Interpreting predictive maps of disease, highlighting the pitfalls of species distribution models in epidemiology

Nicola A Wardrop[1], Matthew Geary[1,2], Patrick E Osborne[3], Peter M Atkinson[1].

[1]Geography and Environment, Faculty of Social and Human Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

[2]Current address: Department of Biological Sciences, University of Chester, Chester, CH1 4BJ, UK

[3]Centre for Environmental Sciences, Faculty of Engineering and the Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK

Corresponding author

Dr Nicola Wardrop
Geography and Environment
Faculty of Social and Human Sciences
University of Southampton
Highfield, Southampton
SO17 1BJ, UK

Tel.: +44 (0)2380 594 612

Email: Nicola.Wardrop@soton.ac.uk.

Abstract

The application of spatial modelling to epidemiology has increased significantly over the past decade, delivering enhanced understanding of the environmental and climatic factors affecting disease distributions and providing spatially continuous representations of disease risk (predictive maps). These outputs provide significant information for disease control programmes, allowing spatial targeting and tailored interventions. However, several factors (e.g. sampling protocols or temporal disease spread) can influence predictive mapping outputs. This paper proposes a conceptual framework which defines several scenarios and their potential impact on resulting predictive outputs, using simulated data to provide an exemplar. It is vital that researchers recognise these scenarios and their influence on predictive models and their outputs, as a failure to do so may lead to inaccurate interpretation of predictive maps. As long as these considerations are kept in mind, predictive mapping will continue to contribute significantly to epidemiological research and disease control planning.

Keywords: spatial epidemiology; predictive modelling; species distribution modelling.

Introduction

In recent years, there has been a significant increase in the application of spatial modelling tools to disease studies. This has been driven by the increasing availability of epidemiological, environmental and climatic datasets with spatial (and temporal) dimensions, increased computational capacity, the development of geographical information systems (GIS) and a growing number of spatial analytical tools and platforms capable of handling spatial and space-time datasets. Traditional, non-spatial methods of epidemiological analysis can fail to adequately address major determinants of disease risk. The spatial distributions of many diseases are linked explicitly to environmental conditions (e.g. climatic factors or land cover) and these relationships are most effectively explored, quantified and utilised via spatial visualisation and analysis (Bergquist, 2001). The increasing application of spatial analysis is not unique to epidemiology; there is a close parallel in biodiversity studies, where species distribution modelling (SDM) has proliferated (Elith and Leathwick, 2009). Pathogens can be considered in this context: the tools and theories developed in SDM have useful applications in epidemiological research and *vice versa*.

The cartographic representation of epidemiological data has many benefits over presentation using tables or plots; images are attention-grabbing, of more interest and allow immediate visual interpretation of spatial patterns (Koch, 2005). Detailed information on the spatial distribution of diseases also provides significant benefits for disease control programmes, particularly for spatially heterogeneous disease distributions (Snow et al., 1996; Simarro et al., 2010). However, just as in the mapping of biodiversity, obtaining comprehensive spatial coverage of a disease within a region of interest is not always possible using disease surveilance data (particularly in developing countries here the infrastructure is often poor). Additionally, the large-scale surveys required to provide complete information are commonly impractical due to financial constraints, logistical issues, security needs and time limitations (Snow et al. 1996; Brooker et al. 2000). These limitations may be overcome, at least in part, using predictive modelling, as described below.

