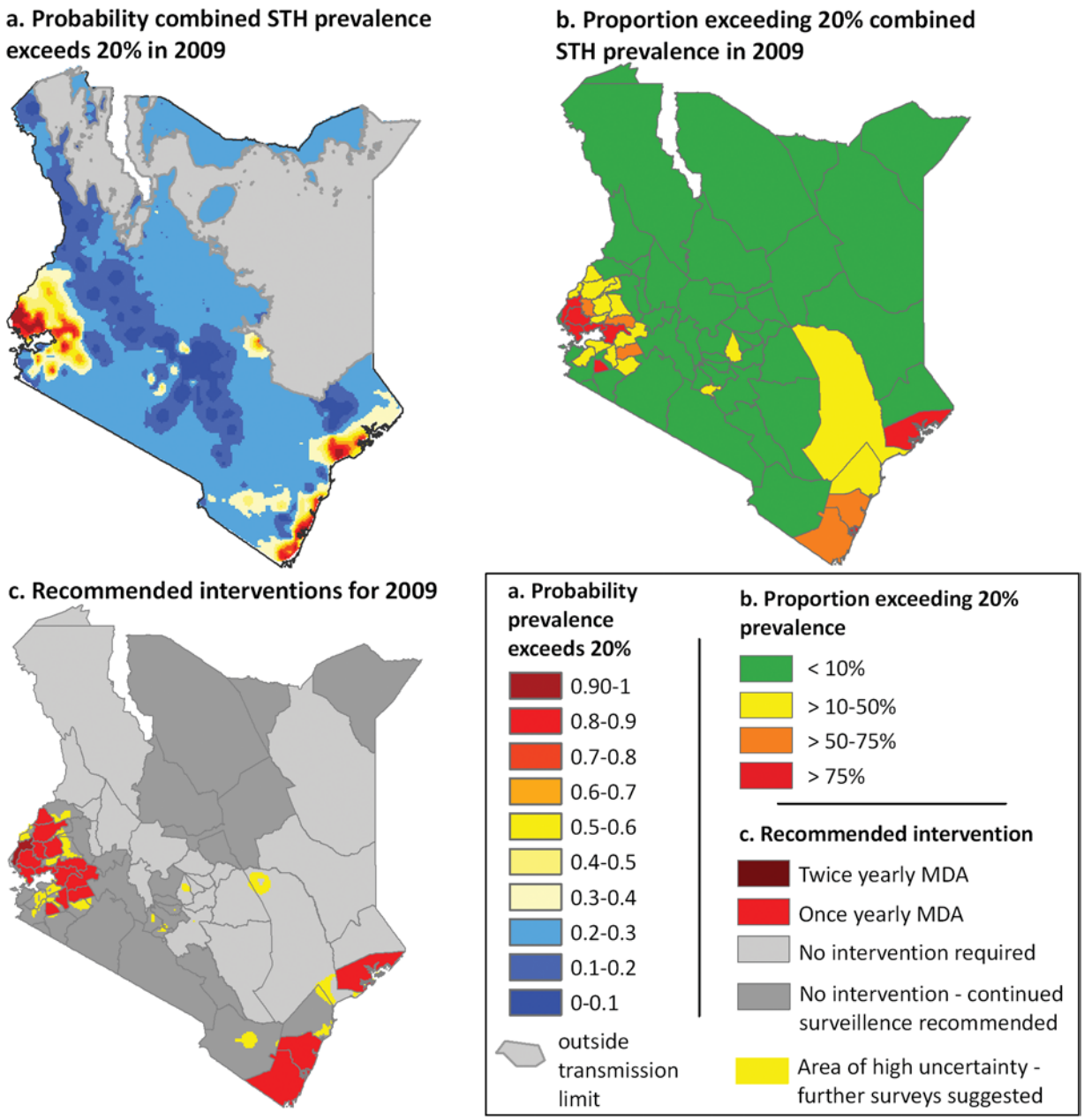

Fig. 3. An example of the practical applications of a model-based geostatistical (MBG) predictive mapping of soil-transmitted helminths (STH). (a) Bayesian space-time geostatistical models were developed for each STH species using survey data from 1980–2009, and were used to interpolate the probability that combined infection prevalence exceeded the 20% level defined by the World Health Organisation as a mass drug administration (MDA) threshold in 2009. (b) Population census data were overlaid with the probability models to estimate the proportion of the population at risk (i.e. >50% probability of exceeding 20% prevalence threshold) and requiring treatment in 2009 for each district. Recommended intervention districts (c) are defined as: once yearly mass drug administration (MDA), at least 33% of the district exceeds 20% prevalence threshold, and twice yearly MDA, at least 33% of the district exceeds a 50% prevalence threshold. Continued surveillance is recommended for districts where historically >75% of the district exceeded the 20% prevalence based on predictions for 1999, and areas of high uncertainty are those where we can only be 50–65% certain that prevalence is lower than 20%. Adapted from Pullan *et al*. 2011.

Although the most common applications of MBG typically use a binomial/logistic regression model (i.e. modelling the prevalence/presence of infection), generalised linear models can handle a variety of data types. For example, density/intensity of infection can provide a more informative indicator of disease burden than simple presence of infection for many parasites. Such data are usually over-dispersed and so better captured using a negative binomial or zero-inflated Poisson distribution. Alexander *et al*. (2000) used a negative binomial MBG to model effectively individual *Wuchereria bancrofti* parasite count data from communities in Papau New Guinea (Alexander *et al*. 2000), whilst Brooker *et al*. (2006) adapted this model to investigate the small-scale spatial heterogeneity in STH and schistosome infections in rural

and urban environments in Brazil. Spatial negative binomial and zero-inflated Poisson models of faecal egg count data have also been developed for *S. mansoni, S. haematobium* and hookworm at community and country levels (Vounatsou *et al*. 2009; Pullan *et al*. 2010; Soares Magalhães *et al*. 2011*b*) and for *S. mansoni* at regional levels (Clements *et al*. 2006*b*). In addition, multinomial models have been built to stratify areas on the basis of prevalence of high- and low-intensity *S. haematobium* infections in West Africa (Clements *et al*. 2009*b*), and to model the distribution of malaria-helminth co-infections at country (Raso *et al*. 2006*b*) and regional levels (Brooker and Clements, 2009). Finally, a Bayesian framework allows the inclusion of multiple imputation steps. For example, the Malaria Atlas Project has incorporated a Bayesian model to predict malaria incidence as a function of parasite prevalence directly (Patil *et al*. 2009) within the geostatistical framework used to model infection prevalence (Hay *et al*. 2010).

Data used for mapping parasitic diseases typically originate from a range of sources using various diagnostics, age groups and sampling methods. A Bayesian inference approach can be adapted to account for these additional sources of uncertainty. For example a number of approaches have been used to adjust for combining data from different age groups, ranging from the inclusion of fixed regression coefficients and random alignment factors (Pullan *et al*. 2011*a*; Schur *et al*. 2011*c*) to the incorporation of mathematical age-standardisation algorithms (Hay *et al*. 2009; Gething *et al*. 2011). Diagnostic tests for a large range of parasites typically have poor sensitivity and specificity, at least in part due to significant day-to-day and intra-specimen variation (Utzinger *et al*. 2001; Booth *et al*. 2003; Engels and Savioli, 2006; Farnert, 2008; Leonardo *et al*. 2008; Tarafder *et al*. 2010). In response, in addition to simply adjusting for the type of diagnostic method used (Pullan *et al*. 2011*a*), authors have explored bivariate outcome spatial models that allow for calibration of spatially correlated data series (Crainiceanu *et al*. 2008), and models that include outcomes as random variables with 'informative' priors defined by observed diagnostic uncertainties (Wang *et al*. 2008).

Another recent extension includes adding a temporal dimension. For example, temporal effects have been handled as random coefficients when modelling STH prevalence across Kenya (Pullan *et al*. 2011*a*) and malaria across Vietnam (Manh *et al*. 2010), explicitly allowing dependency between observations within years. This approach has been extended for mapping malaria at global scales using a sophisticated two-dimensional space-time random coefficient (Hay *et al*. 2010), thus simultaneously modelling correlation between data points in both space and time. Such models can provide more accurate predictions when data are distributed through time as well as space, providing a better understanding of both the contemporary distribution of infection as well as changing risks since the launch of large-scale control. Spatially explicit approaches can also help better capture environmental contexts when investigating co-occurrence of parasite species. For example, studies have used MBG approaches to investigate the geographical distribution of multiple species infection with helminths and malaria at differing spatial scales (Raso *et al*. 2006*b*; Brooker and Clements, 2009; Pullan *et al*. 2011*b*; Soares Magalhaes *et al*. 2011*b*; Brooker *et al*. 2012), facilitating more detailed investigation of associations between species. Lastly, MBG models have been adapted to better capture complex, non-linear relationships with covariates thus providing a deeper understanding of the determinants of infection. For example, authors have used methods such as penalised spline regression (Crainiceanu *et al*. 2005; Gosoniu *et al*. 2009; Soares Magalhaes *et al*. 2011*a*).

Despite their utility, considerable caution must be used when building and interpreting complex MBG models. For example, careful consideration of appropriate model specifications and priors are essential to prevent invalid or inefficient inferences. Systematic changes in diagnostic or sampling methods over time or space can also be misinterpreted as genuine change in disease status. Most MBG models reported in the literature also assume that spatial autocorrelation does not vary with location (so-called stationary models). Whilst such an assumption may be valid across small scales, this may not be true when considering spatial processes over large geographical areas where variation in geography, control, vectors and even parasite strains can cause spatial variation in autocorrelation. To tackle this problem, a number of approaches have been developed (Gemperli, 2003; Kim *et al*. 2005; Raso *et al*. 2006*a*, Beck-Worner *et al*. 2007; Gosoniu *et al*. 2009) although application at large spatial scales is still hindered by practical and computational constraints. To our knowledge, few groups have yet to tackle the issue of geographical anisotropy in parasite epidemiological analyses, although direction is likely to play an important role in observations of spatial dependency for focally transmitted infections such as schistosomiasis and trypanasomiasis (Vounatsou *et al*. 2009). Stein (2005) however has proposed a geographically anisotropic version of the space-time covariance matrix used spatio-temporal MBG, that has since been adapted by Gething *et al*. (2011) to model the global distribution of malaria (Stein, 2005; Gething *et al*. 2011).

