Running title: Spatial statistics for NTD control

Title: The role of spatial statistics in the control and elimination of neglected tropical diseases in sub-Saharan Africa: A focus on human African trypanosomiasis, schistosomiasis and lymphatic filariasis

Michelle C Stanton

Department of Parasitology, Liverpool School of Tropical Medicine, UK

**Abstract**

Disease control and elimination programmes can benefit greatly from accurate information on the spatial variability of disease risk, particularly when risk is highly spatially heterogeneous. Due to advances in statistical methodology, coupled with the increased availability of geospatial technology, this information is becoming increasingly accessible. In this chapter we describe recent advancements in spatial methods associated with the analysis of disease data measured at the point-level, and demonstrate their application to the control and elimination of neglected tropical diseases (NTDs). We further provide information on spatially referenced data sources and software that can be used to create these risk maps, concentrating on those that can be freely obtained. Examples relating to three NTDs affecting populations in sub-Saharan Africa are presented throughout the chapter i.e. human African trypanosomiasis, schistosomiasis and lymphatic filariasis. These three diseases, with differing routes of transmission, control methods and level of spatial heterogeneity, demonstrate the flexibility and applicability of the methods described.

**Keywords**

Neglected tropical diseases, spatial statistics, geostatistics, Maxent, lymphatic filariasis, schistosomiasis, human African trypanosomiasis, species distribution models, GIS, disease mapping

1. **Introduction**

In recent years we have seen great progress being made towards the improvement of human health across the developing world, with notable examples in sub-Saharan Africa including the halving of the number of clinical cases of malaria attributed to *Plasmodium falciparum* (Bhatt et al. 2015), and the near-eradication of Guinea worm. Neglected tropical diseases (NTDs) in particular have received a substantial amount of attention, with the World Health Organization targeting 8 of the 17 NTDs for elimination, in addition to the eradication of Guinea worm i.e. onchocerciasis (in the Americas), lymphatic filariasis (LF), trachoma, leprosy, Chagas (in the Americas), visceral leishmaniasis (in Southeast Asia), Yaws (in Southeast Asia) and human African trypanosomiasis (HAT), whereas others such as schistosomiasis and soil-transmitted helminths are targeted for control (World Health Organization 2010). Tools for control and elimination recommended by WHO include preventive chemotherapy, intensified case management, vector control, the provision of safe water, sanitation and hygiene (Hotez et al. 2010; World Health Organization 2012; Freeman et al. 2013; Bockarie et al. 2013; Golding et al. 2015; World Health Organization 2015). With many of these strategies in place, the geographical disease landscape is changing, and will continue to change as the prevalence of these diseases declines. This decline leads to many operational challenges when designing and implementing control and elimination interventions, as there is an inverse relationship between a diseases level of endemicity and the efforts required to locate those with the disease. As a consequence, the strategies used to control diseases that are currently relatively commonplace will need to be adapted to ensure that they are feasible one the disease becomes rare, else resurgence may occur (Bergquist et al. 2015; Bockarie et al. 2013; Hawkins et al. 2016). To achieve this, a deeper understanding of where a disease is likely to persist is a very valuable resource. This chapter therefore aims to present spatial statistical methodology and spatial resources currently available that may assist in understanding the spatial distribution of NTDs. To demonstrate these advances, their applications to three NTDs found in sub-Saharan Africa are presented: HAT, schistosomiasis and LF (Lutumba et al. 2016; Danso-Appiah 2016; Sodahlon et al. 2016). These three diseases vary in spatial heterogeneity (HAT is most heterogeneous, LF is the least), transmission routes, and approaches to control/elimination and as such serve to prove the diversity of the methods and resources being described in this chapter (see Table 1). In Section 2 we provide an overview of these three NTDs, and intersperse each subsequent section with examples of the statistical methods and resources being applied to these NTDs to further our understanding of their distributions, and potentially guide intervention strategies for controlling and eliminating them. Section 3 focuses on statistical methods that are commonly applied to disease risk mapping, and Section 4 summarises common issues that are encountered when undertaking spatial analyses. Finally, section 5 introduces a range of spatial data resources that are available to aid in mapping the risk of the three NTDs of interest.

1. **Overview of NTDs**

NTDs are a diverse group of diseases, with their commonality being that they have traditionally been a low priority in those countries affected by them, despite their large, but potentially hidden, adverse impact on the population. Key features of the three geographically-overlapping selected diseases are described below, including how they are transmitted, including the factors influencing their spatial distribution, and strategies used for their control (Table 1). References which explore the spatial distribution of these diseases are included in this section, with more detail on the methods used being presented in Section 3.

* 1. **Human African Trypanosomiasis**

HAT, also known as sleeping sickness, is an infection caused by protozoan parasites belonging to the Trypanosoma species. It is a vector-borne disease, with the vector being the tsetse fly i.e. a fly of the *Glossina* genus. There are two forms of HAT depending on the parasite involved i.e. *Trypanosoma brucei gambiense,* also known as Gambian HAT*,* which mainly occurs in West and Central Africa and accounts for the majority of HAT cases and *Trypanosoma brucei rhodesiense*, also known as Rhodesian HAT, which mainly occurs in East and Southern Africa. There are two stages of the disease. Symptoms of stage one are relatively mild and include fever, headaches and joint pain, whereas stage two is much more obvious, causing neurological conditions such as convulsions, confusion, sleep disturbance and eventual death if untreated. *T.b. gambiense* infections result in the chronic form of the disease i.e. a long first stage, whereas those infected with *T.b. rhodesiense* experience neurological signs and symptoms more rapidly (World Health Organization 2013a; Franco et al. 2014).

* + 1. *Geographical influences of transmission*

HAT is the most spatially heterogeneous (focal) of the three diseases being considered, and currently is the closest to elimination with 3,796 cases reported in 2014 (World Health Organization 2016). Approximately 360 HAT foci have previously been identified (300 for Gambian HAT, 60 for Rhodesian HAT), which are generally found in rural and remote areas. Tsetse flies are considered to be very sensitive to environmental conditions, hence even with these defined foci there is heterogeneity in disease risk (Fèvre et al. 2006; World Health Organization 2013a). When determining the geographical distribution of disease prevalence, or more importantly the geographical distribution of transmission, the environmental factors identified are geographically variable, depending on the *Glossina* subgenus and the form of disease under consideration. For example, Gambian HAT is primarily transmitted by riverine species, hence transmission risk is strongly associated with the presence of rivers and the surrounding riparian vegetation (Rogers 2000; Franco et al. 2014). Rhodesian HAT is primarily transmitted by savannah species hence is more associated with forests. Further, as the Rhodesian form is a zoonosis for which animals are the main reservoir, the presence of wild fauna and cattle also influence the distribution of risk (Batchelor et al. 2009; Wardrop et al. 2010). Factors also influencing a person’s likelihood of coming into contact with infected tsetse also need to be considered.

Spatial models relating to the transmission of HAT either focus on the distribution of the vector or HAT cases. There are only a small number of papers in the published literature on Gambian HAT cases that incorporate spatial analysis, with the majority of these being produced by the Atlas of HAT team (Simarro et al. 2011; Simarro et al. 2015; Lumbala et al. 2015). To our knowledge there are no papers that explore the link between the environment and Gambian HAT cases. For Rhodesian HAT, there are several papers that explore the relationship between cases and variables such as vegetation, temperature, elevation and the proximity of livestock to produce predictive maps of prevalence (Batchelor et al. 2009; Wardrop et al. 2010). More prolific in the literature are studies that explore the spatial distribution of the tsetse fly vector. Early examples of this are summarised by Rogers (2000), with more recent work that incorporates risk factors such as elevation, vegetation, temperature rainfall and land cover at a higher spatial resolution includes (Sciarretta et al. 2005; Sciarretta et al. 2010; Matawa et al. 2013; Dicko et al. 2014).

* + 1. *Control and elimination strategy*

Several approaches have been adopted to control HAT, although a universal method has yet to be adopted. Option for control include

* Active case detection and subsequent treatment
* Passive case detection and subsequent treatment
* Vector control
* Control of animal reservoirs

or combinations thereof. In areas where humans are the main reservoir, the primary method of control is the detection and treatment of infected individuals. Active surveillance firstly involves screening the population by checking for clinical sign and performing serological tests (available for the Gambian form only). Diagnosis is then confirmed by establishing whether parasites are present in body fluids. If this is the case, a lumbar puncture is performed to determine the stage of the disease i.e. to determine if the parasite has passed into the central nervous system, the results of which determine which treatment needs to be administered. Current treatment options are however considered to be toxic and complicated to administer (Kennedy 2013). This active approach has been shown to be effective in some Gambian HAT areas, but in other areas where disease transmission continues to persist, vector control is also needed using approaches such as artificial bait methods (traps and targets), ground and aerial spraying, plus sterile insect techniques (World Health Organization 2013b). It has been shown that reductions in the tsetse population of 72% or more can be enough to prevent transmission in areas where riverine tsetse are the main vector (Gouteux & Artzrouni 1996; Solano et al. 2013; Tirados et al. 2015). In Rhodesian HAT areas, animal reservoirs need to be considered. Where domestic livestock is the primary reservoir, treatment with chemotherapy or livestock targeted vector control can be considered, whereas if wild animals are the primary reservoir, vector control and area avoidance are the only viable options available.

* 1. **Schistosomiasis**

In sub-Saharan Africa the two forms of schistosomiasis, urogenital schistosomiasis and intestinal schistosomiasis are caused by *Schistosoma haematobium* and *Schistosoma mansoni* respectively. It is a waterborne disease such that infection occurs when people come into contact with freshwater inhabited by snails (*Bulinus* and *Biomphalaria* spp.) carrying parasitic blood fluke (Fenwick et al. 2006; Colley et al. 2014). More specifically, people are infected by the larval forms of parasite which swim in freshwater after emerging from the snail carrier and penetrate the individual’s skin. Symptoms of schistosomiasis are dependent on the form that the individual infected with but include blood in urine, plus kidney damage, fibrosis of the bladder, bladder cancer and infertility in advanced cases (urogenital) or blood in stool and stomach pain, plus enlarged liver and spleen in advanced cases (intestinal form). Further, prolonged infection may result in anaemia, malnutrition, stunted growth and cognitive impairment anaemia (Colley et al. 2014; Danso-Appiah 2016).

* + 1. *Geographical influences of transmission*

Forty sub-Saharan African countries are considered endemic for schistosomiasis, accounting for 93% of the disease burden with an estimated 192 million cases out of 207 million globally (Steinmann et al. 2006; Adenowo et al. 2015). Transmission of schistosomiasis occurs in areas where people have contact with waterbodies which are a good habitat for the snail host through activities such as bathing, swimming, fishing, and domestic tasks such as cooking, resulting in a very spatially heterogeneous disease landscape.Unlike HAT, in sub-Saharan Africa disease prevalence is still very high, although it is much closer to elimination in other regions with notable success at interrupting transmission in countries such as China and Brazil (World Health Organization 2013e). Environmental factors researchers consider when determining the risk of disease transmission therefore include those associated with snail habitat suitability e.g. precipitation, temperature, various soil properties, vegetation indices, land cover type, and water properties, plus those associated with human water contact e.g. water accessibility(Manyangadze et al. 2015; Lai et al. 2015)**.**

A review of spatial application for schistosomiasis published in 2015 (Manyangadze et al. 2015) identified 36 papers between the period 2009-2013 which focus on the spatial distribution of disease transmission risk in Africa, and supplements a review published in 2009 which focused on the period 1996-2008 (Simoonga et al. 2009). In contrast to HAT, the majority of these studies focus on the distribution of cases as opposed to the intermediate snail host. Of these, prevalence of disease as opposed to infection of intensity is of greater interested, likely due to the control strategy for the disease, described below.

* + 1. *Control and elimination strategy*

Whereas the other two diseases under consideration are targeted for elimination, the goal set of the schistosomiasis by the WHO is to ensure at least 75% of school-aged children living in at-risk areas have access to preventive chemotherapy with praziquantel. Currently, praziquantel is delivered either annually or biennially to school-aged in areas considered to be at high or moderate risk respectively. Risk is assessed using prevalence surveys such that high risk areas are those with an estimated schistosomiasis prevalence (using parasitological diagnostic methods) above 50%, whereas moderate areas are those with prevalence between 10% and 50% (World Health Organization 2006)**.** In recognising the diseases spatially heterogeneous distribution due to its dependence on freshwater sources, the WHO guidelines state “you need to survey specifically some areas that are near lakes, ponds, streams or irrigated areas….select a few schools close to the water and some a little further away”. On a practical level however, praziquantel is administered at the implementation unit level, usually a district, hence strategies are commonly based on the average prevalence for the implementation unit using a sample of 50 children per randomly select school (World Health Organization 2013d; Sousa-Figueiredo et al. 2015). Due to the known spatial heterogeneity of the disease, it has been suggested that this strategy will not enable elimination to be achieved and more targeted approaches are required. Suggestions of suitable strategies include the use of lot quality assurance sampling (LQAS), sampling based on a variogram analysis to measure the degree of spatial heterogeneity in the data, and a strategy based on increasing the ‘mapping resolution’ i.e. the proportion of schools sampled (Sturrock et al. 2011; Sousa-Figueiredo et al. 2015). Snail control, although not commonly used, has also been identified as an effective historical method of reducing disease transmission that should be reconsidered by control programmes (Sokolow et al. 2016).

* 1. **Lymphatic filariasis**

LF is a vectorborne disease caused by the parasitic worms *Wuchereria bancrofti* (90% of cases), *Brugia malayi* and *Brugia timori* that is transmitted by mosquitoes. In sub-Saharan Africa, the disease is primarily transmitted by *Anopheles* mosquitoes in rural areas and *Culex* mosquitoes in urban and peri-urban areas. Whilst a large proportion of infected people are asymptomatic, LF has a wide range of debilitating clinical manifestations the most common of which are lymphoedema, hydrocoele and episodic adenolymphangitis (ADLA) (Hotez & Kamath 2009; Nutman 2013).

* + 1. *Geographical influences of transmission*

In sub-Saharan Africa 36 countries are considered endemic for lymphatic filariasis, with an estimated 40 million cases accounting for 65% of the global burden in 2013 (Ramaiah & Ottesen 2014). Transmission patterns in disease are largely associated with factors relating to the distribution of the vector, plus factors associated with human contact with mosquitoes. As *Anopheles* mosquitoes are also a major vector for malaria, a large amount of research has been undertaken to determine the risk factors associated with its distribution i.e. altitude, rainfall, temperature, humidity, land cover type and vegetation indices (Slater & Michael 2013; Cano et al. 2014; Moraga et al. 2015). With regards to human-mosquito contact, as *Anopheles* are mostly active between dusk and dawn, contact is determined by the use of personal protection methods such as bed nets, and the proximity of the home to suitable *Anopheles* habitat. Less research has been undertaken into the factors affecting LF transmission in urban environments, with linkages being primarily identified with the locations of environmental infrastructure e.g. proximity to sanitation facilities, sewerage and drainage, as *Culex* breeding sites tend to be situated in stagnant water collections (Simonsen & Mwakitalu 2013).

As with schistosomiasis, the majority of papers which apply spatial methods to lymphatic filariasis data focus on lymphatic filariasis prevalence, with prevalence obtained using immunochromatographic test (ICT) cards being most readily available. In a recent paper, Moraga et al. (2015) collated all geographically referenced microfilariae (mf) and ICT prevalence data for sub-Saharan Africa prior to the commencement of MDA to produce regional maps of both prevalence measures. Similar prevalence maps based on historical or baseline survey data can be found at the regional national or sub-national level for many sub-Saharan African countries (Gyapong et al. 2002; Slater & Michael 2012; Slater & Michael 2013; Stanton et al. 2013; Cano et al. 2014; Mwase et al. 2014).