Statistical methods can be used to fit regression models of the relationship between disease and environment; thus, quantifying the effects of covariates (i.e. variables representing environmental, climatic or landscape factors) on epidemiological measures of disease such as occurrence (presence/absence), prevalence or incidence rates. Models based on covariates, which are measured at the same locations for which epidemiological information is available, but where precise geographical coordinates are absent, and their spatial relationships to one another are not accounted for, focus on environmental space (Elith and Leathwick, 2009). Where covariate information is available covering the full area of interest (e.g. as a raster), these models can be interpolated or extrapolated (prediction within or beyond the range of the training data, respectively) over continuous space; hence, predicting disease at locations for which observed data are not available (Elith and Leathwick, 2009). Prediction with respect to new sites is based on the disease's location in environmental space. These types of models provide information regarding factors driving the observed spatial distribution of disease. The resulting output is a predictive map, also known as a "risk map" (Brooker, 2007), which are is widely used (without incorporating the geographical coordinates) in biodiversity studies (Austin, 2002; Elith and Leathwick, 2009). It can be argued that such models are capable of producing predictive (risk) maps because the main processes determining occurrences are aspatial: it is assumed that species do not respond to location *per se*.

One potential problem with the approach discussed above is the inability to account for spatial autocorrelation in the residuals (where values close together in space are more similar than values further apart, which occurs commonly when studying the distributions of infectious diseases). This can (i) violate the underlying assumptions of the statistical methods used; and (ii) result in inaccurate models, biased regression parameters, underestimated standard errors, falsely narrow confidence intervals and an overestimation of the significance of covariates, ultimately leading to misinterpretation of the relationships between observations and covariates (Legendre, 1993;

Thomson et al., 1999). In practice, the effect of spatial autocorrelation on prediction accuracy varies among modelling techniques and represents one source of uncertainty in SDM (Marmion et al., 2009). However, extension of traditional modelling methods allows the explicit inclusion of spatial information in the modelling process, e.g., the inclusion of both environmental and geographic space in the model. Such extension deals appropriately with the potential problem above. One potential solution involves inclusion of geostatistical spatial prediction of the residuals in a mixed regression model (Diggle and Ribeiro Jr, 2007). Geostatistical methods incorporate information on the precise location of each observation in relation to other observations to represent spatial autocorrelation, giving increased accuracy of estimates of covariate effects, measures of uncertainty and predictive outputs (Diggle et al., 2002).

Predictive mapping of disease (or species distributions more generally) can help overcome the problems associated with sparse datasets. Data from a sample of locations (surveys or surveillance) can be used to fit a model, and subsequent interpolation or extrapolation can provide a spatially continuous prediction of disease (Brooker, 2007), alleviating the need for comprehensive and large-scale surveys. These outputs can allow the consideration of spatial heterogeneity in disease distributions during planning, implementation and monitoring of interventions, including targeting interventions to areas with the greatest predicted risk of disease (Clements et al., 2006), identification of areas with a low risk of disease (which can be considered low priority for intervention) (Clements et al., 2010) and recognition of areas in which intervention may be detrimental (Diggle et al., 2007). The consideration of uncertainty in outputs allows the delineation of areas from which additional information is required; thus, allowing targeted data acquisition (Clements et al., 2006).

The integration of predictive maps and population distribution data allows the estimation of populations at risk of disease and disease burden, providing information to support the allocation of resources (e.g., delivery of adequate supplies of drugs) as described by Gething et al. (2011). The types of outputs described above can also provide valuable resources for advocacy purposes, aiding communication to Government bodies, international organisations and the general public. Additional benefits from predictive mapping include enhanced understanding of the ecology of disease transmission, identification of landscape risk factors and the implication of environmental factors in the spread or distribution of disease (Wardrop et al., 2010), each of which can allow the development of tailored interventions for specific epidemiological settings.

The underlying theoretical basis for SDM and predictive mapping is ecological niche theory, particularly Hutchison's model (Austin, 2002). Hutchison (1959) envisaged the niche as a hyper-volume in multi-dimensional space (each axis being an environmental characteristic) that defines the conditions, under which a population can maintain a positive net growth rate (Pearman et al., 2008). The fundamental niche (constrained by genetics and physiology) is defined as distinct from the realised niche (with limitations on resource-use caused by competing species): the realised niche usually seen as a subset of the fundamental niche (see Pulliam, 2000 for exceptions). Vector-borne diseases are interesting in this context since modelling may focus on the vector, the host(s) and/or the disease itself. Furthermore, the vector and host(s) are essentially part of the niche of the disease and, indeed, may control its survival to such an extent that they act as the full niche in certain parts of the life cycle.