Due to the computational burden required to generate predictions at each individual prediction location, most applications of MBG models tend to be via a 'per prediction point' approach, yielding marginal prediction intervals that realistically capture appropriate measures of 'local' uncertainty. However, failing to account for spatial or temporal

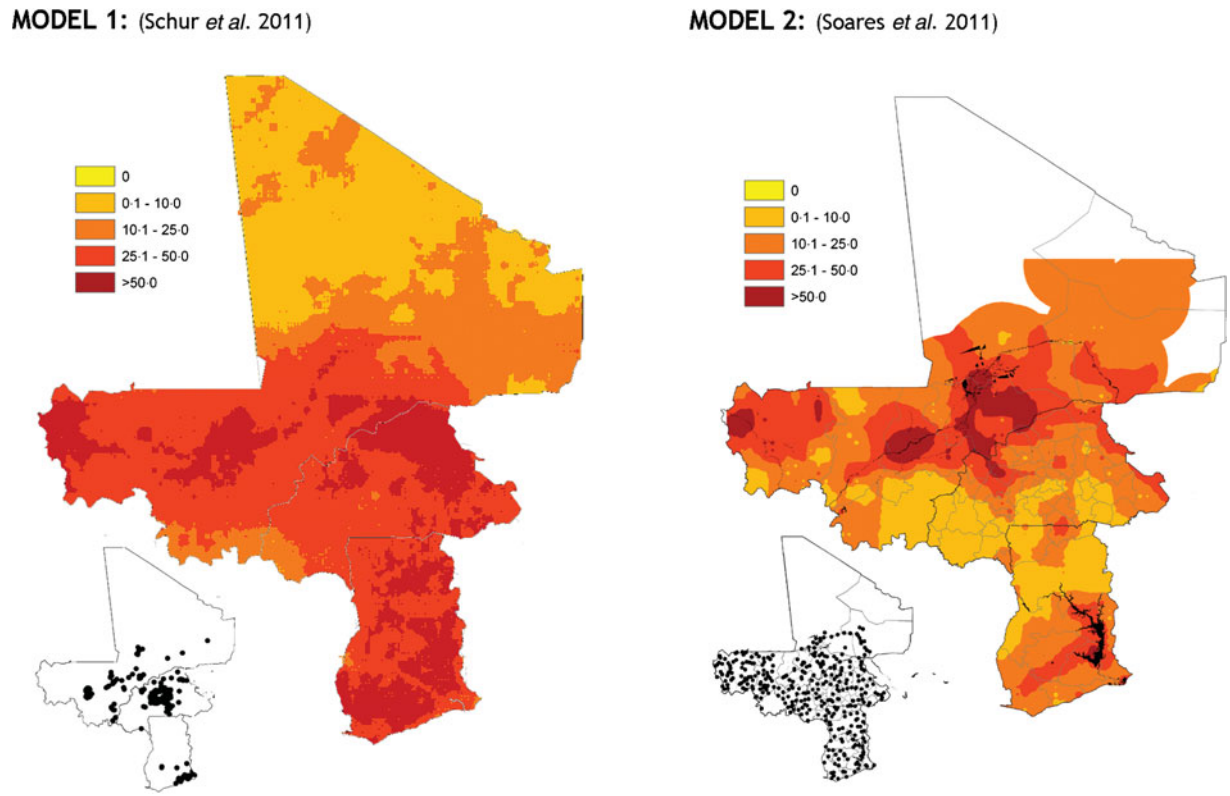

Fig. 4. Contrasting predictions of the distribution of Schistosomiasis haematobium generated using similarly robust MBG regression models, but different data. Model 1: Predicted prevalence of S. haematobium among individuals aged ⩽20 years during the period of 2000–2009, based on survey data from 16 West African countries. Model 2: Predicted prevalence of S. haematobium infection in boys aged 10–15 years in Burkina Faso, Mali and Ghana in 2004–2006. Inset maps show the location of survey data used in each model. Although overall trends are similar, these models show considerable differences in within country distributions, particularly in northern Burkina Faso, central Mali and much of Ghana. Figures are adapted from Schur *et al*. 2011 and Soares *et al*. 2011.

correlation between prediction locations can lead to gross underestimation of uncertainty when aggregating prediction estimates across regions, for example when producing country-level credible intervals (Goovaerts, 2001). A solution is to use joint or simultaneous simulation, which recreates appropriate spatial and temporal correlation in the predictive surface (Goovaerts, 2001), but which can be prohibitively intensive computationally, especially over large areas. Recently however, Gething *et al*. (2010) proposed an approximate algorithm for joint simulation, which they applied to a global scale MBG predictive model for malaria. Importantly, this approach ensured that aggregated estimates of national and regional burdens taken from continuous disease maps still maintained appropriate credible intervals.

Final predictive surfaces are also very dependent upon available data, a problem which becomes more pronounced as spatial heterogeneity increases. For example, a comparison of predictive risk maps of *S. haematobium* in West Africa generated using similarly robust MBG approaches but different datasets gives rise to maps which, whilst having similar regional trends, exhibit large differences in within-country distributions (Clements *et al*. 2009*b*; Schur *et al*. 2011*a*; Soares Magalhaes *et al*. 2011*a*). This point is clearly illustrated in Fig. 4. Similar differences are seen for different maps of *S. mansoni* in East Africa (Clements *et al*. 2010; Schur *et al*. 2011*b*). Such differences can have important implications for the planning of control activities and estimations of populations at risk, and more generally highlight difficulties in interpreting models from highly spatially heterogeneous data, no matter how sophisticated the underlying model.

Quantifying spatial dependency in order to undertake spatial sampling

Data available to the disease mapping community have usually been collected for other purposes, such as to investigate a specific research question or to determine national or sub-national prevalence estimates, using traditional, probability-based sampling methods (Levy and Lemeshow, 2008). Such design-based sampling forms the basis of most prevalence surveys for parasitic infections, including those for *Plasmodium* infection (Roll Back Malaria Monitoring and Evaluation Reference Group, 2005), STHs

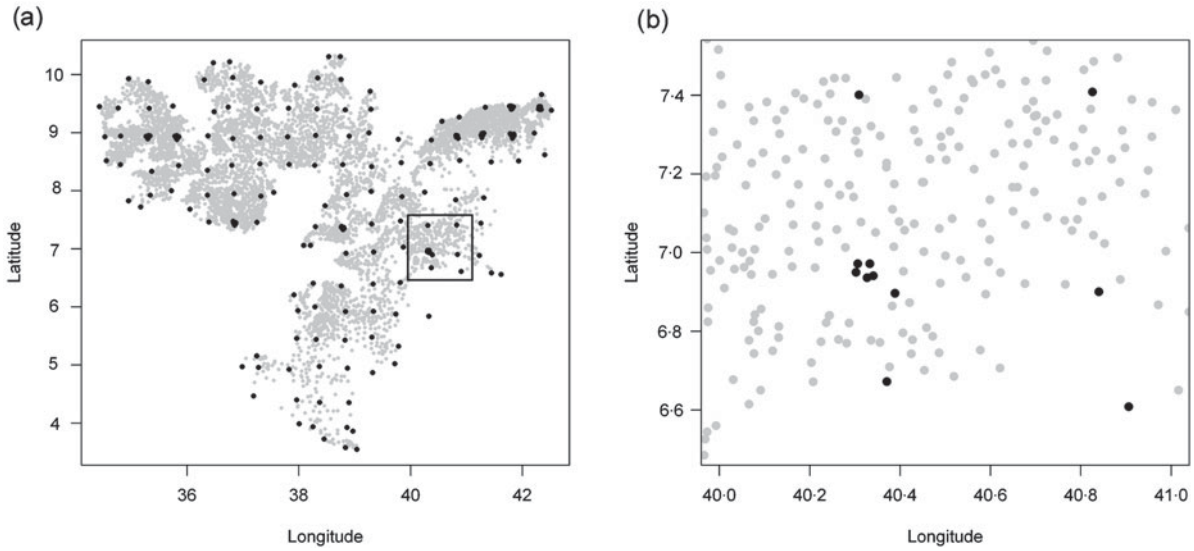

Fig. 5. (a) Illustrative example of the lattice plus close pairs design using a grid size of 50 km to select schools for surveys of *S. mansoni* in Oromia Regional State, Ethiopia. Dark points refer to selected schools and gray points to unselected schools. (b) A close-up of a region (black box in a) showing the locations of some of the clusters of closely located schools. Adapted from Sturrock *et al.* (2011).

(World Health Organization, 2006) and schistosomiasis (World Health Organization, 2006). We now know from other disciplines, such as geology and environmental sciences, that where MBG mapping is the eventual goal, such design-based sampling may be suboptimal and instead purposive (non-probability-based) sampling is generally more efficient (Brus and de Gruijter, 1997; de Gruijter *et al.* 2006). Such purposive sampling does not involve a random selection of sites, rather sites are selected based on their location or characteristics (i.e. selected to represent a particular altitude or ecological zone). The majority of early applications of spatial sampling came from ecological or soil science, but there are now an increasing number of applications in infectious disease epidemiology, which we review here.

Recent surveys of schistosome infection have adopted a stratified cluster random sampling design in order to obtain a spatially representative sample for subsequent risk mapping (Clements *et al.* 2006*a*, *b*, 2009*b*). A qualitative approach to selecting survey locations was adopted in a recent nationwide school survey of *Plasmodium* infection in Kenya (Gitonga *et al.* 2010), whereby the selection of schools in each district was made with a non-probability-based method to ensure a representative spatial spread of points. Furthermore, sites were over-sampled in sparsely populated districts to allow efficient spatial interpolation in these areas. An alternative approach is spatial coverage sampling, whereby survey sites are selected to ensure maximum coverage over a given survey region, accounting for its shape and previously collected data (de Gruijter *et al.* 2006). Van Groenigen *et al.* (2000) have, for example, used an iterative process to determine the configuration of sites for soil sampling that minimises the distance between any point in an area to its nearest survey site, such that an uniform distribution of sites over areas of any shape or size is obtained.