As vector control for LF becomes a greater focus (World Health Organization 2011; van den Berg et al. 2013), knowledge of the spatial distribution of LF vectors is also of importance. As the *Anopheles* mosquitoes that transmit LF are also a vector for malaria, there is a vast amount of information available on their spatial distribution, e.g. the Malaria Atlas Project produces predictive species distribution maps produced using mosquito occurrence data (De’ath 2007; Sinka, Bangs, Manguin, Coetzee, Arbaji, et al. 2010). Predictive maps of *Culex* distribution have also been derived by the VectorMap project (www.vectormap.org) (Moraga et al. 2015).

* + 1. *Control and elimination strategy*

Similar to schistosomiasis, the primary elimination for LF is annual preventive chemotherapy using ivermectin plus albendazole if onchocerciasis is co-endemic, or DEC plus albendazole otherwise (World Health Organization & Global Programme to Eliminate Lymphatic Filariasis 2011). Implementation occurs at the implementation unit level which is primarily district level. The average prevalence of adults (aged 15+) at the implementation unit level is used to determine whether to deliver preventive chemotherapy, using a decision threshold of 1% based on rapid diagnostic tests i.e. ICTs for *Wuchereria bancrofti* antigen. Average prevalence is determined using the results of 50-100 people in two villages that are considered most likely to have ongoing transmission. No guidance is provided on how to identify these villages however. Initially suggested baseline mapping strategies included Rapid Assessment of the Geographical Distribution of Filariasis (RAGFIL), which advocated applying a grid-based sampling scheme to ensure sampled villages were approximately 50km apart, and using the variogram (see Section 3.1.2) of the resulting prevalence estimates to estimate the spatial correlation parameters and subsequently produce the smoothed risk surface (Gyapong & Remme 2001; Gyapong et al. 2002). This approach was not adopted by WHO however (World Health Organization 2000; World Health Organization & Global Programme to Eliminate Lymphatic Filariasis 2011). In addition to preventive chemotherapy, vector control and morbidity management are also advocated (World Health Organization 2011; World Health Organization 2013c).

1. **Statistical methods for disease risk mapping**

Disease risk maps come in a variety of forms depending on their purpose, and crucially the availability of data. In its simplest form, disease maps are visual representations of disease risk (incidence, prevalence, intensity, vector abundance etc.) in geographical space, either at the area-level e.g. using spatially discrete district boundaries or point-level across an area of interest. Historically, area-level maps of disease risk are more common, however due to advances in geographical and computational technology, geographically referenced data is now much more attainable. It is almost the natural progression, once given a visual representation of the disease landscape, to look for patterns in its spatial distribution and subsequently to begin to hypothesise as to what influences of these patterns. An early and well known instance of this took place during a cholera epidemic in Soho, London in 1854. During this period, the now renowned Dr John Snow used a map of cholera deaths to identify the source of the epidemic, namely the Broad Street waterpump (Warner 1996; McLeod 2000). This led to John Snow confirming his hypothesis that cholera was transmitted via polluted drinking water as opposed to miasmata (bad air), which was the dominant theory at the time. Spatial statistical methods therefore exist so that we can go beyond hypothesising whether spatial patterns in diseases exist, and further what causes these spatial patterns ‘by eye’, and allow us to draw more objective conclusions. This area of statistics is vast and continues to grow so in this paper we focus on spatial statistical concepts and approaches involving point-level data, i.e. data observed at a specific point in geographical space, as opposed to area-level spatial data or point process data (Cressie 1993b). In particular we focus on methods which we find commonly applied to tropical disease data, including the three NTDs of interest. The methods described in this section have been separated into two categories, namely methods of assessing for evidence of spatial dependency and methods for mapping spatial variability.

* 1. **Assessing for evidence of spatial dependency**

Spatial dependency is the backbone of many spatial statistical models. The concept of spatial dependency, sometimes termed spatial autocorrelation, is that events that occur close together in geographical space are more similar than events that occur further apart i.e. the first law of geography (Tobler 1970). In a tropical disease context, spatial dependency is likely to be present as nearby locations share the environmental or socio-economic conditions that are favourable for disease transmission e.g. the environment is suitable habitat for the vector, or the lack of appropriate water or sanitation facilities result in human behaviour that increases the risk of transmission. By ignoring the presence of spatial dependency when determining a suitable statistical model for the data, there is a risk that the relationship between the disease and potential risk factors under investigation may be misrepresented, or missed altogether (Legendre 1993; Thomson et al. 1999; F. Dormann et al. 2007).

We describe two common methods for assessing spatial dependency found in the tropical disease literature: Moran’s I and the variogram (Moran 1950; Cressie 1993a). Whilst both serve to identify whether there is a spatial pattern to geographically-referenced data, it is only the latter provides information on the data’s spatial correlation structure. We assume that we are interested in the observed data which are realisations of the spatially continuous process measured at locations in geographical space. In practice, it is common to assess whether there is any evidence of spatial dependency once any known trends have been removed e.g. assess the residuals obtained after fitting a generalised linear model, but for explanatory purposes we shall assume that is the process of interest.

* + 1. *Moran’s I*

Moran’s I provides an overall assessment of whether values associated with specific geographical locations are clustered, random or dispersed (Moran 1950). The (Global) Moran’s I statistic is calculated as follows:

\begin{equation*}
I = \frac{N}{\sum_{i}\sum_{j}w_{ij}}\frac{\sum_{i}\sum_{j}(Y_{i}-\bar{Y})(Y_{j}-\bar{Y})}{\sum_{i}(Y_{i}-\bar{Y})^{2}}
\end{equation*}

where $w_{ij}$ is the spatial weight between the two points measured at location and location , $N$is the total number of observed locations and $\bar{Y}$ is the mean of $Y$. Spatial weight matrices can take many forms, with commonly $w_{ij}=1$ if points and are neighbours, with examples of neighbours of including locations within a fixed distance, or points that when ranked by distance are at most the kth point furthest away from i (K nearest neighbours). As the mean and the variance under the assumption of no spatial dependency are known, the index can be standardised and the resulting Z score can be compared to the Gaussian distribution. The statistic $I$ lies between -1 and 1, and significant positive values indicate spatial clustering, whereas significant negative values indicate dispersion.

* + 1. *The variogram*

The variogram is an important tool for exploring whether there is evidence of spatial dependency in geographically referenced data, and obtaining preliminary estimates of the parameters that represent the spatial correlation structure of the data i.e. the covariance parameters. The theoretical variogram or semi-variance of $Y(x)$ can be written as

\begin{equation*}
V(x_{i},x_{j}) = \frac{1}{2}\mbox{Var}\{Y_{i}-Y_{j}\}
\end{equation*}

is sometimes referred to as the semi-variance. To define this further, we initially assume that the spatially continuous data $Y(x)$ has a normal distribution, and is simply a combination of the true underlying outcome $S(x)$, plus some random variability/measurement error/random $Z(x)$. It is commonly assumed that $S(x)$ is a stationary, isotropic Gaussian process with zero mean and covariance defined as $\mbox{Cov}\{S(x_{i}),S(x_{j})\} = \sigma^2\rho(u)$ where $u$ is the Euclidean (straight-line) distance between $x_{i}$ and $x_{j}$. The correlation function $\rho(\cdot)$ is represented by the Matérn family of spatial correlation functions (Matern 1960; Guttorp & Gneiting 2006), which have two parameters $\kappa$ and $\phi$. The $\kappa$ parameter represents the smoothness of the function, whereas $\phi$ represents the rate at which the correlation between two points decreases with increasing distance. For example, the exponential correlation function $\rho(u) = \mbox{exp}\{-u/\phi\}$ is a special case of the Matérn family, with $\kappa = 0.5$. As $\mbox{Cov}\{S(x_{i}),S(x_{j})\}$ depends only on the separation distance $u$ between two points as opposed to the locations of the points themselves, we say that the process is stationary. Further as the separation distance is independent of direction, we say that the process is isotropic. We assume that $Z(x)$ are identically distributed, independent Gaussian random variables with mean 0, variance $\tau^2$. In the spatial statistics literature, this is referred to as the nugget effect (Diggle & Ribeiro 2007). Figure 1 is a theoretical representation of the variogram using parameters $\sigma^2=0.8$, $\tau^2=0.2$, $\kappa = 0.5$ and $\phi=0.4$. Note that the practical range represents the value of $u$ when the semi-variance is equal to 95% of the sill i.e. $\tau^2+\sigma^2$. We note that in the presence of spatial correlation, the values of the variogram are small for points close together, and increase as the distance between points increases indicating that close points are more similar (less variable) than distant points. At a sufficiently large separation distance there is negligible correlation between the points, and the variogram line levels off, and tends asymptotically to the value of the sill. The value of $v(u)$ at close or coincident points (small $u$) represents the variance of the random noise/measurement error, $\tau^2$. This is sometimes referred to as the nugget effect. The steepness of the initial slope (i.e. the rate at which the spatial correlation decays) is determined by $\phi$ and $\kappa$.



[Insert Figure 1 here]

In practice, in order to make an initial assessment of whether spatial dependency is present, plus obtain initial estimates of the covariance parameters $\sigma^2$ and $\phi$ plus the nugget effect $\tau^2$, the empirical variogram is plotted. Point pairs are grouped by distance bands $h(u) = (u-d,u+d)$, and $h(u)$ is plotted against the average of the empirical variogram ordinates $v_{ij} = \frac{1}{2}(y_{i}-y_{j})^2$, where $y_{i}$ and $y_{j}$ are separated by a distance of $u \pm d$ . The shape of the empirical variogram gives some indication of whether there is spatial dependency in the data, plus estimates of the model parameters can be made either subjectively e.g. fitting a line to the resulting points by eye (see eyefit() in the R package geoR, or using curving-fitting techniques such as least-squares (see variofit() in geoR).This approach should be implemented cautiously however as the fitted variogram is dependent on the width of the distance bands selected, and multiple combinations of parameter estimates may result in equally well-fitting curves (Cressie 1985; Diggle & Ribeiro 2007). Whilst curve-fitting approaches are generally adopted within geographical information systems (GIS) software (e.g. within the Geostatistical Analyst tools in ArcGIS), within the statistical community these parameter estimates are more commonly used as initial values in optimisation algorithms when fitting spatial regression models as outlined below (Section 3.2.2). Monte Carlo approaches are used to assess whether the empirical variogram provides evidence of spatial correlation in the data, or whether the data from which it has been produced are spatially unstructured (Baddeley et al. 2014). In this context, the data are randomly allocated to the observation locations a large number of times (usually at least 99). For each simulation, the empirical variogram is computed and the maximum and minimum value for each value of the distance band are plotted to form a simulation envelope. This envelope represents the range of empirical variogram values that could feasible be obtained under the assumption that the data are spatially independent. Hence, if the observed empirical variogram lies outside of this envelope, there is evidence of spatial dependency in the data. Figure 1 displays the empirical variogram of logit-transformed schistosomiasis prevalence data from 299 schools in Namibia (Sousa-Figueiredo et al. 2015). This figure indicates that spatial correlation persists up to separation distances of 300km.

Variograms have been used to describe the spatial dependency in HAT, schistosomiasis and LF in a number of scenarios. Most commonly, the variogram is used as a preliminary exploratory tool to indicate whether spatial dependency needs to be considered when undertaking a regression analysis to explore the spatial distribution of either disease cases or their vectors. Additional uses in the literature include the investigation of appropriate survey sampling strategies that account for spatial heterogeneity in schistosomiasis and LF cases (Gyapong & Remme 2001; Sturrock et al. 2011), and to determine the appropriate placement of monitoring traps for tsetse (Sciarretta et al. 2005).

* 1. **Mapping spatial variability**

Knowledge of where tropical disease transmission is likely to occur and how many people it is likely to infect is very valuable, particularly when developing and implementing an intervention. Further, as disease programmes develop the technology or the resources to adopt more spatially targeted control strategies, information of disease risk at an increasingly fine spatial scale is becoming more sought after (Carter et al. 2000; Verity et al. 2014; Walz et al. 2015). Below we describe two approaches to mapping spatially variability in disease risk i.e. interpolation methods and spatial regression, specifically model-based geostatistics (Diggle & Ribeiro 2007). Both approaches are used to predict the spatial distribution of disease at locations across geographical space, with the latter having the additional benefit of providing information on the causes of the geographical variability. These approaches make explicit use of the spatial correlation in the data. Other spatial mapping approaches commonly applied in tropical disease research which do not explicitly incorporate spatial dependency, but instead assume that all of the spatial variability in the data can be explained by observed covariates, are briefly described at the end of this section.

* + 1. *Spatial interpolation methods*

Spatial interpolation methods fall into two categories: deterministic and stochastic (probabilistic). Deterministic methods are those which predictions of the value of Y(x) are determined by the values of observed nearby points with no assessment of uncertainty being made, with a common methods for this being weighted moving average approaches such as inverse distance weights (IDW). The IDW method assumes that the value of the variable of interest at an unsampled location is a weighted average of the surrounding sampled point values, with the weight of the contribution of each surrounding point being dependent on the distance between the sampled and unsampled locations. Hence if we’re interested in making a prediction at location x, we assume

\begin{equation*}
\hat{y}(x) = \frac{\sum_{i}w_{i}y(x_{i})}{\sum w_{i}}
\end{equation*}

where $w_{i}$, the inverse distance weight, is simply the inverse of the distance between location $x$ and location $x_{i}$ to some power $p$, and usually $p=1 \mbox{ or }2$. As well we choosing the form of $w_{i}$, the number of neighbouring points that contribute to the estimate also need to be considered. Popular options include including a fixed number of nearest neighbour points, or including all points within a fixed radius (de Smith et al. 2015). Figure 2 presents the inverse distance weighted schistosomiasis prevalence surface for northern Namibia on a 2.5km by 2.5km grid. In this instance, $p=2$ and all points were considered when determining the inverse distance weight.

Kriging is a stochastic spatial interpolation methods that was initially derived to aid those working in the mining industry to estimate the distribution of gold. It is from here which the term *geostatistics* originates (Krige 1951; Matheron 1976; Cressie 1990). Kriging was initially described as a form of weighted averaging i.e. $\hat{Y}(x) = \sum w_{i}(x)Y_{i}$, with the weights being selected to minimise the mean square prediction error in the resulting estimates, but in recent years it has been shown to fit into the spatial regression model framework (Cressie 1990; Diggle & Ribeiro 2007), incorporating the concept of a Gaussian process $S(x)$ as described above. There are a number of forms of kriging. For example, simple kriging refers to an approach which uses the mean of the observed data ($\bar{y}$) as an estimate of the true mean of $Y$ ($\mu$) when calculating the kriging weights, whereas ordinary kriging using the generalized least squares estimator ($\hat{\mu}$) to estimate the mean (Cressie 1993b; Diggle et al. 1998). Universal kriging assumes that the true mean of $Y$ is not a single value, but instead depends on covariates $d_{k}(x)$ such that $\mu(x) = \sum \beta_{k}d_{k}(x)$ where the covariates $d_{k}(x)$ may or may not be spatially varying (Goovaerts 1997). In practice, universal kriging involves applying simple kriging to the residuals of the detrended data $(y' = y-\hat{\mu}(x))$, then adding these back on to the global trend. Trans-Gaussian kriging is the term used to describe the kriging process when the surface of interest itself is not Gaussian, but there exists a transformation such that the transformed data can be considered to be approximately Gaussian. For example, in spatial epidemiology we often deal with prevalence or count data, which do not follow a Gaussian distribution. In this instance, transformations such as the *logit* transformation (*$\mbox{logit}(x) = \mbox{log}(1/(1-x))$* for prevalence data, or the log transform for count data need to be applied before kriging can be performed (Cressie 1993b).