When predictive models are extrapolated (and to some extent interpolated) to new locations (and time periods), two ecological assumptions are necessary: (a) the species is in equilibrium with the environment in the area used to train the model; and (b) the niche is conserved across space and time, i.e. the species-environment relationship is spatially homogeneous (Broennimann and Guisan, 2008; Nogues-Bravo, 2009). Assumption (a) is violated when ranges are expanding (Elith et al., 2010) or where parts of a range are unoccupied by the species (e.g., due to chance or human

intervention), but may otherwise hold. There is considerable uncertainty over the applicability of assumption (b) and, indeed, whether it is the realised niche, the fundamental niche, or both that might vary between areas (Pearman et al., 2008). Additionally, careful consideration should be given to the observed epidemiological data and covariate data used in the modelling process. To illustrate how these theoretical underpinnings affect disease modelling in conjunction with the limitations imposed by incomplete or unrepresentative sampling, we applied predictive modelling methods to a simulated dataset under four scenarios.

Materials and methods

Study area and data

A hypothetical disease was simulated across an area of East Africa (between latitude 27° and 5° and longitude 22° and 42°; Figure 1). This choice was arbitrary and the disease simulated is not meant to represent any particular existing disease. Environmental data for the disease distribution simulations were downloaded from Worldclim as raster layers at the spatial resolution of 10' and cropped to the study area (Hijmans et al., 2005). Data for mean monthly temperature and mean monthly precipitation were converted to annual averages. Altitude and mean temperature of the wettest quarter were also used in the modelling.

The disease was simulated to occur in areas with a mean annual temperature between 18.0 and 22.5ºC and mean annual precipitation between 60 and 170 mm but was not constrained by altitude. As a result of these choices, approximately one quarter of the study area was classified as suitable for disease transmission (26.4 %; Figure 2).

Disease scenario sampling

The four scenarios described in Table 1 were investigated using the hypothetical disease described above. Sampling for each of the disease scenarios was performed using the randomPoints function from the dismo package (Hijmans et al., 2013). In each scenario, 300 presence or absence locations were extracted from a true suitability raster and used for model fitting (see Figure 3). In scenarios a) and c) (full information and missing covariates scenarios, respectively), these points were distributed completely randomly across the study area. For scenario b) (heterogeneous sampling effort) these locations were biased towards Kenya (200 locations) rather than the remaining study area (100 locations). For scenario d) (disease not in equilibrium) the presence or absence values for the locations were manipulated so that the disease was recorded normally in Kenya (present/absent), hile all of the locations in the remaining study area were recorded as absent: this could represent a situation where the disease is not occupying its full niche due to chance or human intervention.

Model fitting and testing

Generalised linear models were fitted to the observed (presence/absence) data from each of the four scenarios: environmental data were extracted for the sample data locations, and logistic regression analysis was applied to quantify the relations between disease presence and the covariates. In each scenario, mean annual temperature, mean annual precipitation and altitude were strongly correlated with one another (Pearson's $c > 0.5$). To avoid problems associated with collinearity, only mean annual precipitation and altitude were included in candidate models for scenarios a), b) and d), and only altitude and mean temperature of the wettest quarter for scenario c). To make meaningful comparisons across scenarios we chose to fit the same model (or its equivalent in the missing covariates scenario) in each case. Based on prior knowledge of the disease distribution, we included an interaction term between altitude and mean annual precipitation (or mean temperature of the wettest quarter for scenario c). For each scenario, 100 simulated sets of sample data were used in the epidemiological distribution models.