The selection of survey sites can also be informed by an understanding of the spatial structure of the outcome to be surveyed. For example, if the spatial structure of the data is known (based on a semi-variogram), it is possible to estimate the kriging variance of any configuration of survey sites *a priori*. This feature makes it possible to optimise the locations of surveys for spatial prediction before data are gathered (Van Groenigen *et al.* 1999; Brus *et al.* 2006). Where the autocorrelation structure is unknown or uncertain, pilot surveys can be conducted to quantify the spatial characteristics, which can then be used to optimize secondary data collection (Stein and Ettema, 2003). Alternatively, Diggle and Lophaven (2006) propose the use of a lattice plus close pairs design which is formed of a regular grid of points with some additional sites clustered around a selected number of grid sites. Fig. 5 provides an illustration of the lattice plus close pairs design for the selection of schools in a survey of *S. mansoni* in Ethiopia. First, a grid of a predefined size is placed over the survey region and those schools lying closest to the interstices of the grid are selected. Second, at some of these schools, the five closest neighbouring schools are selected. This stepwise selection ensures both a good spread of sample sites which are efficient for spatial interpolation and some closely located sites which are important for quantifying the spatial structure of the outcome. Recent work by Sturrock *et al.* (2011) demonstrates that use of a lattice plus close pairs design followed by kriging provided a more cost-effective approach to identify schools with high prevalence of *S. mansoni* compared to sampling a

small number of children in every school using lot quality assurance sampling (LQAS). This work shows that, whilst LQAS performed better than spatial sampling in identifying schools with a high prevalence, its cost-effectiveness in identifying such schools was lower.

Researchers have recently begun to investigate optimal survey designs that also incorporate covariate information (such as environmental and climatic factors) when collecting data for mapping based on MBG techniques. Unlike optimizing sampling for spatial interpolation alone, optimizing sampling for mapping using MBG with covariates requires a spread of points in so-called feature space (i.e. across the full range of included covariates) as well as across geographic space. To find a balance between these differing requirements, Hengl *et al*. (2003) propose an 'equal-range strategy' which ensures that equal numbers of sites are randomly selected in areas stratified by relevant covariates. By repeating this process multiple times, the sampling design with the most comprehensive spatial coverage can be chosen. Researchers have also used universal kriging variance to optimize surveys for soil, groundwater dynamics and radioactive releases (Heuvelink *et al*. 2006; Brus and Heuvelink, 2007; Melles *et al*. 2011). By finding the configuration of sites that minimises the universal kriging variance, a balance is struck between optimizing sampling across feature and geographic space. An appeal of this approach is that any number of covariates can be included, making it theoretically plausible to optimise surveys for multiple species with differing environmental niches. Such stratified sampling designs do however come with an important caveat: over-sampling in areas with particular characteristics (such as known higher infection prevalence) does risk invalidating standard geostatistical inference, as the implicit assumption of this approach is of non-preferential sampling. Nevertheless, given the recent increase in advocacy for integrated control of multiple parasite species, an investigation into optimal survey methods that consider both environmental correlates and spatial dependency for multiple species is clearly warranted.

DISCRETE SPATIAL VARIATION: UNDERSTANDING SPATIAL NEIGHBOURHOOD STRUCTURE

The methods described above depend on two major assumptions: that the underlying spatial process is continuous and that sufficiently detailed point-level data are available to capture this process. Certain data may only be available at a small-area level (for example, routine health surveillance data, access to water and sanitation, quality of local health services), and as such autocorrelation may only be apparent between immediate neighbours (i.e. based upon proximity rather than actual location). Alternatively, it may be difficult to obtain sufficient point-level data to model spatial variation in infection risk effectively, due to financial or practical constraints. In such instances, it may be more appropriate to make best use of spatially discrete data using hierarchical techniques. Whilst the primary focus of authors developing such techniques has often been in improving demographic and sociological data (Hentschel *et al*. 2000), or in modelling the distribution of non-communicable disease (Jackson *et al*. 2008*b*; Yiannakoulias *et al*. 2009; Danaei *et al*. 2011), many of these methods are also applicable to parasitological and related data only available at an area level, for example the number of malaria cases or intervention population coverage. In the next section we discuss some of these applications, drawing on examples from beyond the infectious disease literature where necessary.

Discrete spatial variation: modelling discrete outcome data

A major objective when modelling discrete disease data is obtaining statistically precise local estimates of the outcome of interest, whilst maintaining fine-scale geographic resolution. This can be a considerable challenge when outcomes are rare or sample sizes are small, as small stochastic differences between areas in the number of cases can result in large apparent differences in the distribution of the outcome. By smoothing high-resolution variability, model-based approaches can compromise between (overly) uncertain within-area estimates and (overly) simplified aggregated higher level estimates, thus stabilising estimation from areas with small populations or sample sizes (Goldstein, 1995). Such approaches, based upon the use of generalised linear mixed effect models, form the basis of *small area estimation*, with wide application in the analysis of health and social survey data (Ghosh and Rao, 1994; Ghosh *et al*. 1998; Richardson and Best, 2003; Asiimwea *et al*. 2011). They are inherently non-spatial, borrowing information across all areas without considering spatial location, and smoothing to the global mean. However, they can easily be extended to include additional model complexity such as spatial dependence using discrete spatial smoothing models based on proximity. Such models, which assume positive spatial correlation between observations, essentially borrow more information from close neighbours than those further away, and so smooth local rates towards local, neighbouring values (Waller and Carlin, 2010).

One of the first examples of spatial discrete modelling is provided by Clayton and Kaldor (1987), who developed a Poisson regression model with area-specific random intercepts defined using a *conditional autoregressive* (CAR) structure to model standardised mortality ratios (in this case, cancer rates). By this

approach, the area-specific random effect is generated using a simple adjacency weights matrix, such that for each observation the associated random parameter has a weighted mean given by a simple average of its defined neighbours and a conditional variance inversely proportion to the number of neighbours. This model has since been extended to a fully Bayesian formulation (Besag *et al*. 1991) and can be readily implemented in standard Bayesian inference software. Via this flexible inference platform, spatial CAR models can be structured to allow autocorrelation between adjacent neighbours only, or to allow spatial smoothing to extend to more distant neighbours (Wakefield, 2004; MacNab, 2010), and can be extended to allow for: estimation of spatially varying covariates; prediction of missing data; inclusion of both spatial and non-spatial dependency; and inclusion of spatio-temporal and multivariate outcome covariance structures (Mollie, 1996; Waller and Carlin, 2010).

A similar but less commonly used class of models are the spatial multiple membership (MM) models, which examine to what extent a latent spatially distributed variable can explain the outcomes of interest (Breslow and Clayton, 1993; Goldstein, 1995; Langford *et al*. 1999; Browne *et al*. 2001). In contrast to CAR models, the spatial dependence here is modelled through the multiple membership relationship, with an independent area-level random effect.

CAR and MM approaches have most widely been applied to modelling rates of rare non-communicable diseases, such as cancer and heart disease (Lawson, 2006), usually in a developed country setting where comprehensive disease registry data are available. However in tropical epidemiology, analyses of routinely collected surveillance data have used spatial CAR and MM models on varying scales to investigate incidence of malaria in Zimbabwe, South Africa and China (Kleinschmidt *et al*. 2002; Mabaso *et al*. 2006; Clements *et al*. 2009*c*) and dengue in Rio de Janeiro, Brazil (Teixeria and Cruz, 2011) in addition to assisting in the geographical targeting of schistosomiasis control in Tanzania using collated questionnaire data (Clements *et al*. 2008*b*).

There are a number of substantial methodological challenges when modelling spatially discrete data. The first of these concerns assumptions made regarding the underlying spatial process. Discrete spatial variation models consider the location of data points in terms of proximity only, rather than as literal positions, and as such the model is valid only for included data; validity is not necessarily preserved if further locations are added to the data (Diggle, 2004). For this reason, these models are not appropriate for spatial prediction in new locations (interpolation). As for methods quantifying continuous spatial dependency, discrete spatial approaches also model global spatial structure and thus assume that the degree of correlation between neighbouring units is consistent across the study area, and in all directions. This issue was recently tackled in part by Reich and colleagues (2007), who developed a 2NRCAR (CAR prior with two neighbour relations) model that is able to accommodate two difference classes of neighbour relations (e.g. east-west and north-south) (Reich *et al*. 2007), although to our knowledge this has yet to be applied in an epidemiological context. A third issue concerns the often arbitrarily defined units of representation available for geographical analysis. This is known as the *modifiable areal unit problem*, by which for any specified number of spatial units, there are many ways of defining the boundaries of these units, which can produce very different results (Openshaw and Taylor, 1979). For example, spatial anomalies may go undetected if the scale of the underlying spatial heterogeneity is smaller than the area unit available. Although this cannot be completely overcome, careful consideration of CAR and MM models does allow smoothing between units, thus blurring the concept of a discrete unit of analysis.

Discrete spatial variation: combining point and area level data

In many instances, disease data may be available at a point level (e.g. survey or sentinel site data), although covariate information may be available only at an area level. For example, although recognised as important factors influencing the distribution of parasitic diseases at varying spatial scales, data on factors such as water supply, sanitation and hygiene (WASH), ownership and use of bednets, coverage of interventions such as mass drug administration, and poverty and deprivation indicators may only be available at district and regional levels (Esrey *et al*. 1991; Kazembe *et al*. 2007; Soares Magalhaes *et al*. 2011*a*). Alternatively, disease information may only be available aggregated at area-levels for rare outcomes, although individual-level survey data describing the distribution of explanatory factors may be readily available. Despite the aggregated nature of such data, careful analysis can provide information about the relationships between area-level risks and point-level outcomes. This process is known as ecological inference (Richardson and Monfort, 2000; Jackson *et al*. 2006, 2008*a*), and can be valuable when the effect of a variable is believed to operate through its area-level average (sometimes termed a contextual effect) (Begg and Parides, 2003). For instance, control policies for many parasitic infections implemented at the district level have been shown to benefit indirectly those individuals who have not participated. Helminth infection prevalence in non-compliant or non-targeted individuals, for example, is typically seen to reduce after the administration of community or school-based mass chemotherapy (Bundy *et al*. 1990; Chan *et al*. 1997; Vanamail *et al*. 2005; Mathieu *et al*. 2006;

El-Setouhy *et al*. 2007). This is primarily due to a reduction in the force of transmission, analogous to the 'herd immunity effect' seen for vaccines. However, in many instances area-level exposure-response relationships may not accurately reflect associations at the community or household level, a process known as the ecological fallacy or ecological bias (Morgenstern, 2008). For example, an individual-level association between individual socio-economic indicators and regional rates of disease does not necessarily imply an effect of socio-economic status on individual infection status. It can equally be caused by other confounding factors.