[Insert Figure 2 here]

With the advent of new statistical methods that are able to directly model data from a range of distributions such as model-based geostatistics (see Section 3.2.2), ‘traditional’ kriging methods such as trans-Gaussian kriging has fallen out of favour with the statistical and disease research communities (Pullan et al. 2012), although due to its accessibility within GIS software such as ArcGIS, it is still frequently applied. Within the NTD literature, traditional kriging methods have been used to produce maps of tsetse fly abundance for HAT transmission risk (Sciarretta et al. 2005; Cano et al. 2007), schistosomiasis prevalence and snail abundance maps (Sturrock et al. 2011; Moser et al. 2014) and lymphatic filaraisis prevalence maps (Gyapong & Remme 2001; Onapa et al. 2005; Koroma et al. 2012).

* + 1. *Spatial regression*

Simple spatial interpolation methods such as IDW and simple and ordinary kriging are a valuable way of visualising spatial trends in data, however they rely upon the suitable selection of sampling points such that important changes in the outcome measure of interest may be omitted if there is heterogeneity in risk between sampled locations. Without information as to why the risk varies, these methods may result in these high (‘hotspots’) or low (‘cold spots’) risk areas being ‘smoothed out’, the consequences of which could be that those who are most at risk are excluded from control programmes. Spatial regression techniques incorporate spatial dependency into a regression framework, hence allow for an assessment of the effect of measureable factors on risk to be made without requiring the outcome of interest to be independent. As with generalised linear models, this information can then be used to predict risk at unsampled sites provided it is possible to obtain information on all of the covariates included in the regression model at these sites. In this chapter we focus on the spatial regression method known as model-based geostatistics as this approach is popular in the tropical disease literature (Pullan et al. 2012).

Model-based geostatistics (Diggle & Ribeiro 2007) embeds classical geostatistical techniques such as kriging into the generalised linear modelling framework. The benefit of this is that it relaxes the assumption that the outcomes of interest, or appropriate transformations thereof, must have a Gaussian distribution, and allows us to consider outcomes that follow a Bernoulli distribution (e.g. presence/absence data), a binomial distribution (e.g. prevalence data) and a Poisson distribution (e.g. incidence rates) as well as outcomes from other distributions in the exponential family. This extension of the geostatistical paradigm has led to model-based geostatistics being a popular method of assessing and subsequently mapping disease risk whilst accounting for the influence of (spatially varying) risk factors. For example, the binomial geostatistical model with the inclusion of risk factors is specified as:

\begin{eqnarray*}
Y(x)|S(x) &\sim& \mbox{Bin}(m(x),P(x)) \\
\mbox{logit}\{P(x)\} &=& \sum \beta_{k}d_{k}(x) + S(x) + Z(x)
\end{eqnarray*}

where $S(x)$ is a stationary Gaussian Process as described above and $d_{k}(x)$, $k=1,....,K$ are risk factors which may or may not be spatially varying.

Parameter estimation within the generalised linear models paradigm is undertaken using likelihood-based methods. In simple terms, using likelihood-based methods, the optimal value of the parameters in the model are the value that optimise the likelihood function i.e. a function which represents the chance of observing the data *y* given the parameter values. In extending the generalised linear model framework to include spatial correlation, maximum likelihood approaches become computationally complex (i.e. intractable), and subsequently alternative model fitting approaches are required. This has led to the adoption of a Bayesian framework, such that rather than assuming the parameters of interest in the model are fixed (and unknown), it is assumed that they are random variables, and hence each have their own distribution. More details about Bayesian methods can be found in Gelman et al. (2013) and Diggle & Ribeiro (2007). Within the Bayesian framework, the standard approach for addressing the issue of intractable likelihood has been to use Markov Chain Monte Carlo (MCMC) methods (Gilks et al. 1995). Statistical software commonly used to fit generalised linear geostatistical models, notably WinBUGS (Lunn et al. 2000) and the R package geoRglm utilise MCMC to make inference, and over the last decade a plethora of research articles have been produced which apply these techniques to tropical diseases (Schur et al. 2011; Schur et al. 2013; Oluwole et al. 2015; Lai et al. 2015; O’Hanlon et al. 2016). These analytical developments have coincided with the increased availability of spatially referenced data on both the diseases themselves and their potential risk factors (See Section 4). However, unlike likelihood-based inference for generalised linear model, MCMC methods require the user to make numerous subjective inputs when specifying the model which therefore requires the user having a reasonable working knowledge of the model fitting process. Further, the model fitting process can be computationally-intensive, and relies upon the user’s judgement to determine whether the MCMC algorithm has converged appropriately. As such, without sufficient knowledge and computing power, this modelling approach is considered to be difficult to perform.

In recognition of this, alternatives to MCMC have been sought, and the Integrated Nested Laplace approximation (INLA) algorithm is quickly being adopted as being a much less labour-intensive approach to fitting Bayesian models which contain a latent (i.e. unobserved) Gaussian process, such as geostatistical models (Rue et al. 2009; Lindgren et al. 2011; Lindgren & Rue 2015; Brown 2015). Whilst the user is still required to specify suitable initial values and tuning parameters, the model fitting process is much quicker, and there is far less ambiguity with regards to assessing the reliability of the output. The R-INLA package (www.r-inla.org) provides an interface to INLA (a free-standing external programme), thereby enabling a wide range of spatial latent variable models to be fitted in R using the INLA methodology including model-based geostatistical models (Bivand et al. 2015). Due to the flexibility of R-INLA, the geostatistical model-fitting process is currently relatively difficult for a non-expert however. Software which act as an intermediary between R-INLA and the less experienced geostatistical modeller are therefore being produced e.g. the package geostatsp (Brown 2015).

The benefits of INLA over MCMC model-fitting approaches could potentially have an impact on the accessibility of these modelling approaches to those in resource-poor settings, and further increase the feasibility of using the resulting predictive maps in a more dynamic way. For example, predictive maps could be updated with data collected in the field within a matter of hours, thus control programmes to make decisions based on the most accurate spatial information of the disease to be available.

Whilst geostatistical models are not the only spatial regression approach available for disease mapping, they are useful as they not only result in predictive maps of disease risk, they provide several other key outputs that may be of interest to the disease control community. Firstly, as with all generalised linear models, in addition to a predicted risk surface, information on the uncertainty in predictions can also be easily extracted either in the form of prediction intervals, or perhaps more intuitively, exceedance probabilities, $\mathbb{P}(Y > T)$, where T is a threshold of interest. Secondly, models of this form not only provide information on which of the considered covariates are associated with the disease outcome, but further (assuming parsimony was taken into consideration when undertaking model selection (Lunn et al. 2000; Spiegelhalter et al. 2002; Austin & Tu 2004; Hoeting et al. 2006)) provide information on the nature of those associations. Further, the estimated surface $\hat{S}(x)$ indicates which geographical areas have a higher or lower disease risk after accounting for measured risk factors, which may aid further investigations to identify other unmeasured elements that may be affecting disease risk. Finally, unlike some of the spatially implicit methods referenced below, they are not restricted to considering only environmental variables as risk factors, and other influential measures e.g. demographic, socio-economic can be incorporated, with the caveat being that maps of the resulting risk can only be produced if the risk factor is available at each location where predictions of risk are to be made.

Figure 2 presents the predicted mean schistosomiasis prevalence in Namibia, obtained using the R-INLA package. The binomial geostatistical model included the average maximum normalised difference vegetation index (NDVI) over a three year period obtained at 250m resolution (NASA 2016a), average annual rainfall at 100m resolution (WorldClim 2016) and topographical wetness index (TWI) derived from 90m resolution elevation data (Sørensen et al. 2006; NASA 2016b) as covariates. As with the map produced using IDW, the predictions were made on a 2.5km by 2.5 km grid. In comparison to the IDW map, the predicted prevalence is much more spatially variable as the relationship between the covariates and the prevalence survey data allows predictions to be made at locations far from those surveyed.

There are a plethora of additional examples of model-based geostatistics in the schistosomiasis literature including large scale schistosomiasis prevalence mapping (Schur et al. 2013; Lai et al. 2015), and intensity of infection in East Africa (Clements 2006), plus the distribution of freshwater snails in Lake Victoria (Standley 2012). Model-based geostatistical application to LF for sub-Saharan Africa are less common, with just a small number of papers using this approach to map LF prevalence at the large geographical scale (Slater & Michael 2013; Moraga et al. 2015), whereas (Kelly-Hope et al. 2006; Stensgaard et al. 2011) use model-based geostatistics to explore the spatial association between LF and malaria. Similarly, there are currently few application of model-based geostatistics to HAT transmission data, with publications predominantly focusing on Rhodesian HAT (Batchelor et al. 2009; Wardrop et al. 2010). As model-based geostatistics becomes more accessible through the development of more user-friendly software, it is possible that the popularity of this method will increase.

* + 1. *Common spatially implicit methods*

There are numerous other disease risk mapping approaches that do not explicitly incorporate spatial dependence, and instead assume that all of the spatially-structured variability can be explained by measurable risk factors, commonly environmental variables. These methods often fall under the category of species distribution models or ecological/environmental niche models. Here we briefly describe several approaches that have been applied in the tropical disease literature, predominantly to map the spatial variability in disease vectors, i.e. MaxEnt, discriminant analysis and boosted regression trees. These methods acknowledge that the host-parasite relationship is very complex and non-linear, hence intricate interactions need to be taken into consideration when trying to map disease risk. Although in theory it is possible to incorporate this complexity into spatial regression models such as geostatistical models, this would be at the expense of parsimony, and by extension, interpretability. These approaches generally make use of presence only or presence/absence data with the aim of predicting the probability that the species or disease of focus is present at a given location.

MaxEnt is a popular species distribution model as it only requires information on presence data to be able to elucidate the relationship between environmental variables (Phillips et al. 2006; Elith et al. 2006; Phillips & Dudík 2008). This approach is commonly applied in ecology to determine the spatial patterns of the species of interest using records of species sightings (i.e. presence only data), although it is also an alternative to logistic regression when there is doubt over the accuracy of absence data (Dicko et al. 2014). MaxEnt is a machine learning technique, which in this context refers to automatic approaches which find the best combination (including non-linear, complex combinations) of risk factors that best described the variability in the data whilst controlling for over-fitting (Elith & Leathwick 2009). This process is referred to as regularisation. The MaxEnt software enables users to fit models of this form to their presence only data either using the software interface itself (https://www.cs.princeton.edu/~schapire/maxent/), or by using functions within the R package dismo (Merow et al. 2013; Hijmans et al. 2016). MaxEnt, and its accompanying software, is however a good demonstration of the disconnect between researcher working in different disciplines as it has since been proven that MaxEnt and Poisson regression are equivalent (Renner & Warton 2013). By using regularisation techniques such as lasso or ridge regression (Hoerl & Kennard 1970; Tibshirani 2011), users can fit the equivalent of MaxEnt models to their presence only data using R packages such as glmnet (Friedman et al. 2015), without restricting themselves to the limitations of the MaxEnt software.

With regards to the three NTDs of focus, MaxEnt has been used both in an ecological and an epidemiological context. For example, it has been used to provide species distribution maps for guiding tsetse control strategies (Matawa et al. 2013; Dicko et al. 2014), and for establishing the distribution of snails (Stensgaard et al. 2013; Pedersen et al. 2014), and has further been used to explore the large scale spatial variability in LF (Slater & Michael 2012; Mwase et al. 2014) using disease presence data at the regional and national level respectively.

Popular binary classification methods that are frequently used to explore the spatial distribution of species, and perhaps less commonly, diseases, using both presence and absence data include discriminant analysis and decision trees. As with MaxEnt, the goal of these classification methods are to determine which combination of covariates best explains the data without focusing on the interpretability or parsimony. Non-linear discriminant analysis focuses on identifying a combination of continuous covariates to discriminate locations where the outcome of interest is likely to be present or absent, and has been used to map the distribution of disease vectors including mosquitoes, sandfly and tsetse fly (Robinson 2000; Rogers 2000; Rogers 2006). As the relationship between the outcome and the environment may be inconsistent across the full range of the environmental covariates under consideration, in many applications the environmental is stratified into areas of similar conditions. Non-linear discriminant analysis is then undertaken within each of the strata separately, such that within each strata there is a non-linear axis discriminating between the two groups in multivariate space (Rogers 2000). Non-linear discriminant analysis is equivalent to maximum likelihood classification, which is frequently implemented in GIS software to classify remotely sensed data, however machine learning-based techniques such as decision trees (e.g. random forests) and support vector machines have been shown to consistently outperform this approach (Lu & Weng 2007; Otukei & Blaschke 2010; Cianci et al. 2015).

The boosted regression trees approach is an increasingly popular method being used in species distribution modelling and is a combination of two methods: regression trees and boosting (De’ath 2007; Elith et al. 2008; Stevens & Pfeiffer 2011). Regression trees are a method of classification which aim to predict a continuous outcome. There are two stages to the process. Firstly the data are partitioned into small groups using a recursive partitioning approach such that covariate-based binary decision rules to successively partition the data into smaller groups that are relatively homogeneous with regards to their relationship to the outcome. Once these groups have been obtained, simple regression models are then fitted to the data in each group independently (De’ath & Fabricius 2000). Boosting is a method of improving the accuracy of a model, and are based on the concept that it is easier to build multiple models based on a less strictly accurate decision rules and take an average that it is to build a single model based on a highly accurate ones. Boosting is a sequential process such that models such as regression trees are fitted iteratively to the data, and at each iteration the emphasis shifts to improve those outcomes that were poorly predicted in the previous iteration (Schapire 2003; Elith et al. 2008). The boosted regression tree approach has been implemented in a number of recent prominent tropical disease epidemiology papers including an assessment of the global burden of dengue (Bhatt et al. 2013), the global map of dominant malaria vectors (Sinka, Bangs, Manguin, Coetzee, Mbogo, et al. 2010; Sinka et al. 2012), the global distribution of lymphatic filariasis (Cano et al. 2014). Each of these examples utilises presence and absence data rather than prevalence or abundance, with model-based geostatistics being the method of choice when suitable prevalence or abundance data are available (Hay, Battle, et al. 2013; Bhatt et al. 2015; Moraga et al. 2015; Grimes & Templeton 2015).

1. **Common issues in spatial analysis**

Spatial analyses of point-level data are sensitive to a number of factors which need to be accounted for when collating, analysis and interpreting the data. We highlight below several of the most common in tropical disease applications, i.e. the spatial scale at which the data are recorded, the sampling methods used to collect the data, and the problems associated with combining data from different sources.