Models were tested using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve (Fielding and Bell, 1997), where a threshold probability of occurrence of 0.5 was used to classify predicted disease presence (or suitability). AUC scores range between 0 and 1: those greater than 0.5 are considered to have predictive ability better than random (for predicting presence), while scores above 0.7 indicate a good predictive ability. ROC plots were constructed and AUC values were calculated using the ROCR package (Sing et al., 2005). Along with the AUC score we assessed the predicted binary distribution (predicted presence, based on a threshold probability of 0.5) from each modelling scenario against the true suitability and calculated the proportion of the study area predicted correctly. These testing metrics were calculated for each scenario over 100 simulations to obtain a full picture of the variability in predictions for each scenario.

All modelling was performed in R (R Development Core Team, 2013). Spatial functions from the 'raster' (Hijmans, 2013), 'rgdal' (Bivand et al., 2013), 'sp' (Pebesma and Bivand, 2005; Bivand et al., 2008) and 'maptools' (Bivand and Lewin-Koh, 2013) packages were also used during the model simulations.

Results

Spatial predictions

The models differed with respect to the spatial predictions across the study area (Figure 4) that can be interpreted as predicted probability of occurrence, or predicted suitability for disease. The full information model (scenario a)) predicted an area which broadly matched the actual spatial distribution. However, the predicted area of suitability was slightly larger, particularly in the South of the study area. The missing covariates model (scenario c)) also predicted an area of similar pattern to the simulated disease. However, in this case the area of predicted suitability was broader still and included a patch in the South-west of the study area, which was unsuitable for the disease. The heterogeneous sampling effort model (scenario b)) predicted inaccurately overall with areas on the edges of the study area, outside of the range of the simulated disease, predicted to be suitable. The disease 'not in equilibrium' model (scenario d)) predicted almost all of the study area to be unsuitable. Some small pockets were predicted to be suitable in the north of the region. However, the majority of these pockets were outside the distribution of the simulated disease.

Model testing

Figure 5 shows the ROC curves from each of the scenarios, and Figure 6 shows the proportion of the study area that was correctly predicted. The full information model (scenario a)) produced the highest median scores for both AUC (0.77) and the proportion of the study area predicted correctly (0.71). These scores suggest the model has good predictive power. The missing covariates model (scenario c)) was closest to the full information model in terms of performance (median AUC = 0.71; median proportion of the study area predicted correctly = 0.68). The AUC scores for both the disease 'not in equilibrium' model (scenario d)) and the heterogeneous sampling effort model (scenario b)) suggest that they perform no better than random in terms of prediction. The disease 'not in equilibrium' model performed more accurately in terms of correct prediction of the study area (median = 0.62) than AUC score (median = 0.5). The heterogeneous sampling effort performed less accurately than the other scenarios for both metrics (median AUC = 0.45; median proportion of study area predicted correctly = 0.53).

Overall, the full information scenario (scenario a)) performed the most accurately in terms of both the proportion of the study area predicted correctly and the AUC score, followed by the missing covariates scenario (scenario c)). The least accurate model was the scenario representing heterogeneous sampling effort (scenario b)) which, along with the disease spreading/control programme scenario, failed to predict disease suitability in the majority of the study area.

Discussion

As discussed above, statistical models are now used widely to map disease, supplementing traditional epidemiological methodologies leading to enhanced characterisation and understanding of disease distributions and epidemiology. The four scenarios presented above highlight the dependence of predictive mapping outputs on (i) sampling and data considerations; and (ii) contextual factors, such as temporal disease spread in the study area. It is vital that researchers recognise these factors and their influence on predictive models and their outputs as inadvertent use of incomplete or biased data and thethe use of inadequate covariates may lead to inaccurate interpretation of predictive maps. In addition, absence of consideration for the on-going dynamics of disease transmission and spread within the study area can easily result in erroneous guidance.