The magnitude of ecological bias depends upon the degree of within-area variability in exposures and confounders – if there is no variability, all individuals will experience the same degree of exposure, and so there will be no ecological bias (Wakefield and Lyons, 2010). The only way to truly overcome the problem of ecological bias therefore is to supplement aggregate data with samples of data at the individual level, which on their own may be too sparse to accurately capture geographic variation but can provide an indication of intra-unit variation. Several theoretical methods have been proposed to do this, which can be used to address bias and separate individual and contextual effects when either the outcome or the exposure measure is available at an ecological level (Prentice and Sheppard, 1995; Steel and Holt, 1996; Lasserre *et al*. 2000; Best *et al*. 2001; Wakefield and Salway, 2001; Glynn *et al*. 2008). The so-called aggregate data method, for example, estimates individual-level exposure effects by regressing population-based disease rates on covariate data from survey samples in each population group (Prentice and Sheppard, 1995). An alternative approach, termed hierarchical related regression, assumes a distribution for within-area variability in exposure, and fits the implied model to aggregate data combined with small samples of individual-level exposure and outcome data (Jackson *et al*. 2006, 2008*a*).

Both of these approaches however are not inherently spatial, although the generalised linear model frameworks upon which they are based can in theory be adapted to include multiple levels of aggregation and spatial dependency between baseline risks. For example, spatial CAR models have been combined with aggregate data methods to account better for exposure effect when modelling spatially heterogeneous breast cancer rates (Guthrie *et al*. 2002), and with hierarchical related regression when investigating sensitivity of environmental exposure and childhood leukaemia data to ecological bias (Best *et al*. 2001). These approaches have yet to be applied within an infectious disease context, but they do have great potential for application in spatial parasite epidemiology, for example, to improve evaluation of intervention programmes using implementation-level, sentinel site and cluster-level data. This may be problematic in practice, as obtaining data on the same population from different sources may lead to considerable inconsistencies, such as differences in variable definition and reporting, timing, and even geographical boundaries between levels, which may in turn lead to unreliable conclusions (Jackson *et al*. 2008*a*).

SPATIAL POINT PROCESSES: INVESTIGATING SPATIAL CLUSTERING

The global spatial statistics described above provide an important set of epidemiological tools to inform whether spatial heterogeneity is present throughout spatially sampled measurement data. This in turn informs optimal model building and can be used to conduct spatial interpolation and prediction. These methods cannot be used to delineate explicitly the locations of individual clusters and typically make the assumption that the magnitude and scale of clustering is equal throughout the study region. Identifying the propensity for spatial clustering to occur, or the physical locations of individual clusters, is vital for both identifying areas with higher than expected underlying risk and detecting outbreaks, as well as determining the optimal spatial location and scale of interventions. It is also important to emphasise that, whilst identifying the *presence* of spatial clusters may be useful, it is perhaps more important epidemiologically also to gain a deeper understanding of the *determinants* of this clustering (Rothman, 1990; Alexander and Boyle, 2001).

Point process statistics aims to analyse the explicit location of events distributed in space under an assumption that the spatial pattern is random (i.e. the locations and numbers of points are not fixed). In parasite epidemiology and ecology, such data typically present either as point locations within a given study area (for example, residence of incident cases or location of vector breeding sites) or as counts of cases from administrative districts partitioning the study area (Waller, 2010). Point process approaches typically play two distinct roles in the analysis of such data. First, they can be used to investigate the general tendency of points to exist near points, providing a global measure of clustering averaged across the observed point locations. For example, we may be interested in whether summary global measures of clustering for disease cases differ significantly from those for the general at-risk population, and thus whether there is an overall tendency for cases to occur near other cases rather than to occur homogeneously among the population at risk. Secondly, they can be used to delineate explicitly the locations of individual clusters, or anomalous collections of points. Such clusters might be assumed to occur anywhere within the study area, or may be focused, centred around predefined foci of putatively increased risk (e.g. vector or intermediate host breeding sites) (Besag and Newell,

1991). In both scenarios, analysis strategies usually build upon ideas of testing the hypothesis of complete spatial randomness (CSR), such as the realisation of a homogeneous Poisson process (Isham, 2010). In epidemiology, this is often complicated by the heterogeneous distribution of the at-risk population, with the null model of interest no longer being CSR, but one of spatially constant risk, and thus knowledge of the underlying distribution of the population at-risk, or of appropriate non-infected controls, is essential (Waller, 2010).

There is a rich body of literature addressing various approaches for spatial point processes, although their existing application to the ecology and epidemiology of parasitic diseases has to date been somewhat limited in scope. This is partly due to the fact that model fitting is not straightforward and often computationally complex. In addition, many approaches have been derived from a rather mathematical perspective and are not necessarily appropriate in an ecological or epidemiological context. We do not therefore provide a comprehensive review of all available methods here, but instead compare and illustrate some of the most popular contemporary approaches for detecting clustering and clusters in parasite epidemiology. More complete general reviews of methods and applications appear elsewhere (Diggle, 2003; Waller and Gotway, 2004; Gelfand *et al*. 2010).

Detecting global clustering in point pattern data

Investigations of global clustering usually start by testing benchmark hypotheses regarding the underlying process – i.e. is a point or case equally likely to occur at any location? Such hypothesis testing approaches include Ripley's K function (Ripley, 1976) and the related L function (Besag, 1977), which provide a measure of the (scaled) number of additional events expected within distance h of a randomly selected point. Plotting K as a function of h can thus be used to describe characteristics of the point process at many different spatial scales. More recent application of these methods to infectious disease epidemiology includes investigation of urban dynamics of dengue epidemics in the Brazilian city of Belo Horizone (Almeida *et al*. 2008), identification of epidemic hotspots for malaria in the Kenyan western highlands (Wanjala *et al*. 2011), and investigation of spatial clustering of households with seropositive children during evaluation of targeted screening strategies to detect *Trypanasoma cruzi* infection in Peru (Levy *et al*. 2007).

In recent years, summary test-based measures for spatial point processes have been joined by model-based approaches, using both frequentist and Bayesian inference platforms. As with models of continuous and discrete spatial variation, spatial point process models provide an objective and efficient statistical framework for investigating spatial heterogeneity, whilst adjusting for spatially varying risk factors. The more fundamental of these models are based upon the non-homogeneous Poisson process, which assumes a lack of interaction between points (i.e. complete spatial randomness) but still allows point intensity to vary over space. This assumption of independence between points may be appropriate if all spatial variation can be explained by observed or unobserved risk factors, such as climate, topography and inherited genetic risk, which are themselves spatially correlated, an assumption often fitting for non-infectious diseases (Diggle, 2001). For infectious diseases however, stochastic spatial dependence may still remain between points even after accounting for covariates. For example, foci of parasite transmission can perpetuate and amplify spatial heterogeneity, with heavily infected individuals shedding large numbers of parasites into the environment, increasing risk for those living in close proximity. This can be modelled by hierarchical processes derived from the above non-homogeneous Poisson process, including Poisson cluster and Cox processes. These flexible models are 'doubly stochastic' in that they also include a random intensity function, which may be taken to be any random spatial process. As such, they are very effective for describing residual positive association between points. For example, sophisticated log-Gaussian Cox process models, which assume a Gaussian random field for the logarithm of the density function (analogous to the Gaussian geostatistical models detailed above), have been applied to investigations of tick-borne encephalitis in the Czech Republic (Benes *et al*. 2011), and spatio-temporal surveillance of non-specific gastroenteric disease in the UK (Diggle *et al*. 2005). However, despite increased application in other areas of spatial epidemiology (Lawson, 2006) to our knowledge we have yet to see these potentially exciting methods being applied to parasite ecology.

Detecting local clusters in point pattern data

Although restricted to hypothesis testing, methods for detecting local clusters in point pattern data are perhaps the most popular point process statistical techniques currently used by parasite epidemiologists, with a variety of methods available for exploring and identifying clusters in both point and aggregated data (for an excellent review see Pfeiffer *et al*. 2008). The more popular approaches involve spatial scan statistics, the most developed of these being Kulldorff's spatial scan statistic (Kulldorff and Nagarwalla, 1995). This method constructs a series of circles of increasing size around each data location and compares the level of risk within each circle to that outside using a likelihood ratio test. A computationally convenient Monte Carlo simulation is then

used to generate permutations of the observed number of cases across the entire set of data locations, allowing for testing of the null hypothesis of complete spatial randomness. Other tests that are specifically designed to detect clusters around a source, so called focused tests, include Stone's test and Diggle's test, which have been used to explore clustering of cancers around industrial plants (Stone, 1988; Diggle, 1990). Results for all spatial scanning statistics are conditional only over the discrete set of data locations available, an important factor to bear in mind when data locations are based on only a sample of all potential locations (Waller and Gotway, 2004).

Kulldorff's spatial scan statistic was first used to explore clustering of leukaemia cases in New York (Kulldorff and Nagarwalla, 1995) and has since been used to investigate clustering for a variety of parasitic diseases, including leishmaniasis (Ryan *et al*. 2006; Schriefer *et al*. 2009), lymphatic filariasis (Washington *et al*. 2004) schistosomiasis (Peng *et al*. 2010), and trypanosomiasis (Fèvre *et al*. 2001; Gorla *et al*. 2009). The parasitic disease for which spatial scan statistics have been most widely used is malaria. Brooker *et al*. (2004*a*) used spatial scan statistics to identify clusters of malaria cases during an epidemic in the highlands of Kenya. More recently, Bousema *et al*. (2010) used cluster statistics to illustrate that clustering of seropositive individuals can be used to identify 'hot spots' of high malaria disease incidence in Tanzania. Similarly, Cook *et al*. (2011) used spatial scan statistics to explore clustering of infection and seroprevalence in different age-groups to illustrate spatial heterogeneities in effectiveness of malaria control on Bioko Island in Equatorial Guinea. In a slightly different use of the test, Fevre *et al*. (2001) used spatial scan statistics to show that a cattle market was the likely source of an outbreak of *Trypanosoma brucei rhodesiense* sleeping sickness in Uganda.