* 1. **Spatial scale**

In this context, when we refer to spatial scale we are predominantly concerned with the scale at which we would like to make our spatial predictions. This scale may be influenced by the objectives of the analysis, the characteristics of the disease under consideration, and the spatial scale of the available covariates. For example, the objective of the study may be to map the spatial distribution of disease at the large spatial scale (national, regional level) for descriptive purposes or to provide estimates of global burden, or they may be required at a more detailed spatial scale in order to guide more targeted control interventions. The spatial scale of the former is therefore likely to be coarser than that of the latter. Consider for example a highly spatially heterogeneous disease such as HAT (see Section 1 for more details). Whilst predictions at a relatively coarse spatial scale are useful in providing a general sense of the distribution of disease (Simarro et al. 2010), due to the highly focal nature of the disease, important spatial variability in disease distribution will be masked at this scale and hence may produce inaccurate estimates of disease burden, or insufficient information to guide targeted intervention strategies (Sciarretta et al. 2005; Hackett et al. 2014). However, if disease risk is relatively spatially homogeneous within some geographical limit, for example soil-transmitted helminths, it may be a waste of resources to produce maps at the finer spatial level (Sturrock et al. 2010).

Covariates used in the spatial analyses of tropical disease data are often derived from remotely sensed (RS) data (see Section 5). When incorporating these covariates, it is important to recognise that these data are already a spatial aggregation of a spatially continuous phenomenon, with the initial size and positioning of the boundaries of each RS cell (referred to in the statistical literature as the *support*) being defined by the instrument being used and the organisation from which the data were sourced. Further, when undertaking the spatial analysis, these cells may be aggregated or disaggregated to ensure a consistent (and scientifically relevant) spatial resolution and alignment across all of the RS data being included in the analysis. The method by which the RS data are processed in order to be incorporated into the analysis can have a significant impact on the form of the relationship between RS data and the disease-related point-level data, and as such needs to be taken into consideration when evaluating the output (Atkinson & Graham 2006; Raj et al. 2013; Hamm et al. 2015). This is a known problem in spatial statistics, referred to as the modifiable areal unit problem (MAUP) (Jelinski & Wu 1996; Cressie 1996; Gotway & Young 2002). As a consequence, researchers may find that they draw different and sometimes opposing conclusions about the influence of a spatially aggregated covariate and the disease-related outcome of interest.

Whilst the MAUP may cause inconsistencies in the relationship between covariates and the disease-related outcome, it is often the case that the relative influence of covariates naturally differs at different spatial scales. For example, in epidemiology studies the dominant disease drivers at the large geographical scale tend to be associated with the climate and environment e.g. rainfall, temperature, elevation whereas at the smaller scale, the socio-demographic characteristics tend to become more influential (Simoonga et al. 2009; Hamm et al. 2015). Similarly, in ecological studies e.g. when exploring the spatial distribution of disease vectors, the dominant drivers at the large scale drivers are similar to those found in epidemiological studies, whereas the environmental nuances e.g. fragmentation of habitat, river flow speed, steepness of the landscape need to be taken into consideration at the smaller spatial scale (Guerrini et al. 2008; Jacob et al. 2013; Hardy et al. 2015; Mweempwa et al. 2015).

* 1. **Spatial bias**
     1. Sources of spatial bias

It is often the case that spatially-referenced data that are used to explore the spatial variability in disease risk have been collected for a different purpose, with the spatial analysis being an opportunistic addition. As a result, when undertaking a post hoc spatial analysis, it is essential that the analyst is aware of any potential sources of spatial bias and adapt their models accordingly (Wardrop et al. 2014). For example, when conducting a survey, the location of samples is often influenced by the accessibility of an area such that vector sampling sites may not include densely forested areas, or epidemiological surveys may not be conducted in the more remote, impoverished communities. If collating historical data, differing diagnostic tools with non-equivalent sensitivity and specificity may be used in different geographical areas, or only data considered to be of epidemiological/ecological interest might be reported, excluding those geographical areas where disease prevalence or vector abundance is found to be low. Further, areas may be preferentially sampled i.e. targeted specifically because they are suspected to have high or low disease risk. If such biases are identified, it’s important that these potential biases are accounted for in the subsequent spatial analysis (Phillips et al. 2009; Diggle et al. 2010; Fourcade et al. 2014; Grimes & Templeton 2015; Giorgi et al. 2015; Diggle & Giorgi 2015).

* + 1. Spatial sampling design

When designing studies to explicitly explore spatial variability, the spatial sampling design needs to be carefully considered, which includes the consideration of where to sample, how many locations to sample and how much data to collect at each sampled location e.g. how many school-children to survey per sampled school. Common sampling strategies employed in epidemiological surveys generally do not explicitly consider the spatial variation in the disease distribution, and instead are designed to provide population-level summaries. These are referred to as design-based strategies. For example, at the small geographical scale, commonly employed design-based approaches include simple random sampling (where a suitable sampling frame is available) or systematic sampling using techniques such as the “spin the bottle” approach (Bostoen & Chalabi 2006). At the larger scale, in order to make the data collection more logistically feasible, cluster sampling is frequently employed where firstly the clusters (commonly villages or schools) are selected, following which individuals or households within the selected clusters are sampled. Sample size calculations for these types of survey are based on obtaining a sufficiently precise estimate of an area-level summary measure e.g. prevalence. Whilst these approaches provide unbiased estimates of overall prevalence, they are not optimal for determining the spatial variability in disease risk within the study area. For highly spatially heterogeneous diseases, this may result in the important geographical areas of disease transmission being missed. A simple approach to addressing this is to use stratified sampling where the strata are geographically contiguous areas, with random or cluster sampling being used within the strata. Often, these strata are selected using environmental features, and whilst this may result in reduced within-strata spatial heterogeneity, this approach may not enable spatial variability to be explored in continuous space. Further, neither random, cluster or stratified sampling strategies account for spatial dependency when determining sample size, as in a scenario where there is high spatial dependency, sampling two locations that are geographically close together may not provide much more information than if one point were sampled in that area. Thus, the effective sample size in areas with high spatial dependency is reduced (Griffith 2005).

Model-based approaches are those that aim to learn about the spatial correlation structure of the data and make predictions at unsampled locations, as opposed to determining a single value to represent an entire area such as a mean. Within the spatial statistical literature, the location of the sampled points, as opposed to the quantity, is generally more of a consideration when determining an optimum sampling design (Diggle & Lophaven 2006; Wang et al. 2012; Evangelou & Zhu 2012). A commonly applied model-based sampling strategy for the purpose of spatial prediction is based on sampling at regular intervals in one dimension (transect) or two dimensions (lattice) (Diggle & Ribeiro 2007; Miller et al. 2013; Buckland et al. 2015). Numerous examples of this can be found in the ecological literature, whereas spatial sampling has been less of a focus in tropical disease epidemiology, with examples of its consideration being relatively sparse (Gyapong et al. 2002; Zouré et al. 2011). If the focus of the research is to gain a better understanding of the spatial correlation structure of the data, the lattice design should be supplemented to also include close pairs of points within the lattice in order to estimate the spatial dependency over short distances (Diggle & Ribeiro 2007).

In practice, spatial sampling strategies for diseases in resource-poor settings are likely to be influenced by pragmatism, with the number of sampling locations, plus the number sampled at each individual location, being limited by time and resources as well as scientific rigour (Magalhães et al. 2011). As disease landscapes change, as a result of a change in environment, behaviour as well as due to the influence of control, more consideration needs to be given to sampling strategies that promote the production of dynamic as opposed to static disease risk maps e.g. adaptive spatial sampling (Peyrard et al. 2013; Siegfried & Siegfried 2014) is an approach by which sampling is undertaken sequentially such that results of the previous samples are used to guide the selection of the next.

* 1. **Combining different resources**

Due to the lack of large-scale, spatially dense data from a single survey, disease maps at the national or regional level are often produced using historical data from multiple sources. Combining data from historical surveys can be very challenging, as the data under consideration often includes that collected using a variety of survey designs e.g. different sampling strategies, diagnostic tests, target populations etc. In failing to incorporate these data heterogeneities into the model, the resulting predicted risk surfaces may be misrepresentative of the true underlying risk (Grimes & Templeton 2015; Giorgi et al. 2015; Diggle & Giorgi 2015). For example, it may be more appropriate to split the data into smaller, more homogeneous datasets before undertaking the analysis, to derive additional covariates which attempt to measure these differences, or to extend the model to include a temporal component.

1. **Sources of spatially-referenced data**

Spatially-referenced data is available in two categories i.e. vector (points, lines, polygons) and raster. Due to the prolific increase in geospatial technology and the drive for open access data, there is now an abundance of spatially-referenced data that is relevant to disease control in developing countries both publically and commercially available, plus we have the tools available to easily generate data ourselves. Below we describe sources of data applicable to tropical disease control, with a focus on the three NTDs of interest (see Table 2).

* 1. **Disease transmission data**

With the growing realisation that open access to survey data can have wide public health benefits, many leading scientific journals and research funders now require researchers to make their data available publically available and numerous online global disease databases are being developed (Flueckiger et al. 2015). These databases often collate, and if necessary georeference, available disease data obtained from sources such as the published literature, or directly from the researchers involved in collecting the data.

* + 1. Global disease databases

There are a growing number of initiatives who aim to collate all historical spatially-referenced prevalence/individual case data at the global scale for a specific disease. Example of this include the Global Atlas of Helminth Infections (GAHI, [www.thiswormyworld.org](http://www.thiswormyworld.org)) which includes historical and contemporary data for LF, schistosomiasis and soil-transmitted helminths (Brooker et al. 2010; Cano et al. 2014; Sime et al. 2014), the Global NTD database (GNTD) which focuses on schistosomiasis, and the Atlas of Human African Trypanosomiasis (Simarro et al. 2010; Lumbala et al. 2015). Not all data presented on these sites are available to download however, although some data may be shared on request. NTD Mapper ([www.ntdmap.org](http://www.ntdmap.org)) is an NTD-specific website, currently focused on trachoma, Loa loa, LF, onchocerciasis, schistosomiasis and STH. Data presented come from a variety of sources, including prevalence surveys and literature searches with varying geographical scales. A large proportion of the data presented are available for download.

Contemporary and historical disease data may also be accessible from unpublished sources. For example Healthmap ([www.healthmap.org](http://www.healthmap.org)), founded in 2006, is an online resource for informal disease surveillance data, collating information from sources such as ProMed Mail, WHO and Google News to produce maps of reported cases of a variety of diseases. This resource cannot be considered to be a complete record of cases, but may allow geographical areas of interest to be identified.

* + 1. Vector/intermediate-host databases

There is a wealth of vector distribution data in the published literature, and as with disease data, there are initiatives in place to collate this information and make it more easily accessible. There are however many gaps remaining in data relating to the three NTDs of focus. For example, the Food and Agriculture Organization of the United Nations (FAO) are developing an Atlas of tsetse, which in combination with the disease cases atlas (Atlas of Human African Trypanosomiasis) is intended to provide comprehensive information on disease transmission risk to national control programmes (Cecchi et al. 2014; Cecchi et al. 2015). A similar project has been undertaken by the malaria control community to create maps of malaria parasite prevalence (Guerra et al. 2007). This tsetse data does not however appear to be publically available. Due to its linkages with malaria, there is an extensive amount of information available on the distribution of *Anopheles* which may be of interest to the LF elimination community, notably that produced by the Malaria Atlas Project, although this focuses on mapping presence of mosquitoes at the large geographical scale and may therefore be of limited operational use. Less has been done to develop resources relating to the distribution of *Culex* mosquitoes, despite being transmitters of LF and West Nile virus in sub-Saharan Africa. To our knowledge, there are no resources relating to the spatial distribution of schistosomiasis-transmitting snails available. It is also worth noting that it is rare to find spatial information relating to the infection status of the vectors/intermediate hosts for any of the diseases under consideration.

* + 1. Spatial data collection

As well as using previously collected data, researchers may also want to collect their own spatially-referenced data to analyse. The capacity for the public to record precise geographical location data has been available to the public since May 2000, when the US government switched off the pseudorandom errors in the publically available GPS signal, known as ‘selective availability’. Since this date it has become increasingly easy to record location information. Initially, dedicated GPS receivers were required to record coordinates during field activities, which were often prohibitively expensive. However, since the proliferation of the GPS-enabled smartphone and tablet computers, this is no longer the case. Further, unlike many dedicated GPS receivers, smartphones/tablets have many additional features that can be taken advantage of when collecting disease data in the field. For example, they could be used to collect survey responses which can be automatically uploaded to a database once a data connection is available, thus circumventing the need for manual data entry. There are an increasing number of products and services available to develop these data collection platforms, some of which are designed to be accessible to researchers who do not have any programming skills. One popular example of such a product is the open source, Android-based OpenDataKit (ODK, www.opendatakit.org) (Anokwa et al. 2009; King et al. 2013). ODK is a suite of tools that enable users to develop their own data collection forms that can incorporate text and numeric data, GPS coordinates, photographs, videos are barcodes, which can be used to collect field data without being connected to either the mobile network provider or the internet. An internet connection is only required to upload the data, with the data being uploaded to online servers that are either hosted by Google’s App Engine, or hosted locally. ODK has been successfully used to collect tropical disease data (King et al. 2013; Sime et al. 2014; Tom-Aba et al. 2015; Fähnrich et al. 2015), the largest scale of which has been in relation to trachoma mapping (Pavluck et al. 2014), such that between 2012 to 2015 the Global Trachoma Mapping Project (GTMP, http://www.sightsavers.org/gtmp/) collected data from 2.6 million people across 29 countries using ODK. Whilst large scale mapping projects tend to be led by dedicated survey teams, as smartphones become more ubiquitous in developing countries, there is the growing potential to collect spatially-referenced health data directly from the affected communities (Stanton, Molineux, et al. 2016). Community members can either be requested to submit data related to specific field activities, or data can be obtained on a more voluntary or passive basis. Participatory data collection approaches have the potential to provide vast amounts of real-time spatially-referenced data that could be of relevance to disease control (Hay, George, et al. 2013). The most successful examples of the use of crowd-sourced spatial data to improve health outcomes in the developing world primarily relates to improving geographical maps to aid disaster response (Zook et al. 2010; Médecins Sans Frontières 2014), and detecting epidemics (Broniatowski et al. 2014; Milinovich et al. 2015) although there are also examples in the literature of participatory mapping of malaria vector exposure (Fuller et al. 2014; Mwangungulu et al. 2016).

* 1. **Population movement data**

A limitation of spatial methods is that they make the implicit assumption that the subject of interest is associated with a single geographical location e.g. survey participants are usually assigned the coordinates location of their homes, schools or villages when undertaking prevalence surveys, when it is of course recognised that exposure to disease risk may not occur at that location. In a purely spatial models this limitation is often accounted for by considering environmental risk factors within a fixed sensible buffer of the assigned location, or asking participants questions that might identify activities that expose them to risk outside of their homes/schools/villages. These approaches cannot however capture all of the intricacies of the participant’s movements that may be linked to disease risk, and could weaken the spatial signal in the outcome of interest. There is therefore a growing interest in improving measures of human mobility both in terms of large scale patterns such as migration and small scale day-to-day movement (Kraemer et al. 2015). One research area of growing interest is the harnessing of mobility information collected by mobile phone operators via anonymised call logs. This human mobility data has been shown to be complementary to travel survey data and provide a deeper understanding of the spatial dynamics of infectious diseases in developing countries over time (Buckee et al. 2013; Wesolowski et al. 2014; Wesolowski et al. 2015; The Flowminder Foundation n.d.). On a smaller geographical scale, GPS-enabled wearable technology, which in its most basic form consists of a small GPS data logger which records coordinates at regular time intervals, can provide detailed information on human movement that might be of relevance to disease control. For example, GPS data loggers have been used to measure differences infected and uninfected mother and children’s spatial movement patterns in relation to schistosomiasis (Stothard et al. 2011), and assess the impact of LF-related morbidity on mobility (Stanton, Yamauchi, et al. 2016).