Scenario a), which represents the ideal situation where the disease is in equilibrium, representative samples are available and appropriate covariates are being used, is the ideal situation for predictive modelling of disease, although it is likely that many practical examples do not fulfil these criteria. Scenario b) (heterogeneous sampling effort) should be avoided where possible. Indeed, disease prediction studies do not normally make use of spatially biased data as described in this scenario. In addition to the spatial coverage of sampling, statistical models should not be used to provide predictions in areas which are materially different from the area for which training data are available, as the modelled relationships may not be the same (Fitzpatrick and Hargrove, 2009). The example disease provided in this paper was a simulated disease; hence, full information was available on the covariates driving its spatial distribution. However, in real applications, the precise factors which drive the observed distribution are not necessarily known in advance; thus, the subset of potential covariates is selected based on (1) biological understanding; and (2) statistical modelling. This subset may not always represent the most appropriate subset for the disease under consideration, so a lack of data often results in important covariates being omitted from the modelling altogether. Thus, scenario c) (missing covariates) can be considered a frequent occurrence in practical applications.

The final scenario (disease not occupying its full niche) is likely to be the most common scenario encountered in spatial epidemiology applications. As an example, Rhodesian sleeping sickness (caused by the parasite *Trypanosoma brucei rhodesiense*) has been spreading in Uganda over the past two to three decades, with the movement of infected livestock implicated in the most recent introductions (Fèvre et al., 2001; Wardrop et al., 2010). This indicates that historically, the recorded spatial distribution of Rhodesian sleeping sickness did not cover all areas environmentally suitable for the disease and any predictive modelling based upon this distribution would not necessarily be providing the output intended. Most SDM are blind to the mechanisms that promote dispersal from affected to unaffected areas (e.g., human movements, contact patterns and trade), or factors that may inhibit spatial spread of a disease (e.g., human intervention), so resulting predictions are at best maps of potential risk. Most ecosystems are dynamic, and the spatial dispersal of a disease over time is not uncommon, enabling the disease to occupy a larger proportion of its potential range (Reisen, 2010). The identification and quantification of factors influencing this expansion would be required to ascertain the future risk of disease within currently unaffected areas. As Soberon (2010) argues, the fundamental ecological factors that determine species distributions are environment, biotic interactions and movements; without all three of these, modelled outputs of predicted occurrence and hence risk are compromised.

The four scenarios developed here should be taken into consideration when designing surveys and collecting data, fitting statistical models and during subsequent interpretation of predictive outputs. The goal of mapping should be clear from the outset (e.g., to map the present distribution or to map suitability) due to the impact of data acquisition choices on the final outputs. The consideration of whether an epidemiological situation may incorporate one (or more) of these scenarios should

provide greater awareness of the potential impacts on the modelling process and predictive maps. Model coefficients and estimates of uncertainty can only take us so far; the interpretation of these outputs needs to be undertaken with the four scenarios presented in this framework in mind to ensure accurate comprehension of meaning and consequent sound action in relation to decision-making. The premise of SDM is that predictive outputs will represent environmental suitability. However, where input data is not comprehensive, or where dynamic factors have not been taken into account, the predictive outputs may not represent environmental suitability, but may more accurately be described as representing the current distribution of the disease of interest.

Using a simulated dataset, this paper provides an overview of predictive mapping of disease and the linkages with ecological SDM, and has introduced some important considerations, which are rarely discussed in the predictive mapping literature. Care must be taken when carrying out predictive mapping when the distribution of the disease of interest is changing, and a full understanding of the disease's ecology alongside historical, recent and current spatial distributions of the disease should be used to inform the process of modelling and interpretation. Every mapping scenario will have different complexities which may influence the interpretation of resulting predictions, but time spent considering what the observed data represent and the implications of the possible scenarios detailed above will provide a starting point for more accurate interpretation of predictive maps. As long as the considerations introduced here are kept in mind, predictive mapping will continue to contribute significantly to epidemiological research and disease control planning.

Acknowledgements:
This work was supported by the Medical Research Council (PMA, NAW - projects G0902445 and MR/J012343/1). The funders had no role in the decision to publish or in preparation of the manuscript.

References

Austin MP, 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol Model 157, 101-118

Bergquist NR, 2001. Vector-borne parasitic diseases: new trends in data collection and risk assessment. Acta Trop 79, 13-20

Bivand R, Keitt T, Rowlingson B, 2013. Bindings for the geospatial data abstraction library. R package version 0.8-10.