Over the last two decades there have been several developments in spatial scan statistics allowing an exploration of clustering in a variety of different data types including multinomial (Jung *et al*. 2010) and ordinal (Jung *et al*. 2007) data as well as detection of non-spherical clusters (Tango and Takahashi, 2005; Kulldorff *et al*. 2006; Cançado *et al*. 2010). Where data allow, it is also possible to explore clustering through time as well as space, either by exploring years separately (Bejon *et al*. 2010) or by considering circular clusters as cylinders that span time (Kulldorff, 2001; Kulldorff *et al*. 2005). Such spatio-temporal cluster analyses have been used to explore space-time clustering in diarrhoea surveillance data from emergency departments in New York (Kulldorff, 2001) and malaria in highland Kenya (Ernst *et al*. 2006) and South Africa (Coleman *et al*. 2009). Recent advances in the use of space-time cluster detection analyses additionally include the use of statistical models to better define underlying risk (Robertson *et al*. 2010*a*). For example, Kleinman *et al*. (2005) use a generalized linear mixed model, which uses information on census tract, day of the week, month of the year, holiday status and secular time, to describe the underlying spatio-temporal distribution of reported lower respiratory tract infections in Eastern Massachusetts. Spatial scan statistics are then applied to the data to detect anomalies from model-predicted risk, thus helping to avoid false alarms in areas of explained high incidence.

Other current areas of research include the incorporation of human movement data (Tatem *et al*. 2009; Robertson *et al*. 2010*b*), which allows researchers to account for the fact that detection of risk in one area does not necessarily correspond with the initial location of infection. With the increasing availability of data from mobile phones and other technological devices with inbuilt GPS devices, such analyses will no doubt become progressively incorporated into disease surveillance systems.

Despite increasingly sophisticated cluster detection methods, a couple of important limitations should be borne in mind when embarking on disease cluster detection. First, systematic bias in study design – including inaccurate and non-standardised case definition, error in exposure measurement or inadequate control of confounding variables – can all lead to high possibility of false-positive clusters (Rothman, 1990). Second, such analyses can be very sensitive to the rather arbitrary assumptions required, such as selection of the scanning window shape and size and upper cluster size threshold, which in turn can lead to different results.

CONCLUSION

With the widespread use of GPS and GIS, and the availability of high resolution environmental data, spatial aspects of parasites and their vectors and intermediate hosts are becoming increasingly well understood. Such an understanding has been facilitated by the development of a wealth of statistical methods that have improved our ability both to describe and to predict parasite distributions over varying spatial scales. A growing number of these methods sit within a Bayesian framework, providing a more flexible approach to modelling that is able to account for both spatial and non-spatial uncertainty effectively. The choice of statistical method and approach used however should be guided both by the type of data available and the scientific questions of interest. A central tenet of statistics that still holds true for spatial analyses should always be borne in mind: that there is no such thing as a "correct" model, and that instead the best model is one that provides a good fit of the data as economically as possible (Bailey and Gatrell, 1995; Pfeiffer, 1996; Diggle, 2004). Judgement is required to reach the correct balance between an over-simplistic model, which may

risk invalid inferences, and an over-elaborate model, which may be inefficient and difficult to validate (Altham, 1984). Nevertheless, applied spatial statistics remains an active area of research, continually providing parasitologists, ecologists and epidemiologists with novel approaches. As such, careful consideration of spatial location is rapidly becoming a routine component of parasite ecology and epidemiology.

ACKNOWLEDGEMENTS

We are grateful to the reviewers for their expert comments which greatly improved the quality of this review. RLP and HJWS are supported by grants from the Bill & Melinda Gates Foundation and GlaxoSmithKline, respectively. RJSM is funded by a Post-doctoral Research Fellowship from the University of Queensland (41795457). ACAC is funded by an Australian National Health and Medical Research Council Career Development Award (631619). SJB is funded by a Research Career Development Fellowship (081673) from the Wellcome Trust.

REFERENCES

Alexander, F. E. and Boyle, P. (2001). Do cancers cluster? In *Spatial Epidemiology, Methods and Applications* (eds. Elliott, P., Wakefield, J. C., Best, N. G. & Briggs, D. J.), pp. 302–316. Oxford University Press, Oxford.

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

Almeida, M. C., Assuncao, R. M., Proietti, F. A. and Caiaffa, W. T. (2008). [Intra-urban dynamics of dengue epidemics in Belo Horizonte, Minas Gerais State, Brazil, 1996–2002]. *Cadernos de Saúde Pública* **24**, 2385–2395.

Altham, P. M. E. (1984). Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society (Series B)* **46**, 118–119.

Anderson, R. M. (1993). Epidemiology. In *Modern Parasitology* (ed. Cox, F. E. G.), pp. 75–116. Blackwell Science, Oxford.

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford.

Asiimwea, J. B., Jehopioa, P., Atuhairea, K. L. and Mbonye, A. K. (2011). Examining small area estimation techniques for public health intervention: Lessons from application to under-5 mortality data in Uganda. *Journal of Public Health Policy* **32**, 1–15.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Longman, Harlow.

Beck-Worner, C., Raso, G., Vounatsou, P., N'Goran, E. K., Rigo, G., Parlow, E. and Utzinger, J. (2007). Bayesian spatial risk prediction of *Schistosoma mansoni* infection in western Cote d'Ivoire using a remotely-sensed digital elevation model. *American Journal of Tropical Medicine and Hygiene* **76**, 956–963.

Begg, M. D. and Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine* **22**, 2591–2602.

Bejon, P., Williams, T. N., Liljander, A., Noor, A. M., Wambua, J., Ogada, E., Olotu, A., Osier, F. H. A., Hay, S. I., Farnert, A. and Marsh, K. (2010). Stable and unstable malaria hotspots in longitudinal cohort studies in Kenya. *PLoS Medicine* **7**, e1000304.

Benes, V., Bodlak, K., Moller, J. and Waagepetersen, R. (2011). A case study on point process modelling in disease mapping. *Image Analysis and Stereology* **24**, 159–168.

Besag, J. (1977). Discussion of "Modelling spatial patterns" by B.D.Riley. *Journal of the Royal Statistical Society (Series B)* **39**, 193–195.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society (Series A)* **154**, 237–333.

Besag, J., York, J. C. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

Best, N., Cockings, S., Bennett, J., Wakefield, J. and Elliott, P. (2001). Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society (Series A)* **164**, 155–174.

Booth, M., Vounatsou, P., N'Goran, E. K., Tanner, M. and Utzinger, J. (2003). The influence of sampling effort and the performance of the Kato-Katz technique in diagnosing *Schistosoma mansoni* and hookworm co-infections in rural Cote d'Ivoire. *Parasitology* **127**, 525–531.

Bousema, T., Drakeley, C., Gesase, S., Hashim, R., Magesa, S., Mosha, F., Otieno, S., Carneiro, I., Cox, J., Msuya, E., Kleinschmidt, I., Maxwell, C., Greenwood, B., Riley, E., Sauerwein, R., Chandramohan, D. and Gosling, R. (2010). Identification of hot spots of malaria transmission for targeted malaria control. *Journal of Infectious Diseases* **201**, 1764–1774.

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

Brooker, S. (2007). Spatial epidemiology of human schistosomiasis in Africa: risk models, transmission dynamics and control. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **101**, 1–8.

Brooker, S., Alexander, N., Geiger, S., Moyeed, R. A., Stander, J., Fleming, F., Hotez, P. J., Correa-Oliveira, R. and Bethony, J. (2006). Contrasting patterns in the small-scale heterogeneity of human helminth infections in urban and rural environments in Brazil. *International Journal for Parasitology* **36**, 1143–1151.

Brooker, S., Clarke, S., Njagi, J. K., Polack, S., Mugo, B., Estambale, B., Muchiri, E., Magnussen, P. and Cox, J. (2004*a*). Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. *Tropical Medicine & International Health* **9**, 757–766.

Brooker, S. and Clements, A. C. (2009). Spatial heterogeneity of parasite co-infection: Determinants and geostatistical prediction at regional scales. *International Journal for Parasitology* **39**, 591–597.

Brooker, S., Kabatereine, N. B., Tukahebwa, E. M. and Kazibwe, F. (2004*b*). Spatial analysis of the distribution of intestinal nematode infections in Uganda. *Epidemiology and Infection* **132**, 1065–1071.

Brooker, S., Pullan, R. L., Gitonga, C. W., Ashton, R. A., Kolaczinski, J. H., Kabatereine, N. B. and Snow, R. W. (2012). Epidemiology of *Plasmodium*-helminth coinfection in contrasting transmission settings across East Africa. *Journal of Infectious Diseases* **205**(5), 841–852.

Browne, W. J., Goldstein, H. and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling* **1**, 103–124.

Brus, D. J. and de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma* **80**, 1–44.

Brus, D. J., de Gruijter, J. J. and van Groenigen, J. W. (2006). Designing Spatial Coverage Samples Using the k-means Clustering Algorithm. In *Developments in Soil Science* Vol. 31 (eds. Lagacherie, P., McBratney, A. B. & Voltz, M.), pp. 183–192. Elsevier, Amsterdam.

Brus, D. J. and Heuvelink, G. B. M. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* **138**, 86–95.

Bundy, D. A., Wong, M. S., Lewis, L. L. and Horton, J. (1990). Control of geohelminths by delivery of targeted chemotherapy through schools. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **84**, 115–120.

Cançado, A. L., Duarte, A. R., Duczmal, L. H., Ferreira, S. J., Fonseca, C. M. and Gontijo, E. C. (2010). Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International Journal of Health Geographics* **9**, 55.

Chan, M. S., Bradley, M. and Bundy, D. A. (1997). Transmission patterns and the epidemiology of hookworm infection. *International Journal of Epidemiology* **26**, 1392–1400.

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

Clements, A. C., Bosque-Oliva, E., Sacko, M., Landoure, A., Dembele, R., Traore, M., Coulibaly, G., Gabrielli, A. F., Fenwick, A. and Brooker, S. (2009*a*). A comparative study of the spatial distribution of schistosomiasis in Mali in 1984–1989 and 2004–2006. *PLoS Neglected Tropical Diseases* **3**, e431.

Clements, A. C., Firth, S., Dembele, R., Garba, A., Toure, A., Sacko, M., Landoure, A., Bosque-Oliva, E., Barnett, A. G., Brooker, S. and Fenwick, A. (2009*b*). Use of Bayesian geostatistical prediction to estimate local variations in *Schistosoma haematobium* infection in West Africa. *Bulletin of the World Health Organization* **87**, 921–929.