* 1. **Geographical disease risk factor data**

Here we restrict ourselves to consider physical measures of the natural and man-made environment as opposed to other risk factors that can be spatially-referenced e.g. socio-economic measures associated with a particular location.

* + 1. *Natural geographical features*

Remotely sensed (RS) data refers to data acquired from a distance, with the term being most commonly associated with data obtained from satellites. In areas where in situ environmental data is sparse, such as sub-Saharan Africa, satellite-derived RS data are often the only environmental data available at a spatially and temporally consistent scale, hence are an exceptionally valuable resource. Whilst the exact number of operational satellites is unknown, approximately 940 spacecraft have been launched since October 1957 for the purpose of earth science (Belward & Skøien 2014; NASA n.d.). These satellites tend to operate at a relatively low altitude (~700-800km) and provide information at a range of spatial resolutions. As the value of these data continues to be realised, RS datasets that were once only available at considerable financial cost are being released to researchers or the general public for free. The Landsat and ASTER programmes are good examples of this.

NASA’s Landsat programme has been providing data on a global scale since its launch in 1972, and in 2008 decided to release all data to the public (Woodcock & Allen 2008). Landsat 1 was equipped with a multispectral scanner (MSS) and initially provided data in four spectral bands (red, green and two infrared) at a spatial resolution of 80m, with data for each geographical location being provided every 18 days. Most recently, Landsat 8 was launched in 2013 and is equipped with two instruments, namely the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS) which provide data on the visible spectrum (red, green, blue), plus infrared, near infrared and short-wave infrared at 30m resolution, thermal infrared at 100m resolution and a grayscale/panchromatic image is produced at a resolution of 15m. Data are provided every 16 days, and are available to download using the USGS Earth Explorer (<http://earthexplorer.usgs.gov/>) shortly after collection.

ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) is a Japanese owned sensor on board NASA’s Terra satellite. It was launched in 1999 and is still providing data today, despite its initial five year design life. All ASTER imagery was made available to the public in April 2016, prior to which only digital topographical maps were available. ASTER data is comprised of 14 spectral bands: four bands in the visible and near infrared ranges (15m resolution), six bands in the shortwave infrared ranges (30m resolution) and five bands in the thermal infrared ranges (90m) resolution (Abrams et al. 2015). As with Landsat, data re provided every 16 days, and are available to download using the USGS Earth Explorer.

Products such as Landsat require the user to be well-versed in remote sensing and/or GIS in order to extract information that may be of relevance to their disease of focus e.g. Landsat data can be used to derive measures of vegetation such as the Normalised Difference Vegetation Index (NDVI), or classification algorithms can be applied to identify land cover type. In acknowledgement of this limitation, there are an increasing number of pre-processed resources available to non-RS/GIS experts. With regards to land cover for example, whilst users continue to have the opportunity to use RS data such as AVHRR (Advanced Very High Resolution Radiometer, ~1km resolution) , MODIS (Moderate Resolution Imaging Spectroradiometer, 250m-1km resolution) and Landsat to use image classification techniques to derive land cover categories of interest (Townshend et al. 1991; Friedl & Brodley 1997), there are also a number of global-scale land cover products freely available. The disadvantage of using these products is that they provide land cover information for only one time point, which may not match that of the epidemiological data under consideration. Examples of global land cover products include MODIS Land Cover (250m resolution, annual data 2001-2012; Friedl et al. 2002; Friedl et al. 2010), GLC-SHARE (1km, data collated between 1998-2012), GlobCover (300m resolution, 2009 RS data; (Congalton et al. 2014)) and GlobeLand30 (30m resolution, 2010). These data have been used extensively in the development of predictive disease models and maps, including the NTDs of focus e.g. tsetse mapping (Kitron et al. 1996; Odiit et al. 2006; Guerrini et al. 2008; DeVisser & Messina 2009).

Elevation is an important factor when considering the distribution of many tropical diseases, as it often impacts transmission e.g. through influencing vector habitat suitability. The highest resolution global, publically available elevation data is currently 30m, and was collected by the Shuttle Radar Topography Mission (SRTM) in February 2000. This data was made available to the public at its finest scale in 2015, and can be accessed using the USGS Earth Explorer (http://earthexplorer.usgs.gov). As well as elevation, this data can be used to derive information on river networks using hydrological modelling. For example, the HydroSHEDS (HYDROlogical data based on Shuttle Elevation Derivatives at multiple Scales) has produced river network and drainage basin datasets using 90m resolution SRTM data (http://hydrosheds.cr.usgs.gov/).

Data from civilian earth observation satellite sensors provide a vast amount of contemporary information on the global environment, however given that the multispectral spatial resolution using rarely exceeds 30m, important small-scale details associated with disease risk may be masked (see Section 4). When considering disease risk at a reasonable large geographical scale, this loss of information may not be of importance. However, as interest in more targeted disease control efforts, more detailed environmental information may be required. Commercial earth observation satellites include RapidEye (6.5m multispectral resolution), SPOT (6m multispectral resolution), IKONOS (3.2m multispectral resolution), QuickBird (2.62m, decommissioned in 2015), Pleiades (2m multispectral resolution), GeoEye (1.84m multispectral resolution) and WorldView (1.24 multispectral resolution), and archived data can be purchased from the satellites operators (DigitialGlobe, GeoEye, Blackbridge etc) or from various value added resellers. The price of these products is very variable, depending on the resolution of the image, the size of the area required and the time period it is required for, and as such it has not been extensively applied to NTD research (Hamm et al. 2015), although the potential of this resource is becoming increasingly recognised as mapping disease risk at the small (micro) spatial scale becomes increasingly important, particularly for diseases where targeted control is possible (Bousema et al. 2012; Jacob et al. 2013; Walz et al. 2015; Shaw et al. 2015; Rayaisse et al. 2015). There are a number of initiatives to release archived very high resolution data for free however. For example the European Space Agency (ESA, https://earth.esa.int) has made their own archives of SPOT and RapidEye data available to the public following a brief registration process, whereas other datasets (e.g. QuickBird, GeoEye, WorldView, PLEIADES) are available to researchers on the submission and acceptance of a research proposal. The ESA does not however store all data produced by these commercial satellites in their archives hence there are still gaps in both geographical space and time. A small amount of very high resolution data can also be obtained from Open Aerial Map (<https://openaerialmap.org>), which is a platform for sharing open-licenced RS imagery, although this resource is in its early stages of development.

Google Earth is another valuable resource for researchers as it enables the user to visualise high resolution products such as those listed above and to create simple vector features (points, lines, polygons). Whilst this is an excellent resource, it’s limited in use as users only have access to data from a small subset of time points, the resolution at which data are available are not geographically consistent, and users are not able to access the raw data, and subsequently any of the additional spectral bands (Lozano-Fuentes et al. 2008; Jacobson et al. 2015). Finally, unmanned aerial vehicles (UAVs), commonly known as drones, are also now recognised as an accessible method of generating high resolution remotely sensed data that is applicable to public health (Fornace et al. 2014).

* + 1. *Man-made geographical features*

Whilst historically there is a disparity in the availability of accurate, digitised, georeferenced data on the man-made landscape (roads, settlements, health facilities etc) between the developed and developing world, advances in geospatial technologies are reducing this knowledge gap. We focus on two types of data resources in this paper, namely voluntary geographic information (VGI) resources and remotely sensed resources.

VGI resources are geographical information resources that are compiled from geographically referenced data provided by volunteers. Other terms used for this are participatory mapping or crowdsourced mapping (Goodchild 2007; Sui et al. 2012). An excellent example of a VGI resource on a global scale is OpenStreetMaps (OSM, www.openstreetmap.org) (Budhathoki & Haythornthwaite 2012; Neis & Zielstra 2014). OSM was founded in 2004 with the goal of creating a geographic database that was both free to use and contribute to, thus harnessing the power of local knowledge, and the enthusiasm of experienced and amateur mappers. Data contributed to OSM can be that collected directly in the field using GPS-enabled devices, or digitised from sources such as aerial images or locally generated paper-based maps. Vector data created by users are publically available via numerous resources such as Planet OSM (http://planet.openstreetmap.org/) or using the OSM import functionality within QGIS (http://wiki.openstreetmap.org/wiki/QGIS). See <http://wiki.osmfoundation.org/wiki/How_To_Get_OpenStreetMap_Data> a more comprehensive list of sources. The level of detail available within a geographical area of interest, and the quality of the data, is however dependent on the contributions made by the OSM community and hence is not globally consistent (Haklay 2010; Neis et al. 2011; Arsanjani et al. 2013; Barron et al. 2014).

In recognition of the power of maps for humanitarian aid and economic development in the developing world, the Humanitarian OpenStreetMap Team (HOT) was established (www. hotosm.org). Through the medium of the HOT Task Manager, OSM users are asked to assist in improving the quality of maps in particular areas of interest with regards to humanitarian response. This was first implemented in response to the 2010 earthquake in Haiti, and more recently the HOT has been involved in mapping areas in West Africa affected by Ebola (Teng et al. 2014; Soden & Palen 2014; Médecins Sans Frontières 2014; Palen & Soden 2015; Koch 2015). Not only do these maps provide valuable information to those directly responding to these crises, but they leave behind a legacy of freely available geographical information that can be used to benefit these population further (Eckle & Albuquerque 2015; Stevens & Pfeiffer 2015). In addition to providing maps for immediate humanitarian response, HOT also support community development. Current projects supported by HOT include mapping all residential areas, all building plus water and sanitation facilities southern parts of Malawi affected by extensive flooding in January 2015 in order to improve flood risk management in area, and similarly mapping infrastructure relating to flood risk in Dar es Salaam, Tanzania following a series of flooding events over the last five years. These community development initiatives are part Missing Maps Project ([www.missingmaps.org](http://www.missingmaps.org)), which supports HOT through promoting the mapping of vulnerable populations pre-emptively rather than responsively. Figure 3 presents a comparison between Google Maps and OSM of the historic downtown of Monrovia, Liberia, highlighting the level of detail available using OSM as a result of the Ebola crisis.

As these spatial data increase in accuracy, we speculate that older resources for man-made features will be replaced by more contemporary OSM data, which have the capacity to improve our knowledge on the spatial distribution of disease risk. For example, the population mapping project WorldPop (http://www.worldpop.org.uk/) has started to integrate OSM data into their maps in order to improve the precision of their 100m spatial resolution population density maps (Linard et al. 2014; Sorichetta et al. 2015). These population data can subsequently be used to obtain more accurate estimate of disease burden. An improved knowledge of infrastructure, and subsequently the spatial distribution of the human population, can also aid in understanding the distribution of disease risk at finer spatial scales through the derivation of relevant risk indices. For example, detailed land use and transport network data can assist in determining the accessibility of an area, or the proximity between an at-risk population and geographical risk factors. More practically, resources such as OSM can be used to better plan the distribution of interventions and treatments.

[Insert Figure 3 here]

1. **Conclusions**

It is encouraging to see the quantity of spatially-referenced data that is being made publically available, particularly in terms of disease prevalence and remotely sensed data. Not only is the spatial resolution of the available data becoming increasingly fine, some data, particularly remotely sensed data such as Landsat and ASTER are available at a continuous temporal resolution in near-real time. Thus detailed contemporary spatio-temporal models and maps are increasingly achievable. This trend needs to continue to allow researchers and control programme implementers to maximize on their potential to benefit disadvantaged populations. There are however still gaps in the data that need to be addressed in order to understand the changing epidemiology of the diseases of interest and increase the likelihood of sustained control and elimination. For example, more regularly collected geographically referenced disease monitoring and surveillance data is required at the spatial scale of interest to the respective control programmes. Further, and crucially, in order to avoid premature cessation of control efforts, more sensitive diagnostics, plus sustained vector/intermediate host surveillance are required for all three diseases (Hawkins et al. 2016; Hotez et al. 2016; Stothard et al. 2017).

This chapter has demonstrated the utility of spatial methods in tropical epidemiology, showing that a great deal has been learnt about the spatial patterns of diseases targeted for control and elimination in recent years as a result. The focus should now be on putting the information acquired from spatial statistical methods into operational use (Bergquist et al. 2015). There have been some examples of this within the context of a HAT control, a disease that is close to elimination and highly spatially heterogeneous, and hence engenders a more targeted approach (Sciarretta et al. 2005; Sciarretta et al. 2010). Current control strategies for the other two diseases of focus treat the selected at-risk population with preventive chemotherapy indiscriminately at present. However, as prevalence decreases as a result of control a more targeted approach may be necessary, particularly if the decrease is not spatially uniform. It is therefore essential to consider how statistical models and spatially-referenced data can be fully utilised to make the end stages of these programmes as resource efficient as possible. It is however worth bearing in mind that in situations where indiscriminate control strategies are much more operationally feasible than the costs associated with developing a targeted approach, that the former may be a more valid elimination strategy than the latter for some diseases (Stothard et al. 2014). In developing risk maps to inform a public health intervention it is always necessary to be mindful on the fine balance between producing risk maps that are incredibly detailed and accurate, and producing maps that are operationally useful. Thus, the lines of communication between the data providers, the statistical modellers and the disease control implementers should always be open to maximise the potential impact of the end product, and ultimately reduce the burden of disease.

**References**

Abrams, M. et al., 2015. The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) after fifteen years: Review of global products. *International Journal of Applied Earth Observation and Geoinformation*, 38, pp.292–301.

Adenowo, A.F. et al., 2015. Impact of human schistosomiasis in sub-Saharan Africa. *The Brazilian Journal of Infectious Diseases*, 19(2), pp.196–205.

Anokwa, Y., Hartung, C. & Brunette, W., 2009. Open source data collection in the developing world. *Computer*, 42(10).

Arsanjani, J. et al., 2013. Assessing the quality of OpenStreetMap contributors together with their contributions. In *Proceedings of the AGILE conference*.

Atkinson, P.M. & Graham, A.J., 2006. Issues of scale and uncertainty in the global remote sensing of disease. *Advances in parasitology*, 62, pp.79–118.

Austin, P.C. & Tu, J. V, 2004. Bootstrap Methods for Developing Predictive Models. *The American Statistician*, 58(2), pp.131–137.

Baddeley, A. et al., 2014. On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*, 84(3), pp.477–489.

Barron, C., Neis, P. & Zipf, A., 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, 18(6), pp.877–895.

Batchelor, N.A. et al., 2009. Spatial predictions of Rhodesian Human African Trypanosomiasis (sleeping sickness) prevalence in Kaberamaido and Dokolo, two newly affected districts of Uganda. J. Raper, ed. *PLoS neglected tropical diseases*, 3(12), p.e563.

Belward, A.S. & Skøien, J.O., 2014. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103, pp.115–128.

van den Berg, H., Kelly-Hope, L.A. & Lindsay, S.W., 2013. Malaria and lymphatic filariasis: the case for integrated vector management. *The Lancet. Infectious diseases*, 13(1), pp.89–94.

Bergquist, R. et al., 2015. Surveillance and response: Tools and approaches for the elimination stage of neglected tropical diseases. *Acta tropica*, 141(Pt B), pp.229–34.

Bhatt, S. et al., 2015. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, 526, pp.207–211.

Bhatt et al., 2013. The global distribution and burden of dengue. *Nature*, 496, pp.504–507.

Bivand, R.S., Gómez-Rubio, V. & Rue, H., 2015. Spatial Data Analysis with R - INLA with Some Extensions. *Journal of Statistical Software*, 63(20), pp.1–31.