Bivand, R. & Lewin-Koh, N. Maptools: tools for reading and handling spatial objects. R package version 0.8-25. 2013.

Bivand, R., Pebesma, E.J., & Gomez-Rubio, V. 2008. Applied Spatial Data Analysis with R New York, Springer.

Broennimann, O. & Guisan, A. 2008. Predicting current and future biological invasions: both native and invaded ranges matter. Biol Lett, 4, 585-589

Brooker, S. 2007. Spatial epidemiology of human schistosomiasis in Africa: risk models, transmission dynamics and control. Trans R Soc Trop Med Hyg, 101, 1-8

Brooker, S., Rowlands, M., Haller, L., Savioli, L., & Bundy, D.A.P. 2000. Towards an atlas of human helminth infection in sub-Saharan Africa: The use of geographical information systems

(GIS). Parasitol Today, 16, 303-307

Clements, A.C.A., Kur, L.W., Gatpan, G., Ngondi, J.M., Emerson, P.M., Lado, M., Sabasio, A., & Kolaczinski, J.H. 2010. Targeting Trachoma Control through Risk Mapping: The Example of Southern Sudan. PLoS Negl Trop Dis, 4,

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

Diggle, P., Moyeed, R., Rowlingson, B., & Thomson, M. 2002. Childhood malaria in the Gambia: a case-study in model-based geostatistics. J Roy Stat Soc C-App, 51, 493-506

Diggle, P. J. & Ribeiro Jr, P. J. 2007, "An overview of model-based geostatistics," In Model-based Geostatistics, New York: Springer, pp. 27-45.

Diggle, P.J., Thomson, M.C., Christensen, O.F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J.H., Boussinesq, M., & Molyneux, D.H. 2007. Spatial modelling and the prediction of Loa loa risk: Decision making under uncertainty. Ann Trop Med Parasitol, 101, 499-509

Elith, J., Kearney, M., & Phillips, S. 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution, 1, 330-342

Elith, J. & Leathwick, J.R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu Rev Ecol Syst, 40, 677-697

Fèvre, E.M., Coleman, P.G., Odiit, M., Magona, J.W., Welburn, S.C., & Woolhouse, M.E.J. 2001. The origins of a new Trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern Uganda. Lancet, 358, 625-628

Fielding, A.H. & Bell, J.F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv, 24, 38-49

Fitzpatrick, M.C. & Hargrove, W.W. 2009. The projection of species distribution models and the problem of non-analog climate. Biodivers Conserv, 18, 2255-2261

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

Hijmans, R. J. raster: Geographic data analysis and modeling. R package version 2.1-49. 2013.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol, 25, 1965-1978

Hijmans, R. J., Phillips, S., Leathwick, J. R., & Elith, J. dismo: Species distribution modeling. R package version 0.8-17. 2013.

Hutchinson, G.E. 1959. Homage to Santa Rosalia, or why are there so many kinds of animals. Am Nat, 93, 245-249

Koch, T. 2005. Cartographies of Disease: Maps, Mapping, and Medicine, 1 ed. Redlands, CA, ESRI

384 Press.

385 Legendre, P. 1993. Spatial autocorrelation - trouble or new paradigm. Ecology, 74, 1659-1673

386 Marmion, M., Luoto, M., Heikkinen, R.K., & Thuiller, W. 2009. The performance of state-of-the-art
387 modelling techniques depends on geographical distribution of species. Ecol Model, 220, 3512-3520

388 Nogues-Bravo, D. 2009. Predicting the past distribution of species climatic niches. Global Ecol
389 Biogeogr, 18, 521-531

390 Pearman, P.B., Guisan, A., Broennimann, O., & Randin, C.F. 2008. Niche dynamics in space and
391 time. Trends Ecol Evol, 23, 149-158

392 Pebesma, E. J. & Bivand, R. Classes and methods for spatial data in R. R News 5. 2005.

393 Pulliam, H.R. 2000. On the relationship between niche and distribution. Ecol Lett, 3, 349-361

394 R Development Core Team. R: A language and environment for statistical computing. 2013.
395 Vienna, Austria, R Foundation for Statistical Computing.