Clements, A. C., Garba, A., Sacko, M., Toure, S., Dembele, R., Landoure, A., Bosque-Oliva, E., Gabrielli, A. F. and Fenwick, A.

(2008*a*). Mapping the probability of schistosomiasis and associated uncertainty, West Africa. *Emerging Infectious Diseases* **14**, 1629–1632.

Clements, A. C., Lwambo, N. J., Blair, L., Nyandindi, U., Kaatano, G., Kinung'hi, S., Webster, J. P., Fenwick, A. and Brooker, S. (2006*a*). Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. *Tropical Medicine and International Health* **11**, 490–503.

Clements, A. C., Moyeed, R. and Brooker, S. (2006*b*). Bayesian geostatistical prediction of the intensity of infection with *Schistosoma mansoni* in East Africa. *Parasitology* **133**, 711–719.

Clements, A. C. A., Barnett, A. G., Cheng, Z. W., Snow, R. W. and Zhou, H. N. (2009*c*). Space-time variation of malaria incidence in Yunnan Province, China. *Malaria Journal* **8**, 180.

Clements, A. C. A., Brooker, S., Nyandindi, U., Fenwick, A. and Blair, L. (2008*b*). Bayesian spatial analysis of a national urinary schistosomiasis questionnaire to assist geographic targeting of schistosomiasis control in Tanzania, East Africa. *International Journal for Parasitology* **38**, 401–415.

Clements, A. C. A., Deville, M., Ndayishimiye, O., Brooker, S. and Fenwick, A. (2010). Spatial co-distribution of neglected tropical diseases in the east African great lakes region: revisiting the justification for integrated control. *Tropical Medicine and International Health* **15**, 198–207.

Coleman, M., Coleman, M., Mabuza, A. M., Kok, G., Coetzee, M. and Durrheim, D. N. (2009). Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. *Malaria Journal* **8**, 68.

Cook, J., Kleinschmidt, I., Schwabe, C., Nseng, G., Bousema, T., Corran, P. H., Riley, E. M. and Drakeley, C. J. (2011). Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, Equatorial Guinea. *PLoS One* **6**, e25137.

Crainiceanu, C. M., Diggle, P. J. and Rowlingson, B. (2008). Bivariate binomial spatial modeling of *Loa loa* prevalence in tropical Africa. *Journal of the American Statistical Association* **103**, 21–37.

Crainiceanu, C. M., Ruppert, R. and Wand, M. P. (2005). Bayesian analysis for penalised spline regression using WinBUGS. *Journal of Statistical Software* **14**, 1–24.

Cressie, N. (1991). *Statistics for Spatial Data*. Wiley, New York.

Cressie, N., Calder, C. A., Clark, J. S., van Hoes, J. M. and Wilke, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications* **19**, 553–570.

Danaei, G., Finucane, M. M., Lin, J. K., Singh, G. M., Paciorek, C. J., Cowan, M. J., Farzadfar, F., Stevens, G. A., Lim, S. S., Riley, L. M. and Ezzati, M. (2011). National, regional, and global trends in systolic blood pressure since 1980: systematic analysis of health examination surveys and epidemiological studies with 786 country-years and 5·4 million participants. *Lancet* **6736**, 62036–62033.

de Gruijter, J. J., Brus, D. J., Bierkens, M. F. P. and Knotters, M. (2006). *Sampling for Natural Resource Monitoring*. Springer-Verlag, Berlin.

Diggle, P. J. (1990). A point process modeling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society (Series A)* **153**, 349–362.

Diggle, P. J. (1996). Spatial analysis in biometry. In *Advances in Biometry* (eds. Armitage, P. & David, H. A.), pp. 363–384. Wiley, New York.

Diggle, P. J. (2001). Overview of statistical methods for disease mapping and its relationship to cluster detection. In *Spatial Epidemiology* (eds. Elliot, P., Wakefield, J., Best, N. & Briggs, D.), Oxford University Press, Oxford, UK.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. 2nd edition. Arnold, London.

Diggle, P. J. (2004). Spatial Statistics in the Biomedical Sciences. In *GIS and Spatial Analysis in Veterinary Science* (eds. Durr, P. A. & Gatrell, A. C.), CABI Publishing, Wallingford, UK.

Diggle, P. and Lophaven, S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics* **33**, 53–64.

Diggle, P. J., Moyeed, R. A. and Tawn, J. A. (1998). Model-based geostatistics. *Applied Statistics* **47**, 299–350.

Diggle, P. J., Rowlingson, B. and Su, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* **16**, 423–434.

Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J. H., Boussinesq, M. and Molyneux, D. H. (2007). Spatial modelling and the prediction of *Loa loa* risk: decision making under uncertainty. *Annals of Tropical Medicine and Parasitology* **101**, 499–509.

El-Setouhy, M., Abd Elaziz, K. M., Helmy, H., Farid, H. A., Kamal, H. A., Ramzy, R. M. R., Shannon, W. D. and Weil, G. J. (2007). The effect of compliance on the impact of mass drug administration for elimination of Lymphatic Filariasis in Egypt. *American Journal of Tropical Medicine and Hygiene* **77**, 1069–1073.

Engels, D. and Savioli, L. (2006). Reconsidering the underestimated burden caused by neglected tropical diseases. *Trends in Parasitology* **22**, 363–366.

Ernst, K. C., Adoka, S. O., Kowuor, D. O., Wilson, M. L. and John, C. C. (2006). Malaria hotspot areas in a highland Kenya site are consistent in epidemic and non-epidemic years and are associated with ecological factors. *Malaria Journal* **5**, 78.

Esrey, S. A., Potash, J. B., Roberts, L. and Shiff, C. (1991). Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, schistosomiasis and trachoma. *Bulletin of the World Health Organization* **69**, 609–621.

Farnert, A. (2008). *Plasmodium falciparum* population dynamics: only snapshots in time? *Trends in Parasitology* **24**, 340–344.

Fèvre, E. M., Coleman, P. G., Odiit, M., Magona, J. W., Welburn, S. C. and Woolhouse, M. E. (2001). The origins of a new *Trypanosoma brucie rhodesiense* sleeping sickness outbreak in eastern Uganda. *Lancet* **358**, 625–628.

Gelfand, A. E., Diggle, P. J., Fuentes, M. and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Chapman & Hall/CRC Press, Boca Raton, USA.

Gemperli, A. (2003). Development of spatial statistical methods for modeling point-referenced spatial data in malaria epidemiology. *Swiss Tropical Institute* Vol. Doctoral Dissertation pp. 111–127. University of Basel.

Gething, P. W., Patil, A. P. and Hay, S. I. (2010). Quantifying aggregated uncertainty in *Plasmodium falciparum* malaria prevalence and populations at risk via efficient space-time geostatistical joint simulations. *PloS Computational Biology* **6**(4), e1000724.

Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I. R., Johnston, G. L., Tatem, A. J. and Hay, S. I. (2011). A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malaria Journal* **10**, 378.

Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998). Generalised linear models for small area estimation. *Journal of the American Statistics Association* **93**, 273–282.

Ghosh, M. and Rao, J. K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science* **7**, 457–511.

Gitonga, C., Karanja, P., Kihara, J., Mwanje, M., Juma, E., Snow, R., Noor, A. and Brooker, S. (2010). Implementing school malaria surveys in Kenya: towards a national surveillance system. *Malaria Journal* **9**, 306.

Glynn, A., Wakefield, J., Handcock, M. and Richardson, T. (2008). Alleviating linear ecological bias and optimal design with subsample data. *Journal of the Royal Statistical Society (Series A)* **171**, 179–202.

Goldstein, H. (1995). *Multilevel Statistical Models*, second ed., Arnold, London, UK.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma* **103**, 3–26.

Gorla, D. E., Porcasi, X., Hrellac, H. and Catala, S. S. (2009). Spatial stratification of house infestation by *Traitoma infestans* in La Rioja, Argentina. *American Journal of Tropical Medicine and Hygiene* **80**, 405–409.

Gosoniu, L., Veta, A. M. and Vounatsou, P. (2010). Bayesian geostatistical modeling of Malaria Indicator Survey data in Angola. *PLoS One* **5**, e9322.

Gosoniu, L., Vounatsou, P., Sogoba, N., Maire, N. and Smith, T. (2009). Mapping malaria risk in West Africa using a Bayesian nonparametric non-stationary model. *Computational Statistics and Data Analysis* **53**, 3358–3371.

Guthrie, K. A., Sheppard, L. and Wakefield, J. (2002). A hierarchical aggregate data model with spatially correlated disease rates. *Biometrics* **58**, 898–905.

Hay, S. I., Guerra, C. A., Gething, P. W., Patil, A. P., Tatem, A. J., Noor, A. M., Kabaria, C. W., Manh, B. H., Elyazar, I. R., Brooker, S., Smith, D. L., Moyeed, R. A. and Snow, R. W. (2009). A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Medicine* **24**, e1000048.

Hay, S. I., Okiro, E. I., Gething, P. W., Patil, A. P., Tatem, A. J., Guerra, C. A. and Snow, R. W. (2010). Estimating the Global Clinical Burden of *Plasmodium falciparum* Malaria in 2007. *PLoS Medicine* **7**, e1000290.

Hay, S. I., Omumbo, J. A., Craig, M. H. and Snow, R. W. (2000). Earth observation, geographic information systems and *Plasmodium falciparum* malaria in sub-Saharan Africa. *Advances in Parasitology* **47**, 173–215.

Hengl, T., Rossiter, D. G. and Stein, A. (2003). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research* **41**, 1403–1422.

Hentschel, J., Lanjouw, J. O., Peter, L. and Poggi, J. (2000). Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador. *World Bank Economic Review* **14**, 147–165.

Heuvelink, G. B. M., Brus, D. and de Gruijter, J. J. (2006). Optimization of sample configurations for digital mapping of soil properties with universal kriging. In *Digital Soil Mapping: An Introductory Perspective* (eds. Lagacherie, P., McBratney, A. & Voltz, M.), pp. 1–17. Elsevier, Amsterdam.

Isham, V. (2010). Spatial point process models. In *Handbook of Spatial Statistics* (eds. Gelfand, A. E., Diggle, P. J., Fuentes, M. & Guttorp, P.), pp. 283–298. Chapman & Hall/CRC Press, Boca Raton, USA.