Bockarie, M.J. et al., 2013. Preventive chemotherapy as a strategy for elimination of neglected tropical parasitic diseases: endgame challenges. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368(1623).

Bostoen, K. & Chalabi, Z., 2006. Optimization of household survey sampling without sample frames. *International Journal of Epidemiology*, 35(3), pp.751–755.

Bousema, T. et al., 2012. Hitting hotspots: spatial targeting of malaria for control and elimination. *PLoS medicine*, 9(1), p.e1001165.

Broniatowski, D.A., Paul, M.J. & Dredze, M., 2014. Twitter: big data opportunities. *Science (New York, N.Y.)*, 345(6193), p.148.

Brooker, S., Hotez, P.J. & Bundy, D.A.P., 2010. The global atlas of helminth infection: mapping the way forward in neglected tropical disease control. *PLoS neglected tropical diseases*, 4(7), p.e779.

Brown, P.E., 2015. Model-Based Geostatistics the Easy Way. *Journal of Statistical Software*, 63(12), pp.1–24.

Buckee, C.O. et al., 2013. Mobile phones and malaria: Modeling human and parasite travel. *Travel medicine and infectious disease*, 11(1), pp.15–22.

Buckland, S.T. et al., 2015. *Distance Sampling: Methods and Applications*, Springer.

Budhathoki, N.R. & Haythornthwaite, C., 2012. Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist*, 57(5), pp.548–575.

Cano, J. et al., 2007. Spatial and temporal variability of the Glossina palpalis palpalis population in the Mbini focus (Equatorial Guinea). *International Journal of Health Geographics*, 6(1), p.36.

Cano, J. et al., 2014. The global distribution and transmission limits of lymphatic filariasis: past and present. *Parasites & vectors*, 7(1), p.466.

Carter, R., Mendis, K.N. & Roberts, D., 2000. Spatial targeting of interventions against malaria. *Bulletin of the World Health Organization*, 78(12), pp.1401–1411.

Cecchi, G. et al., 2014. Assembling a geospatial database of tsetse-transmitted animal trypanosomosis for Africa. *Parasites & Vectors*, 7(1), p.39.

Cecchi, G. et al., 2015. Developing a continental atlas of the distribution and trypanosomal infection of tsetse flies (Glossina species). *Parasites & Vectors*, 8(1), p.284.

Cianci, D. et al., 2015. Modelling the potential spatial distribution of mosquito species using three different techniques. *International Journal of Health Geographics*, 14(1), p.10.

Colley, D.G. et al., 2014. Human schistosomiasis. *The Lancet*, 383(9936), pp.2253–2264.

Congalton, R. et al., 2014. Global Land Cover Mapping: A Review and Uncertainty Analysis. *Remote Sensing*, 6(12), pp.12070–12093.

Cressie, N., 1996. Change of support and the modifiable areal unit problem.

Cressie, N., 1993a. Estimation of the variogram. In *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics, pp. 69–82.

Cressie, N., 1985. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5), pp.563–586.

Cressie, N., 1993b. *Statistics for Spatial Data*, John Wiley & Sons.

Cressie, N., 1990. The origins of kriging. *Mathematical Geology*, 22(3), pp.239–252.

Danso-Appiah, T., 2016. Schistosomiasis. In Springer International Publishing, pp. 251–288.

De’ath, G., 2007. Boosted trees for ecological modeling and prediction. *Ecology*, 88(1), pp.243–51.

De’ath, G. & Fabricius, K., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11), pp.3178–3192.

DeVisser, M.H. & Messina, J.P., 2009. Optimum land cover products for use in a Glossina-morsitans habitat model of Kenya. *International journal of health geographics*, 8(1), p.39.

Dicko, A.H. et al., 2014. Using species distribution models to optimize vector control in the framework of the tsetse eradication campaign in Senegal. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28), pp.10149–54.

Diggle, P. & Lophaven, S., 2006. Bayesian Geostatistical Design. *Scandinavian Journal of Statistics*, 33(1), pp.53–64.

Diggle, P.J. & Giorgi, E., 2015. Geostatistical mapping of helminth infection rates. *The Lancet. Infectious diseases*, 15(1), pp.9–11.

Diggle, P.J., Menezes, R. & Su, T., 2010. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2), pp.191–232.

Diggle, P.J. & Ribeiro, P.J., 2007. *Model-based Geostatistics*, Springer.

Diggle, P.J., Tawn, J.A. & Moyeed, R.A., 1998. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3), pp.299–350.

Eckle, M. & Albuquerque, J. de, 2015. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. In *Proceedings of the ISCRAM Conference*.

Elith, J. et al., 2006. Novel methods improve prediction of species’ distributions from occurrence data. *Ecography*, 29(2), pp.129–151.

Elith, J. & Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1), pp.677–697.

Elith, J., Leathwick, J.R. & Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), pp.802–813.

Evangelou, E. & Zhu, Z., 2012. Optimal predictive design augmentation for spatial generalised linear mixed models. *Journal of Statistical Planning and Inference*.

F. Dormann, C. et al., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5), pp.609–628.

Fähnrich, C. et al., 2015. Surveillance and Outbreak Response Management System (SORMAS) to support the control of the Ebola virus disease outbreak in West Africa. *Eurosurveillance*, 20(12).

Fenwick, A., Rollinson, D. & Southgate, V., 2006. Implementation of human schistosomiasis control: Challenges and prospects. *Advances in parasitology*, 61, pp.567–622.

Fèvre, E.M. et al., 2006. Human African trypanosomiasis: Epidemiology and control. *Advances in parasitology*, 61, pp.167–221.

Flueckiger, R.M. et al., 2015. Integrating data and resources on neglected tropical diseases for better planning: the NTD mapping tool (NTDmap.org). *PLoS neglected tropical diseases*, 9(2), p.e0003400.

Fornace, K.M. et al., 2014. Mapping infectious disease landscapes: unmanned aerial vehicles and epidemiology. *Trends in parasitology*, 30(11), pp.514–9.

Fourcade, Y. et al., 2014. Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias J. F. Valentine, ed. *PLoS ONE*, 9(5), p.e97122.

Franco, J.R. et al., 2014. Epidemiology of human African trypanosomiasis. *Clinical epidemiology*, 6, pp.257–75.

Freeman, M. et al., 2013. Integration of water, sanitation, and hygiene for the prevention and control of neglected tropical diseases: a rationale for inter-sectoral collaboration. *PLoS Negl Trop*.

Friedl, M.. et al., 2002. Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, 83(1–2), pp.287–302.

Friedl, M.A. et al., 2010. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, 114(1), pp.168–182.

Friedl, M.A. & Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61(3), pp.399–409.

Friedman, J. et al., 2015. glmnet: Lasso and Elastic-Net Regularized Generalised Linear Models. R package version 2.0-2.

Fuller, D.O. et al., 2014. Participatory Risk Mapping of Malaria Vector Exposure in Northern South America using Environmental and Population Data. *Applied geography (Sevenoaks, England)*, 48, pp.1–7.

Gelman, A. et al., 2013. *Bayesian data analysis* 3rd ed., CPC Press.

Gilks, W.., Richardson, S. & Spiegelhalter, D., 1995. *Markov Chain Monte Carlo in Practice*, CRC Press.

Giorgi, E. et al., 2015. Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(2), pp.445–464.

Golding, N. et al., 2015. Integrating vector control across diseases. *BMC Medicine*, 13(1), p.249.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp.211–221.

Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*, Oxford University Press.

Gotway, C. & Young, L., 2002. Combining incompatible spatial data. *Journal of the American Statistical*.

Gouteux, P.J. & Artzrouni, M., 1996. Is vector control needed in the fight against sleeping sickness? A biomathematical approach. *Bulletin de la Société de pathologie exotique*, 89(4), pp.299–305.

Griffith, D.A., 2005. Effective Geographic Sample Size in the Presence of Spatial Autocorrelation. *Annals of the Association of American Geographers*, 95(4), pp.740–760.

Grimes, J.E.T. & Templeton, M.R., 2015. Geostatistical modelling of schistosomiasis prevalence. *The Lancet. Infectious diseases*, 15(8), pp.869–70.

Guerra, C.A. et al., 2007. Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. *Malaria Journal*, 6(1), p.17.

Guerrini, L. et al., 2008. Fragmentation analysis for prediction of suitable habitat for vectors: example of riverine tsetse flies in Burkina Faso. *Journal of medical entomology*, 45(6), pp.1180–6.

Guttorp, P. & Gneiting, T., 2006. Studies in the history of probability and statistics XLIX On the Matern correlation family. *Biometrika*, 93(4), pp.989–995.

Gyapong, J. & Remme, J., 2001. The use of grid sampling methodology for rapid assessment of the distribution of bancroftian filariasis. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 95(6), pp.681–686.

Gyapong, J.O. et al., 2002. The use of spatial analysis in mapping the distribution of bancroftian filariasis in four West African countries. *Annals of tropical medicine and parasitology*, 96(7), pp.695–705.

Hackett, F. et al., 2014. Incorporating scale dependence in disease burden estimates: the case of human African trypanosomiasis in Uganda. *PLoS neglected tropical diseases*, 8(2), p.e2704.

Haklay, M., 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design*, 37(4), pp.682–703.

Hamm, N.A.S., Soares Magalhães, R.J. & Clements, A.C.A., 2015. Earth Observation, Spatial Data Quality, and Neglected Tropical Diseases. *PLoS neglected tropical diseases*, 9(12), p.e0004164.

Hardy, A. et al., 2015. Mapping hotspots of malaria transmission from pre-existing hydrology, geology and geomorphology data in the pre-elimination context of Zanzibar, United Republic of Tanzania. *Parasites & Vectors*, 8(1), p.41.

Hawkins, K.R. et al., 2016. Diagnostic Tests to Support Late-Stage Control Programs for Schistosomiasis and Soil-Transmitted Helminthiases. *PLoS neglected tropical diseases*, 10(12), p.e0004985.

Hay, S.I., George, D.B., et al., 2013. Big data opportunities for global infectious disease surveillance. *PLoS medicine*, 10(4), p.e1001413.

Hay, S.I., Battle, K.E., et al., 2013. Global mapping of infectious disease. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1614), p.20120250.

Hijmans, R.J. et al., 2016. dismo: Species Distribution Modeling. R package version 1.0-15.

Hoerl, A.E. & Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), pp.55–67.

Hoeting, J.A. et al., 2006. Model selection for geostatistical models. *Ecological applications : a publication of the Ecological Society of America*, 16(1), pp.87–98.

Hotez, P.J. et al., 2016. Eliminating the Neglected Tropical Diseases: Translational Science and New Technologies S. Lustigman, ed. *PLOS Neglected Tropical Diseases*, 10(3), p.e0003895.

Hotez, P.J. et al., 2010. “Manifesto” for Advancing the Control and Elimination of Neglected Tropical Diseases. *PLoS Neglected Tropical Diseases*, 4(5), p.e718.

Hotez, P.J. & Kamath, A., 2009. Neglected tropical diseases in sub-saharan Africa: review of their prevalence, distribution, and disease burden. *PLoS neglected tropical diseases*, 3(8), p.e412.

Hürlimann, E. et al., 2011. Toward an open-access global database for mapping, control, and surveillance of neglected tropical diseases. *PLoS neglected tropical diseases*, 5(12), p.e1404.

Jacob, B.G. et al., 2013. Validation of a Remote Sensing Model to Identify Simulium damnosum s.l. Breeding Sites in Sub-Saharan Africa. A. C. A. Clements, ed. *PLoS neglected tropical diseases*, 7(7), p.e2342.

Jacobson, A. et al., 2015. A novel approach to mapping land conversion using Google Earth with an application to East Africa. *Environmental Modelling & Software*, 72, pp.1–9.

Jelinski, D. & Wu, J., 1996. The modifiable areal unit problem and implications for landscape ecology. *Landscape ecology*, 11(3), pp.129–140.

Kelly-Hope, L.A. et al., 2006. Short communication: Negative spatial association between lymphatic filariasis and malaria in West Africa. *Tropical medicine & international health : TM & IH*, 11(2), pp.129–35.

Kennedy, P.G., 2013. Clinical features, diagnosis, and treatment of human African trypanosomiasis (sleeping sickness). *The Lancet Neurology*, 12(2), pp.186–194.

King, J.D. et al., 2013. A novel electronic data collection system for large-scale surveys of neglected tropical diseases. A. M. Noor, ed. *PloS one*, 8(9), p.e74570.

Kitron, U. et al., 1996. Spatial Analysis of the Distribution of Tsetse Flies in the Lambwe Valley, Kenya, Using Landsat TM Satellite Imagery and GIS. *Journal of Animal Ecology*, 65(3), pp.371–380.

Koch, T., 2015. Mapping medical disasters: Ebola makes old lessons, new. *Disaster medicine and public health preparedness*, 9(1), pp.66–73.

Koroma, J.B. et al., 2012. Lymphatic filariasis mapping by immunochromatographic test cards and baseline microfilaria survey prior to mass drug administration in Sierra Leone. *Parasites & vectors*, 5, p.10.

Kraemer, M.U.G. et al., 2015. Progress and Challenges in Infectious Disease Cartography. *Trends in Parasitology*, 32(1), pp.19–29.

Krige, D.G., 1951. *A statistical approach to some mine valuations and allied problems at the Witwatersrand*. University of Witwatersrand.

Lai, Y.-S. et al., 2015. Spatial distribution of schistosomiasis and treatment needs in sub-Saharan Africa: a systematic review and geostatistical analysis. *The Lancet. Infectious diseases*, 15(8), pp.927–40.

Legendre, P., 1993. Spatial Autocorrelation: Trouble or New Paradigm? *Ecology*, 74(6), pp.1659–1673.

Linard, C. et al., 2014. Use of active and passive VGI data for population distribution modelling: experience from the WorldPop project. In *GIScience 2014 workshop*.

Lindgren, F. & Rue, H., 2015. Bayesian Spatial Modelling with R - INLA. *Journal of Statistical Software*, 63(19), pp.1–25.

Lindgren, F., Rue, H. & Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4), pp.423–498.

Lozano-Fuentes, S. et al., 2008. Use of Google Earth to strengthen public health capacity and facilitate management of vector-borne diseases in resource-poor environments. *Bulletin of the World Health Organization*, 86(9), pp.718–725.

Lu, D. & Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5), pp.823–870.

Lumbala, C. et al., 2015. Human African trypanosomiasis in the Democratic Republic of the Congo: disease distribution and risk. *International journal of health geographics*, 14, p.20.

Lunn, D.J. et al., 2000. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4), pp.325–337.

Lutumba, P., Matovu, E. & Boelaert, M., 2016. Human African Trypanosomiasis (HAT). In *Neglected Tropical Diseases - Sub-Saharan Africa*. pp. 63–85.

Magalhães, R.J.S. et al., 2011. The applications of model-based geostatistics in helminth epidemiology and control. *Advances in parasitology*, 74, pp.267–96.

Manyangadze, T. et al., 2015. Application of geo-spatial technology in schistosomiasis modelling in Africa: a review. *Geospatial Health*, 10(2).

Matawa, F., Murwira, K.S. & Shereni, W., 2013. Modelling the Distribution of Suitable Glossina Spp. Habitat in the North Western parts of Zimbabwe Using Remote Sensing and Climate Data. *Geoinformatics and Geostatistics: An Overview*, S1, pp.S1-S016.

Matern, B., 1960. *Spatial variation*, Stockholm.