Jackson, C., Best, N. and Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in Medicine* **25**, 2136–2159.

Jackson, C., Best, N. and Richardson, S. (2008*a*). Hierarchical related regression for combining aggregate and individual level data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society (Series A)* **171**, 159–178.

Jackson, C., Richardson, S. and Best, N. (2008*b*). Studying space effects on health by synthesising individual and area-level outcomes. *Social Science and Medicine* **67**, 1995–2006.

Jung, I., Kulldorff, M. and Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine* **26**, 1594–1607.

Jung, I., Kulldorff, M. and Richard, O. J. (2010). A spatial scan statistic for multinomial data. *Statistics in Medicine* **29**, 1910–1918.

Kazembe, L. N., Appleton, C. C. and Kleinschmidt, I. (2007). Geographical disparities in core population coverage indicators for roll back malaria in Malawi. *International Journal of Equity in Health* **6**, 5.

Kim, H.-M., Mallick, B. K. and Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistics Association* **100**, 653–668.

Kleinman, K. P., Abrams, A. M., Kulldorff, M. and Platt, R. (2005). A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection* **133**, 409–419.

Kleinschmidt, I., Bagayoko, M., Clarke, G. P. Y., Craig, M. and Le Sueur, D. (2000). A spatial statistical approach to malaria mapping. *International Journal of Epidemiology* **29**, 355–361.

Kleinschmidt, I., Sharp, B., Mueller, I. and Vounatsou, P. (2002). Rise in malaria incidence rates in South Africa: A small-area spatial analysis of variation in time trends. *American Journal of Epidemiology* **155**, 257–264.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society (Series A)* **164**, 61–72.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. and Mostashari, F. (2005). A Space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine* **2**, e59.

Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine* **25**, 3929–3943.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine* **14**, 799–819.

Langford, I. H., Leyland, A. H., Rasbash, J. and Goldstein, H. (1999). Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistics Society C* **48**, 253–268.

Lasserre, V., Guihenneuc-Jouyauc, C. and Richardson, S. (2000). Biases in ecological studies: utility of including within-area distribution of confounders. *Statistics in Medicine* **19**, 45–59.

Lawson, A. B. (2006). *Statistical Methods in Spatial Epidemiology*. John Wiley, Chichester, UK.

Legendre, P. and Fortin, M.-J. (1989). Spatial pattern and ecological analysis. *Vegetation* **80**, 107–138.

Leonardo, L. R., Rivera, P., Saniel, O., Villacorte, E., Crisostomo, B., Hernandez, L., Baquilod, M., Erce, E., Martinez, R. and Velayudhan, R. (2008). Prevalence survey of schistosomiasis in Mindanao and the Visayas, The Philippines. *Parasitology International* **57**, 246–251.

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology* **73**, 1943–1967.

Levy, M. Z., Kawai, V., Bowman, N. M., Waller, L. A., Cabrera, L., Pinedo-Cancino, V. V., Seitz, A. E., Steurer, F. J., Cornejo del Carpio, J. G., Cordova-Benzaquen, E., Maguire, J. H., Gilman, R. H. and Bern, C. (2007). Targeted screening strategies to detect *Trypanasoma cruzi* infection in children. *PLoS Neglected Tropical Diseases* **1**, e103.

Levy, P. S. and Lemeshow, S. (2008). *Sampling of Populations: Methods and Applications*. Wiley-Blackwell, Oxford.

Mabaso, M. L., Vounatsou, P., Midzi, S., Da Silva, J. and Smith, T. (2006). Spatio-temporal analysis of the role of climate in inter-annual variation of malaria incidence in Zimbabwe. *International Journal of Health Geographics* **5**, 20.

Machault, V., Vignolles, C., Borchi, F., Vounatsou, P., Pages, F., Briolant, S., Lacaux, J. P. and Rogier, C. (2011). The use of remotely sensed environmental data in the study of malaria. *Geospatial Health* **5**, 151–168.

MacNab, Y. C. (2010). On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research* **20**, 49–68.

Manh, B. H., Clements, A. C. A., Thieu, N. Q., Hung, N. M., Hung, L. X., Hay, S. I., Hien, T. T., Wertheim, H. F. L., Snow, R. W. and Horby, P. (2010). Social and environmental determinants of malaria in space and time in Vietnam. *International Journal for Parasitology* **41**, 109–116.

Mathieu, E., Direny, A. N., de Rochars, M. B., Streit, T. G., Addis, D. G. and Lammie, P. J. (2006). Participation in three consecutive mass drug administrations in Leogane, Haiti. *Tropical Medicine and International Health* **11**, 862–868.

Melles, S. J., Heuvelink, G. B. M., Twenhofel, C. J. W., van Dijk, A., Hiemstra, P. H., Baume, O. and Stohlker, U. (2011). Optimizing the spatial pattern of networks for monitoring radioactive releases. *Computers and Geosciences* **37**, 280–288.

Mollie, A. (1996). Bayesian mapping of disease. In *Marok Chain Monte Carlo in Practice* (eds. Richardson, S. & Spiegelhalter, D.), Chapman & Hall/CRC Press, Boca Ranton, USA.

Morgenstern, H. (2008). Ecologic studies. In *Modern Epidemiology* (eds. Rothman, K. J., Greenland, S. & Lash, T. L.), pp. 511–531. Lippincott Williams & Wilkins, Philadelphia, PA.

Openshaw, S. and Taylor, P. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In *Statistical Applications in the Spatial Sciences* (ed. Wrigley, N.), Pion, London.

Patil, A. P., Gething, P. W., Piel, F. B. and Hay, S. I. (2011). Bayesian geostatistics in health cartography: the perspective of malaria. *Trends in Parasitology* **27**, 246–253.

Patil, A. P., Okiro, E. A., Gething, P. W., Guerra, C. A., Snow, R. W. and Hay, S. I. (2009). Defining the relationship between *Plasmodium falciparum* parasite rate and clinical disease: statistical models for disease burden estimation. *Malaria Journal* **8**, 186.

Peng, W. X., Tao, B., Clements, A. C., Jiang, Q. L., Zhang, Z. J., Zhou, Y. B. and Jiang, Q. W. (2010). Identifying high-risk areas of schistosomiasis and associated risk factors in the Poyang Lake region, China. *Parasitology* **137**, 1099–1107.

Pfeiffer, D. U. (1996). Issues related to handling of spatial data. In *New Zealand Veterinary Association/Australian Veterinary Association Second Pan Pacific Veterinary Conference* (ed. McKenzie, J.), Christchurch.

Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J. and Clements, A. C. A. (2008). *Spatial Analysis in Epidemiology*. Oxford University Press, New York.

Prentice, R. L. and Sheppard, L. (1995). Aggregate data studies of disease risk factors. *Biometrika* **82**, 113–125.

Pullan, R. L., Gething, P. W., Smith, J. L., Mwandawiro, C. S., Sturrock, H. J., Gitonga, C. W., Hay, S. I. and Brooker, S. (2011*a*). Spatial modelling of soil-transmitted helminth infections in Kenya: a disease control planning tool. *PLoS Neglected Tropical Diseases* **5**, e958.

Pullan, R. L., Kabatereine, N., Bukirwa, H., Staedke, S. G. and Brooker, S. (2011*b*). Heterogeneities, determinants and consequences of co-infection with *Plasmodium* malaria and hookworm: a spatial Bayesian analysis of a population study in Uganda. *Journal of Infectious Diseases* **203**, 406–417.

Pullan, R. L., Kabatereine, N., Quinnell, R. J. and Brooker, S. (2010). Spatial and genetic epidemiology of hookworm in a rural Ugandan community. *PLoS Neglected Tropical Diseases* **4**, e713.

Raso, G., Vounatsou, P., Gosoniu, L., Tanner, M., N'Goran, E. K. and Utzinger, J. (2006*a*). Risk factors and spatial patterns of hookworm infection among schoolchildren in a rural area of western Cote d'Ivoire. *International Journal for Parasitology* **36**, 201–210.

Raso, G., Vounatsou, P., Singer, B. H., N'Goran, E. K., Tanner, M. and Utzinger, J. (2006*b*). An integrated approach for risk profiling and spatial prediction of *Schistosoma mansoni*-hookworm coinfection. *Proceedings of the National Academy of Sciences, USA* **103**, 6934–6939.

Reich, B. J., Hodges, J. S. and Carlin, B. P. (2007). Spatial analysis of periodontal data using conditionally autoregressive priors having two classes of neighbor relations. *Journal of the American Statistics Association* **102**, 44–55.

Reid, H., Haque, U., Clements, A. C. A., Tatem, A. J., Vallely, A., Syed Masud, A., Islam, A. and Haque, R. (2010*a*). Mapping malaria risk

in Bangladesh using Bayesian geostatistical models. *American Journal of Tropical Medicine and Hygiene* **83**, 861–867.

Reid, H., Vallely, A., Taleo, G., Tatem, A., Kelly, G., Riley, I., Harris, I., Iata, H., Yama, S. and Clements, A. C. A. (2010*b*). Baseline spatial distribution of malaria prior to an elimination program in Vanuatu. *Malaria Journal* **9**, 150.

Richardson, S. and Best, N. (2003). Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics* **14**, 129–147.

Richardson, S. and Monfort, C. (2000). Chapter 11: Ecological correlation studies. In *Spatial Epidemiology* Oxford University Press, Oxford, UK.

Ripley, B. D. (1976). The second-order analysis of stationary point patterns. *Journal of Applied Probability* **13**, 255–266.

Robertson, C., Nelson, T. A., MacNab, Y. C. and Lawson, A. B. (2010*a*). Review of methods for space-time disease surveillance. *Spatial and Spatio-temporal Epidemiology* **1**, 105–116.

Robertson, C., Sawford, K., Daniel, S. L., Nelson, T. A. and Stephen, C. (2010*b*). Mobile phone-based infectious disease surveillance system, Sri Lanka. *Emerging Infectious Diseases* **16**, 1524–1531.

Roll Back Malaria Monitoring and Evaluation Reference Group (2005). *Malaria indicator survey: basic documentation for survey design and implementation*. World Health Organization, Geneva.

Rothman, K. J. (1990). A sobering start for the cluster busters' conference. *American Journal of Epidemiology* **132**, S6–13.