Matheron, G., 1976. A simple substitute for conditional expectation: The disjunctive kriging. In M. Guarascio, M. David, & C. Huijbregts, eds. *Advanced Geostatistics in the Mining Industry*. Dordrecht: Springer Netherlands, pp. 221–236.

McLeod, K.S., 2000. Our sense of Snow: the myth of John Snow in medical geography. *Social Science & Medicine*, 50(7–8), pp.923–935.

Médecins Sans Frontières, 2014. *GIS support for the MSF Ebola response in Guinea in 2014: Case study*,

Merow, C., Smith, M.J. & Silander, J.A., 2013. A practical guide to MaxEnt for modeling species’ distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10), pp.1058–1069.

Milinovich, G.J., Magalhães, R.J.S. & Hu, W., 2015. Role of big data in the early detection of Ebola and other emerging infectious diseases. *The Lancet Global Health*, 3(1), pp.e20–e21.

Miller, D.L. et al., 2013. Spatial models for distance sampling data: recent developments and future directions O. Gimenez, ed. *Methods in Ecology and Evolution*, 4(11), pp.1001–1010.

Moraga, P. et al., 2015. Modelling the distribution and transmission intensity of lymphatic filariasis in sub-Saharan Africa prior to scaling up interventions: integrated use of geostatistical and mathematical modelling. *Parasites & Vectors*, 8(1), p.560.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37(1–2), pp.17–23.

Moser, W. et al., 2014. The spatial and seasonal distribution of Bulinus truncatus, Bulinus forskalii and Biomphalaria pfeifferi, the intermediate host snails of schistosomiasis, in N’Djamena, Chad. *Geospatial health*, 9(1), pp.109–18.

Mwangungulu, S.P. et al., 2016. Crowdsourcing Vector Surveillance: Using Community Knowledge and Experiences to Predict Densities and Distribution of Outdoor-Biting Mosquitoes in Rural Tanzania. *PloS one*, 11(6), p.e0156388.

Mwase, E.T. et al., 2014. Mapping the geographical distribution of lymphatic filariasis in Zambia. *PLoS neglected tropical diseases*, 8(2), p.e2714.

Mweempwa, C. et al., 2015. Impact of habitat fragmentation on tsetse populations and trypanosomosis risk in Eastern Zambia. *Parasites & vectors*, 8, p.406.

NASA, 2016a. MODIS Vegetation Indices. Available at: http://modis-land.gsfc.nasa.gov/vi.html.

NASA, 2016b. Shuttle Radar Topography Mission. Available at: http://www2.jpl.nasa.gov/srtm/.

NASA, The NASA Master Directory Held at the NASA Space Science Data Center. *2016*. Available at: http://nssdc.gsfc.nasa.gov/nmc/SpacecraftQuery.jsp [Accessed March 3, 2016].

Neis, P. & Zielstra, D., 2014. Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet*, 6(1), pp.76–106.

Neis, P., Zielstra, D. & Zipf, A., 2011. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4(4), pp.1–21.

Nutman, T.B., 2013. Lymphatic Filariasis: Progress and Challenges in the Move Toward Elimination. In I. W. Fong, ed. *Challenges in Infectious Diseases*. New York, NY: Springer New York, pp. 233–246.

O’Hanlon, S.J. et al., 2016. Model-Based Geostatistical Mapping of the Prevalence of Onchocerca volvulus in West Africa. *PLoS neglected tropical diseases*, 10(1), p.e0004328.

Odiit, M. et al., 2006. Using remote sensing and geographic information systems to identify villages at high risk for rhodesiense sleeping sickness in Uganda. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 100(4), pp.354–62.

Oluwole, A.S. et al., 2015. Bayesian geostatistical model-based estimates of soil-transmitted helminth infection in Nigeria, including annual deworming requirements. *PLoS neglected tropical diseases*, 9(4), p.e0003740.

Onapa, a W. et al., 2005. Rapid assessment of the geographical distribution of lymphatic filariasis in Uganda, by screening of schoolchildren for circulating filarial antigens. *Annals of tropical medicine and parasitology*, 99(2), pp.141–53.

Otukei, J.R. & Blaschke, T., 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *International Journal of Applied Earth Observation and Geoinformation*, 12, pp.S27–S31.

Palen, L. & Soden, R., 2015. Success & scale in a data-producing organization: The socio-technical evolution of OpenStreetMap in response to humanitarian events. *Proceedings of the 33rd …*.

Pavluck, A. et al., 2014. Electronic data capture tools for global health programs: evolution of LINKS, an Android-, web-based system. *PLoS neglected tropical diseases*, 8(4), p.e2654.

Pedersen, U.B. et al., 2014. Modelling spatial distribution of snails transmitting parasitic worms with importance to human and animal health and analysis of distributional changes in relation to climate. *Geospatial health*, 8(2), pp.335–43.

Peyrard, N. et al., 2013. Model-based adaptive spatial sampling for occurrence map construction. *Statistics and Computing*, 23(1), pp.29–42.

Phillips, S.J. et al., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1), pp.181–197.

Phillips, S.J., Anderson, R.P. & Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3–4), pp.231–259.

Phillips, S.J. & Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2), pp.161–175.

Pullan, R.L. et al., 2012. Spatial parasite ecology and epidemiology: a review of methods and applications. *Parasitology*, 139(14), pp.1870–1887.

Raj, R., Hamm, N. & Kant, Y., 2013. Analysing the effect of different aggregation approaches on remotely sensed data. *International journal of remote sensing*.

Ramaiah, K.D. & Ottesen, E.A., 2014. Progress and impact of 13 years of the global programme to eliminate lymphatic filariasis on reducing the burden of filarial disease. *PLoS neglected tropical diseases*, 8(11), p.e3319.

Rayaisse, J.-B. et al., 2015. Baited-boats: an innovative way to control riverine tsetse, vectors of sleeping sickness in West Africa. *Parasites & vectors*, 8, p.236.

Renner, I.W. & Warton, D.I., 2013. Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics*, 69(1), pp.274–281.

Robinson, T.P., 2000. Spatial statistics and geographical information systems in epidemiology and public health. *Advances in Parasitology*, 47, pp.81–128.

Rogers, D.J., 2006. Models for vectors and vector-borne diseases. *Advances in parasitology*, 62, pp.1–35.

Rogers, D.J., 2000. Satellites, Space, Time and the African Trypanosomiasis. *Advances in Parasitology*, 47, pp.129–171.

Rue, H., Martino, S. & Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), pp.319–392.

Schapire, R.E., 2003. The Boosting Approach to Machine Learning An Overview. *Nonlinear Estimation and Classification*.

Schur, N. et al., 2011. Geostatistical model-based estimates of Schistosomiasis prevalence among individuals aged ≤ 20 years in West Africa. *PLoS neglected tropical diseases*, 5(6), p.e1194.

Schur, N. et al., 2013. Spatially explicit Schistosoma infection risk in eastern Africa using Bayesian geostatistical modelling. *Acta Tropica*, 128(2), pp.365–377.

Sciarretta, A. et al., 2005. Development of an adaptive tsetse population management scheme for the Luke community, Ethiopia. *Journal of medical entomology*, 42(6), pp.1006–19.

Sciarretta, A. et al., 2010. Spatial clustering and associations of two savannah tsetse species, Glossina morsitans submorsitans and Glossina pallidipes (Diptera: Glossinidae), for guiding interventions in an adaptive cattle health management framework. *Bulletin of entomological research*, 100(6), pp.661–70.

Shaw, A.P.M. et al., 2015. Costs of using “tiny targets” to control Glossina fuscipes fuscipes, a vector of gambiense sleeping sickness in Arua District of Uganda. *PLoS neglected tropical diseases*, 9(3), p.e0003624.

Siegfried, G. & Siegfried, G., 2014. Adaptive and Spatial Sampling Designs. In *Wiley StatsRef: Statistics Reference Online*. Chichester, UK: John Wiley & Sons, Ltd.

Simarro, P.P. et al., 2015. Monitoring the Progress towards the Elimination of Gambiense Human African Trypanosomiasis E. Matovu, ed. *PLOS Neglected Tropical Diseases*, 9(6), p.e0003785.

Simarro, P.P. et al., 2011. Risk for Human African Trypanosomiasis, Central Africa, 2000–2009. *Emerging Infectious Diseases*, 17(12), pp.2322–2324.

Simarro, P.P. et al., 2010. The Atlas of human African trypanosomiasis: a contribution to global mapping of neglected tropical diseases. *International journal of health geographics*, 9(1), p.57.

Sime, H. et al., 2014. Integrated mapping of lymphatic filariasis and podoconiosis: lessons learnt from Ethiopia. *Parasites & vectors*, 7(1), p.397.

Simonsen, P.E. & Mwakitalu, M.E., 2013. Urban lymphatic filariasis. *Parasitology research*, 112(1), pp.35–44.

Simoonga, C. et al., 2009. Remote sensing, geographical information system and spatial analysis for schistosomiasis epidemiology and ecology in Africa. *Parasitology*, 136(13), pp.1683–93.

Sinka, M.E. et al., 2012. A global map of dominant malaria vectors. *Parasites & Vectors*, 5(1), p.69.

Sinka, M.E., Bangs, M.J., Manguin, S., Coetzee, M., Mbogo, C.M., et al., 2010. The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasites & Vectors*, 3(1), p.117.

Sinka, M.E., Bangs, M.J., Manguin, S., Coetzee, M., Mbogo, C.M., et al., 2010. The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasites & vectors*, 3(1), p.117.

Slater, H. & Michael, E., 2013. Mapping, bayesian geostatistical analysis and spatial prediction of lymphatic filariasis prevalence in Africa. *PloS one*, 8(8), p.e71574.

Slater, H. & Michael, E., 2012. Predicting the current and future potential distributions of lymphatic filariasis in Africa using maximum entropy ecological niche modelling. *PloS one*, 7(2), p.e32202.

de Smith, M.J., Goodchild, M.F. & Longley, P.A., 2015. *Geospatial analysis* 5th ed.,

Sodahlon, Y., Malecela, M. & Gyapong, J.O., 2016. Lymphatic Filariasis (Elephantiasis). In *Neglected Tropical Diseases - Sub-Saharan Africa*. pp. 159–186.

Soden, R. & Palen, L., 2014. From crowdsourced mapping to community mapping: The post-earthquake work of openstreetmap haiti. *COOP 2014-Proceedings of the 11th International …*.

Sokolow, S.H. et al., 2016. Global Assessment of Schistosomiasis Control Over the Past Century Shows Targeting the Snail Intermediate Host Works Best. *PLoS neglected tropical diseases*, 10(7), p.e0004794.

Solano, P., Torr, S.J. & Lehane, M.J., 2013. Is vector control needed to eliminate gambiense human African trypanosomiasis? *Frontiers in cellular and infection microbiology*, 3, p.33.

Sørensen, R., Zinko, U. & Seibert, J., 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. *Hydrology and Earth System Sciences Discussions European Geosciences Union*, 10(1), pp.101–112.

Sorichetta, A. et al., 2015. High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific data*, 2, p.150045.

Sousa-Figueiredo, J.C. et al., 2015. Mapping of schistosomiasis and soil-transmitted helminths in Namibia: the first largescale protocol to formally include rapid diagnostic tests. *PLoS neglected tropical diseases*, 9(7), p.e0003831.

Spiegelhalter, D.J. et al., 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), pp.583–639.

Stanton, M.C. et al., 2013. Baseline drivers of lymphatic filariasis in Burkina Faso. *Geospatial health*, 8(1), pp.159–73.

Stanton, M.C., Yamauchi, M., et al., 2016. Measuring the physical and economic impact of filarial lymphoedema in Chikwawa district, Malawi: A case-control study. *Under submission*.

Stanton, M.C., Molineux, A., et al., 2016. Mobile Technology for Empowering Health Workers in Underserved Communities: New Approaches to Facilitate the Elimination of Neglected Tropical Diseases. *JMIR Public Health and Surveillance*, 2(1), p.e2.

Steinmann, P. et al., 2006. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *The Lancet Infectious Diseases*, 6(7), pp.411–425.

Stensgaard, A.-S. et al., 2011. Bayesian geostatistical modelling of malaria and lymphatic filariasis infections in Uganda: predictors of risk and geographical patterns of co-endemicity. *Malaria journal*, 10, p.298.

Stensgaard, A.-S. et al., 2013. Large-scale determinants of intestinal schistosomiasis and intermediate host snail distribution across Africa: does climate matter? *Acta tropica*, 128(2), pp.378–90.

Stevens, K.B. & Pfeiffer, D.U., 2015. Sources of spatial animal and human health data: Casting the net wide to deal more effectively with increasingly complex disease problems. *Spatial and Spatio-temporal Epidemiology*, 13, pp.15–29.

Stevens, K.B. & Pfeiffer, D.U., 2011. Spatial modelling of disease using data- and knowledge-driven approaches. *Spatial and Spatio-temporal Epidemiology*, 2(3), pp.125–133.

Stothard, J.R. et al., 2014. Diagnostics for schistosomiasis in Africa and Arabia: a review of present options in control and future needs for elimination. *Parasitology*, 141(14), pp.1947–1961.

Stothard, J.R. et al., 2011. Investigating the spatial micro-epidemiology of diseases within a point-prevalence sample: a field applicable method for rapid mapping of households using low-cost GPS-dataloggers. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 105(9), pp.500–6.

Stothard, R.J. et al., 2017. Towards interruption of schistosomiasis transmission in sub-Saharan Africa: developing an appropriate environmental surveillance framework to guide and to support “end game” interventions. *Infectious Diseases of Poverty*, 6(10).

Sturrock, H.J.W. et al., 2010. Optimal survey designs for targeting chemotherapy against soil-transmitted helminths: effect of spatial heterogeneity and cost-efficiency of sampling. *The American journal of tropical medicine and hygiene*, 82(6), pp.1079–87.

Sturrock, H.J.W. et al., 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(3), pp.165–75.

Sui, D., Elwood, S. & Goodchild, M., 2012. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, Springer Science & Business Media.

Teng, J.E. et al., 2014. Using Mobile Health (mHealth) and geospatial mapping technology in a mass campaign for reactive oral cholera vaccination in rural Haiti. J. Clemens, ed. *PLoS neglected tropical diseases*, 8(7), p.e3050.

The Flowminder Foundation, Flowminder. Available at: http://www.flowminder.org/.

Thomson, M.C. et al., 1999. Predicting malaria infection in Gambian children from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results. *The American journal of tropical medicine and hygiene*, 61(1), pp.2–8.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), pp.273–282.

Tirados, I. et al., 2015. Tsetse Control and Gambian Sleeping Sickness; Implications for Control Strategy. *PLoS neglected tropical diseases*, 9(8), p.e0003822.

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

Tom-Aba, D. et al., 2015. Innovative Technological Approach to Ebola Virus Disease Outbreak Response in Nigeria Using the Open Data Kit and Form Hub Technology D. M. Harper, ed. *PLOS ONE*, 10(6), p.e0131000.

Townshend, J. et al., 1991. Global land cover classification by remote sensing: present capabilities and future possibilities. *Remote Sensing of Environment*, 35(2–3), pp.243–255.

Verity, R. et al., 2014. Spatial targeting of infectious disease control: identifying multiple, unknown sources D. Warton, ed. *Methods in Ecology and Evolution*, 5(7), pp.647–655.

Walz, Y. et al., 2015. Risk profiling of schistosomiasis using remote sensing: approaches, challenges and outlook. *Parasites & vectors*, 8(1), p.163.

Wang, J.-F. et al., 2012. A review of spatial sampling. *Spatial Statistics*, 2, pp.1–14.

Wardrop, N.A. et al., 2010. Bayesian geostatistical analysis and prediction of Rhodesian human African trypanosomiasis. *PLoS neglected tropical diseases*, 4(12), p.e914.

Wardrop, N.A. et al., 2014. Interpreting predictive maps of disease: highlighting the pitfalls of distribution models in epidemiology. *Geospatial Health*, 9(1), p.237.

Warner, J., 1996. The case books of Dr. John Snow Medical History. *Medical history*.

Wesolowski, A. et al., 2015. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences of the United States of America*, 112(38), pp.11887–92.

Wesolowski, A. et al., 2014. Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. *Scientific reports*, 4, p.5678.

Woodcock, C. & Allen, R., 2008. Free access to Landsat imagery. *Science (New York, …*.

World Health Organization, 2012. *Accelerating work to overcome the global impact of neglected tropical diseases: A roadmap for implementation*,

World Health Organization, 2013a. *Control and surveillance of human African trypanosomiasis: Report of a WHO Expert Committee, WHO technical report series 984*,

World Health Organization, 2013b. Control and surveillance of human African trypanosomiasis. *World Health Organization technical report series*, (984), pp.1–237.

World Health Organization, 2013c. *Lymphatic filariasis: Managing Morbidity and Preventing Disability*,

World Health Organization, 2000. *Operational guidelines for rapid mapping of Bancroftian filariasis in Africa*, Geneva.

World Health Organization, 2006. *Preventive chemotherapy in human helminthiasis*, Geneva.

World Health Organization, 2013d. *Report of an informal consultation on schistosomiasis control*,

World Health Organization, 2013e. *Sustaining the drive to overcome the global impact of neglected tropical diseases: Second WHO report on neglected tropical diseases*, World Health Organization.

World Health Organization, 2016. Trypanosomiais, human African (Sleeping sickness) Fact sheet. Available at: http://www.who.int/mediacentre/factsheets/fs259/en/.

World Health Organization, 2015. *Water sanitation and hygiene for accelerating and sustaining progress on neglected tropical diseases: a global strategy 2015-2020*,

World Health Organization, 2011. WHO position statement on integrated vector management to control malaria and lymphatic filariasis. *Weekly Epidemiological Record*, 13, pp.121–127.

World Health Organization, 2010. Working to overcome the global impact of neglected tropical disease: First WHO report on neglected tropical diseases. Available at: http://whqlibdoc.who.int/publications/2010/9789241564090\_eng.pdf?ua=1 [Accessed June 5, 2014].

World Health Organization & Global Programme to Eliminate Lymphatic Filariasis, 2011. *Monitoring and epidemiological assessment of mass drug administration*,

WorldClim, 2016. WorldClim - Global Climate Data. Available at: http://www.worldclim.org/.

Zook, M. et al., 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*, 2(2), pp.6–32.

Zouré, H.G.M. et al., 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(6), p.e1210.

**Tables**

Table 1: Environmental influences on the spatial distribution of human African trypanosomiasis, schistosomiasis, lymphatic filariasis in sub-Saharan Africa (SSA).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Disease | Parasite | Vector/intermediate host | SSA burden | Primary methods of control | Environmental influences | Key environmental references |
| Human African trypanosomiasis (HAT) | Trypanosoma brucei gambiense, Trypanosoma brucei rhodesiense | Tsetse flies (Glossina spp.) | 3796 cases reported in 2014 | Case detection, vector control and control of animal reservoirs | Presence of rivers, vegetation, land cover, temperature, elevation, rainfall, proximity of livestock (Rhodesian HAT only) | (Rogers 2000; Sciarretta et al. 2005; Batchelor et al. 2009; Wardrop et al. 2010; Matawa et al. 2013; Franco et al. 2014; Dicko et al. 2014) |
| **Schistosomiasis** | *Schistosoma haematobium, Schistosoma mansoni* | Freshwater snails (*Biomphalaria, Bulinus*) | ~192 million cases, estimated in 2003 | Preventive chemotherapy | Presence of water bodies, vegetation, land cover, temperature, elevation, rainfall, soil properties, water properties, water accessibility | (Lai et al. 2015; Manyangadze et al. 2015) |
| **Lymphatic filariasis** | *Wuchereria bancrofti* | *Anopheles* and *Culex* mosquitoes | ~40 million cases, estimated in 2013 | Preventive chemotherapy | Presence of water bodies (transient and permanent), vegetation, land cover, temperature, elevation, rainfall, humidity | (Slater & Michael 2013; Cano et al. 2014; Moraga et al. 2015; Simonsen & Mwakitalu 2013) |

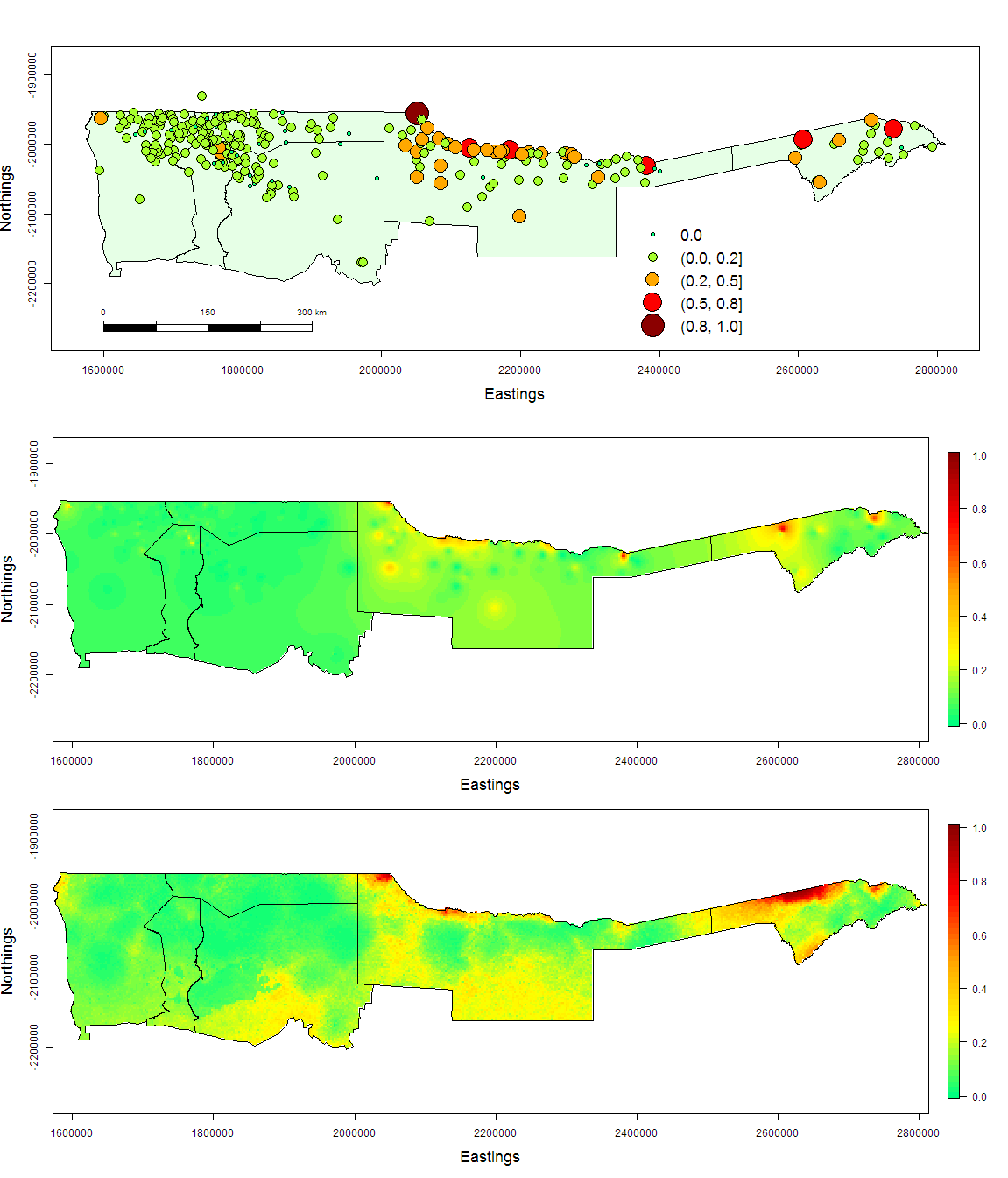
Table 2: Resources for spatially-referenced data relating to the risk of human African trypanosomiasis, schistosomiasis, lymphatic filariasis. Note that this list is not designed to be exhaustive, and other resources are available.

|  |  |  |  |
| --- | --- | --- | --- |
| Resource | Description | Data availability | References |
|  |  |  |  |
| Disease case data |  |  |  |
| Global Atlas of Helminth Infections | Point prevalence survey data for **schistosomiasis**, **LF** and soil-transmitted helminths | All point prevalence maps, plus the raw data for a subset of countries are available to download | <http://www.thiswormyworld.org/>  (Cano et al. 2014) |
| Global NTD database | Point prevalence survey data for **schistosomiasis** | All point prevalence data available. Registration required | <http://www.gntd.org/>  (Hürlimann et al. 2011) |
| Atlas of Human African Trypanosomiasis | Number of HAT cases since 2000 by village | All case maps are available, and case data is available on request | <http://www.who.int/trypanosomiasis_african/country/foci_AFRO/>  (Simarro et al. 2010) |
| NTD Mapper | Point prevalence survey data for **schistosomiasis, LF**, soil-transmitted helminths, onchocerciasis, Loa loa and trachoma | Schistosomiasis, LF, soil-transmitted helminths and trachoma prevalence data are available to download | <http://www.ntdmap.org/>  (Flueckiger et al. 2015) |
| Healthmap | Various diseases | Locations of reported cases, and the source of the report are available to visualise online | http://www.healthmap.org/ |
|  |  |  |  |
| Vector data |  |  |  |
| Atlas of tsetse and African animal trypanosomiasis | Point-level tsetse count data | Maps and raw data currently unavailable, although preliminary maps found in (Cecchi et al. 2015) | (Cecchi et al. 2014) |
| Malaria Atlas Project | Predicted probability of occurrence of Anopheles mosquitoes as a 5 x 5 km resolution | Maps plus raster data are available to download | <http://www.map.ox.ac.uk/>  (Sinka, Bangs, Manguin, Coetzee, Arbaji, et al. 2010) |
|  |  |  |  |
| Remotely sensed data |  |  |  |
| Landsat | Landsat data is available at 30m resolution and includes 11 spectral bands include the visible spectrum, infrared, near infrared, short-wave infrared and thermal infrared.  The data collection period is 1972 – present. | The raw data are available to download at 30m resolution using the USGS Earth Explorer. Landsat data has also been used to create a number of landcover products including GlobeLand30 | <http://landsat.usgs.gov/>  <http://earthexplorer.usgs.gov/> (download raw data)<http://www.globallandcover.com/> (Globeland30) |
| MODIS | MODIS (Moderate Resolution Imaging Spectroradiometer) is available at 250m resolution (2 bands), 500m (5 bands) and 1km (29 bands). The data collection period is 1999 – present. | The raw data are available at their respective resolutions from LAADS Web. Real-time data can also be viewed using World View. Numerous products derived from MODIS data are also available to download, including MODIS land cover and NDVI. | <https://ladsweb.nascom.nasa.gov/data> (download raw data)  <https://worldview.earthdata.nasa.gov/> (view real-time data)  <http://modis.gsfc.nasa.gov/data/dataprod/> (list of products) |
| AVHRR | AVHRR (Advanced Very High Resolution Radiometer) data is available at 1.1km resolution and currently includes 5 spectral bands in the red, near infrared and thermal infrared portions of the electromagnetic spectrum. The data collection period is 1979 – present. | The raw data, plus products derived from the raw data such as NDVI are available from Earth Explorer | <http://earthexplorer.usgs.gov/> (download raw data and NDVI products) |
| ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) | ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) data is available at 15m-90m resolution and includes 14 spectral bands in the visible, near infrared, shortwave infrared and thermal infrared portions of the electromagnetic spectrum. The data collection period is February 2000 - present | The raw data at are currently available to download from multiple sources including NASA Reverb | <https://asterweb.jpl.nasa.gov/>  <http://reverb.echo.nasa.gov/reverb> (download data) |
| SPOT | SPOT (Satellite Pour l’Observation de la Terre) is a commercial satellite with a resolution of 6m, measuring 4 spectral bands (visible spectrum and near infrared) | A small subset of data is available to download for free from the European Space Agency. Registration is required. | https://earth.esa.int/web/guest/data-access |
| RapidEye | RapidEye is a commercial satellite cluster with a resolution of 5m, measuring 5 spectral bands (visible spectrum, red edge and near infrared) | A small subset of data is available to download for free from the European Space Agency. Registration is required. | https://earth.esa.int/web/guest/data-access |
| Open Aerial Map | Open Aerial Map is a set of tools for sharing and using openly licenced satellite and unmanned aerial vehicle imagery. | Data availability is currently limited, with low geographical coverage. | openaerialmap.org |
| Google Earth | Google Earth displays imagery collated from a number of sources, particularly commercial satellite vendors. The majority of the globe is covered in at least a spatial resolution of 15m, with some areas having a resolution of 0.15m. | Imagery accessed using Google Earth can only be used for visualisation purposes. | https://www.google.com/earth/ |
|  |  |  |  |
| **Other data** |  |  |  |
| OpenStreetMap | OpenStreetMap (OSM) is a community-driven mapping project, which provides open data on landscape features such as transport routes (e.g. roads, footpaths, railways), water features (e.g. rivers, lakes), buildings, amenities, land use categories and administrative boundaries. | All data within OSM is available to download using a number of ways, including the OSM website download area. Planet.osm can be used to download the whole dataset and is updated once a week. | <http://www.openstreetmap.org/>  <http://planet.osm.org/> (download entire dataset) |
| WorldPop | WorldPop provides 100m resolution data on human population distributions, modelled from census, survey, satellite and GIS datasets, including OSM. | Estimated population data are available at 1km for Africa, Asia and South and Central America, whereas 100m data is available for a subset of countries within these continents. | <http://www.worldpop.org.uk/> |

**Figures**

C:\Users\michelle.stanton\AppData\Local\Microsoft\Windows\INetCacheContent.Word\Figure1_variogram.tif

**Figure 1**: A depiction of the theoretical variogram with an exponential correlation function (, with (nugget) and , and a practical range of 1.2 (left), and the empirical variogram of logit-transformed schistosomiasis prevalence data obtained for 299 schools in Namibia (Sousa-Figueiredo et al. 2015), indicating the presence of spatial correlation up to distances of 300km. See Figure 2 for a map of surveyed prevalence.



**Figure 2**: Surveyed schistosomiasis prevalence in Namibia (top), plus predictive map of schistosomiasis prevalence using inverse-distance weights (middle) and the posterior mean schistosomiasis prevalence obtained from fitting a binomial geostatistical model to the prevalence data, including average maximum normalised difference vegetation index (NDVI), average annual rainfall and topographical wetness index as covariates (bottom).

C:\Users\michelle.stanton\AppData\Local\Microsoft\Windows\INetCacheContent.Word\Figure3_mapcomparison.tif

**Figure 3**: A comparison of the level of detail provided by Google Maps (left) and OpenStreetMap (right) for Monrovia, Liberia