Ryan, J. R., Mbui, J., Rashid, J. R., Wasunna, M. K., Kirigi, G., Marigi, C., Kinoti, D., Ngumbi, P. M., Martin, S. K., Odera, S. O., Hochberg, L. P., Bautista, C. T. and Chan, A. S. T. (2006). Spatial clustering and epidemiological aspects of visceral leishmaniasis in two endemic villages, Baringo District, Kenya. *American Journal of Tropical Medicine and Hygiene* **74**, 308–317.

Schriefer, A., Guimarães, L. H., Machado, P. R. L., Lessa, M., Lessa, H. A., Lago, E., Ritt, G., Góes-Neto, A., Schriefer, A. L. F., Riley, L. W. and Carvalho, E. M. (2009). Geographical clustering of leishmaniasis in Northeastern Brazil. *Emerging Infectious Diseases* **15**, 871–876.

Schur, N., Hürlimann, E., Garba, A., Traoré, M. S., Ndir, O., Ratard, R. C., Tchuem Tchuenté, L. A., Kristensen, T. K., Utzinger, J. and Vounatsou, P. (2011*a*). Geostatistical model-based estimates of Schistosomiasis prevalence among individuals aged $\leqslant$20 years in West Africa. *PLoS Neglected Tropical Diseases* **5**, e1194.

Schur, N., Hurlimann, E., Stensgaard, A. S., Chimfwembe, K., Mushinge, G., Simoonga, C., Kabatereine, N., Kristensen, T. K., Utzinger, J. and Vounatsou, P. (2011*b*). Spatially explicit *Schistsoma* infection risk in eastern Africa using Bayesian geostatistical modelling. *Acta Tropica* (in press). doi:10.1016/jactatropica.2011.10.006

Schur, N., Utzinger, J. and Vounatsou, P. (2011*c*). Modelling age-heterogenous *Schistsoma haematobium* and *S. mansoni* survey data via alignment factors. *Parasites and Vectors* **4**, 142.

Simoonga, C., Utzinger, J., Brooker, S., Vounatsou, P., Appleton, C. C., Stensgaard, A. S., Olsen, A. and Kristensen, T. K. (2009). Remote sensing, geographical information system and spatial analysis for schistosomiasis epidemiology and ecology in Africa. *Parasitology* **136**, 1683–1693.

Soares Magalhães, R. J., Barnett, A. G. and Clements, A. C. (2011*a*). Geographic analysis of the role of water supply and sanitation in the risk of helminth infections of children in West Africa. *Proceedings of the National Academy of Sciences, USA* **108**, 20084–20089.

Soares Magalhães, R. J., Biritwum, N. K., Gyapong, J. O., Brooker, S., Zhang, Y., Blair, L., Fenwick, A. and Clements, A. C. (2011*b*). Mapping helminth co-infection and co-intensity: geostatistical prediction in Ghana. *PLoS Neglected Tropical Diseases* **5**, e1200.

Soares Magalhães, R. J., Clements, A. C., Patil, A. P., Gething, P. W. and Brooker, S. (2011*c*). The applications of model-based geostatistics in helminth epidemiology and control. *Advances in Parasitology* **74**, 267–296.

Srividya, A., Michael, E., Palaniyandi, M., Pani, S. and Das, P. K. (2002). A geostatistical analysis of the geographical distribution of lymphatic filariasis prevalence in southern India. *American Journal of Tropical Medicine and Hygiene* **67**, 480–489.

Steel, D. G. and Holt, D. (1996). Analysing and adjusting aggregation effects: The ecological fallacy revisited. *International Statistics Review* **64**, 39–60.

Stein, A. and Ettema, C. (2003). An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. *Agriculture, Ecosystems and Environment* **94**, 31–47.

Stein, M. L. (2005). Space-time covariance functions. *Journal of the American Statistical Association* **10**, 310–321.

Stensgaard, A. S., Vounatsou, P., Onapa, A. W., Simonsen, P. E., Pedersen, E. M., Rahbek, C. and Kristensen, T. K. (2011). Bayesian geostatistical modelling of malaria and lymphatic filariasis infections in Uganda: predictors of risk and geographical patterns of co-endemicity. *Malaria Journal* **10**, 298.

Stone, R. A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine* **7**, 649–660.

Sturrock, H. J. W., Gething, P. W., Ashton, R., Kolaczinski, J. H., Kabatereine, N. B. and Brooker, S. (2011). Planning schistosomiasis control: investigation of alternative sampling strategies for *Schistosoma mansoni* to target mass drug administration of praziquantel in East Africa. *International Health* **3**, 165–175.

Sturrock, H. J. W., Gething, P. W., Clements, A. C. A. and Brooker, S. (2010). Optimal survey designs for targeting chemotherapy against soil-transmitted helminths: effect of spatial heterogeneity and cost-efficiency of sampling. *American Journal of Tropical Medicine and Hygiene* **82**, 1079–1087.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* **4**, 11.

Tarafder, M. R., Carabin, H., Joseph, L., Balolong, E. J., Olveda, R. and McGarvey, S. T. (2010). Estimating the sensitivity and specificity of Kato-Katz stool examination technique for detection of hookworms, *Ascaris lumbricoides* and *Trichuris trichiura* infections in humans in the absence of a 'gold standard'. *International Journal for Parasitology* **40**, 399–404.

Tatem, A. J., Qiu, Y., Smith, D. L., Sabot, O., Ali, A. S. and Moonen, B. (2009). The use of mobile phone data for the estimation of the travel patterns and imported *Plasmodium falciparum* rates among Zanzibar residents. *Malaria Journal* **8**, 287.

Teixeria, T. R. and Cruz, O. G. (2011). Spatial modeling of dengue and socio-environmental indicators in the city of Rio de Janeiro, Brazil. *Cadernos de Saúde Pública* **37**, 591–602.

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

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46**, 234–240.

Utzinger, J., Booth, M., N'Goran, E. K., Muller, I., Tanner, M. and Lengeler, C. (2001). Relative contribution of day-to-day and intra-specimen variation in faecal egg counts of *Schistosoma mansoni* before and after treatment with praziquantel. *Parasitology* **122**, 537–544.

Van Groenigen, J. W., Gandah, M. and Bouma, J. (2000). Soil Sampling Strategies for Precision Agriculture Research under Sahelian Conditions. *Soil Science Society of America Journal* **64**, 1674–1680.

Van Groenigen, J. W., Siderius, W. and Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma* **87**, 239–259.

Vanamail, P., Ramaiah, K. D., Subramanian, S., Pani, S. P., Yuvaraj, J. and Das, P. K. (2005). Patterns of community compliance with spaced, single-dose, mass administrations of diethylcarbamazine or ivermectin, for the elimination of lymphatic filariasis from rural areas of southern India. *Annals of Tropical Medicine and Parasitology* **99**, 237–242.

Vounatsou, P., Raso, G., Tanner, M., N'Goran, E. K. and Utzinger, J. (2009). Bayesian geostatistical modelling for mapping schistosomiasis transmission. *Parasitology* **136**, 1695–705.

Wakefield, J. (2004). Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistics Society (Series A)* **167**, 385–445.

Wakefield, J. and Lyons, H. (2010). Spatial aggregation and the ecological fallacy. In *Handbook of Spatial Statistics* (eds. Gelfand, A. E., Diggle, P. J., Fuentes, M. & Guttorp, P.), Chapman & Hall/CRC Press, Boca Raton, USA.

Wakefield, J. and Salway, R. (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistics Society (Series A)* **164**, 119–137.

Waller, L. (2010). Point process models and methods in spatial epidemiology. In *Handbook of Spatial Statistics* (eds. Gelfand, A. E., Diggle, P. J., Fuentes, M. & Guttorp, P.), pp. 403–423. Chapman & Hall/CRC Press, Boca Raton, USA.

Waller, L. and Carlin, B. P. (2010). Disease Mapping. In *Handbook of Spatial Statistics* (eds. Gelfand, A. E., Diggle, P. J., Fuentes, M. & Guttorp, P.), pp. 217–243. Chapman & Hall/CRC Press, Boca Raton, USA.

Waller, L. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Hoboken, USA.

Wang, X. H., Zhou, X. N., Vounatsou, P., Chen, Z., Utzinger, J., Yang, K., Steinmann, P. and Wu, X. H. (2008). Bayesian spatio-temporal modeling of *Schistosoma japonicum* prevalence data in the absence of a diagnostic 'gold' standard. *PLoS Neglected Tropical Diseases* **2**, e250.

Wanjala, C. L., Waitumbi, J., Zhou, G. and Githeko, A. K. (2011). Identification of malaria transmission and epidemic hotspots in the western Kenya highlands: its application to malaria epidemic prediction. *Parasites and Vectors* **4**, 81.

Wardrop, N. A., Atkinson, P. M., Gething, P. W., Fevre, E. M., Picozzi, K., Kakermo, A. S. L. and Welburn, S. C. (2010). Bayesian geostatistical analysis and prediction of Rhodesian Human African Trypanosomiasis. *PLoS Neglected Tropical Diseases* **4**, e914.

Washington, C. H., Radday, J., Streit, T. G., Boyd, H. A., Beach, M. J., Addiss, D. G., Lovince, R., Lovegrove, M. C., Lafontant, J. G., Lammie, P. J. and Hightower, A. W. (2004). Spatial clustering of filarial transmission before and after a Mass Drug Administration in a setting of low infection prevalence. *Filaria Journal* **3**, 3.

Weins, J. A. (1989). Spatial scaling in ecology. *Functional Ecology* **3**, 385–397.

World Health Organization (2006). *Preventive chemotherapy in human helminthiasis. Coordinated use of anthelmintihic drugs in control interventions: a manual for health professionals and programme managers*. World Health Organization, Geneva.

Yiannakoulias, N., Svenson, L. W. and Schopflocher, D. P. (2009). An integrated framework for the geographic surveillance of chronic disease. *International Journal of Health Geographics* **8**, 69.

Zouré, H. G. M., Wanji, S., Noma, M., Amazigo, U. V., Diggle, P. J., Tekle, A. H. and Remme, J. H. (2011). The Geographic Distribution of *Loa loa* in Africa: Results of Large-Scale Implementation of the Rapid Assessment Procedure for Loiasis (RAPLOA). *PLoS Neglected Tropical Diseases* **5**, e1210.