396 Reisen, W.K. 2010. Landscape Epidemiology of Vector-Borne Diseases. Annu Rev Entomol, 55,
397 461-483

398 Simarro, P.P., Cecchi, G., Paone, M., Franco, J.R., Diarra, A., Ruiz, J.A., Fevre, E.M., Courtin, F.,
399 Mattioli, R.C., & Jannin, J.G. 2010. The Atlas of human African trypanosomiasis: a contribution to
400 global mapping of neglected tropical diseases. International Journal of Health Geographics, 9,

401 Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. 2005. ROCR: visualizing classifier
402 performance in R. Bioinformatics, 21, 3940-3941

403 Snow, R.W., Marsh, K., & leSueur, D. 1996. The need for maps of transmission intensity to guide
404 malaria control in Africa. Parasitol Today, 12, 455-457

405 Soberon, J.M. 2010. Niche and area of distribution modeling: a population ecology perspective.
406 Ecography, 33, 159-167

407 Thomson, M.C., Connor, S.J., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M., &
408 Greenwood, B. 1999. Predicting malaria infection in Gambian children from satellite data and bed
409 net use surveys: The importance of spatial correlation in the interpretation of results. Am J Trop
410 Med Hyg, 61, 2-8

411 Wardrop, N., Atkinson, P.M., Gething, P.W., Fèvre, E.M., Picozzi, K., Kakembo, A., & Welburn, S.
412 2010. Bayesian Geostatistical Analysis and Prediction of Rhodesian Human African
413 Trypanosomiasis. PLoS Negl Trop Dis, 4, e914
414
415

Table titles:

Table 1. Four scenarios for disease modelling

Figure titles:

Figure 1. Map of Africa showing the bounding box of the study area in green.

Figure 2. Environmental suitability for hypothetical disease: suitable areas are shown in green and unsuitable areas in grey.

Figure 3. Example sample data used for the disease modelling scenarios showing (a) the full information scenario, (b) the heterogeneous sampling effort scenario, (c) the missing covariates scenario and (d) the disease not in equilibrium scenario: 100 simulated datasets were created for each scenario. Presence records are shown in red and absence records in black. Actual environmental suitability for disease transmission is shown in green.

Figure 4. Actual suitability for disease occurrence (top) and predicted probability of disease presence across the study area for each of the four scenarios: scenario a) full information (centre left), scenario b) heterogeneous sampling effort (centre right), scenario c) missing covariates (bottom left) and scenario d) disease not in equilibrium scenario (bottom right).

Figure 5. Mean ROC curves for 100 simulations of the four scenarios with 95% confidence intervals (dotted lines) showing (a) the full information scenario, (b) the heterogeneous sampling effort scenario, (c) the missing covariates scenario and (d) the disease not in equilibrium scenario.

Figure 6. Results of 100 simulations for the four scenarios showing (a) proportion of the study area for which predictions were correct (based on a cut-off probability of 0.5) and (b) AUC scores.

Table 1.

Scenario	Situation
a) Full information	The disease is in equilibrium with its environment and data are available for a spatially representative sample of its range.
b) Heterogeneous sampling effort	The disease is in equilibrium with its environment, but there is spatial bias in the detection of the disease (i.e. a heterogeneous sampling effort).
c) Missing covariates	The disease is in equilibrium with its environment and there is a spatially representative sample available, but the covariates used for prediction do not fully reflect the species environmental constraints.
d) Disease not in equilibrium with the environment	The disease is not in equilibrium with its environment due to either successful disease control (disease no longer occupying its full niche) or on-going spatial spread (the disease does not yet occupy its full niche).

Figure 1

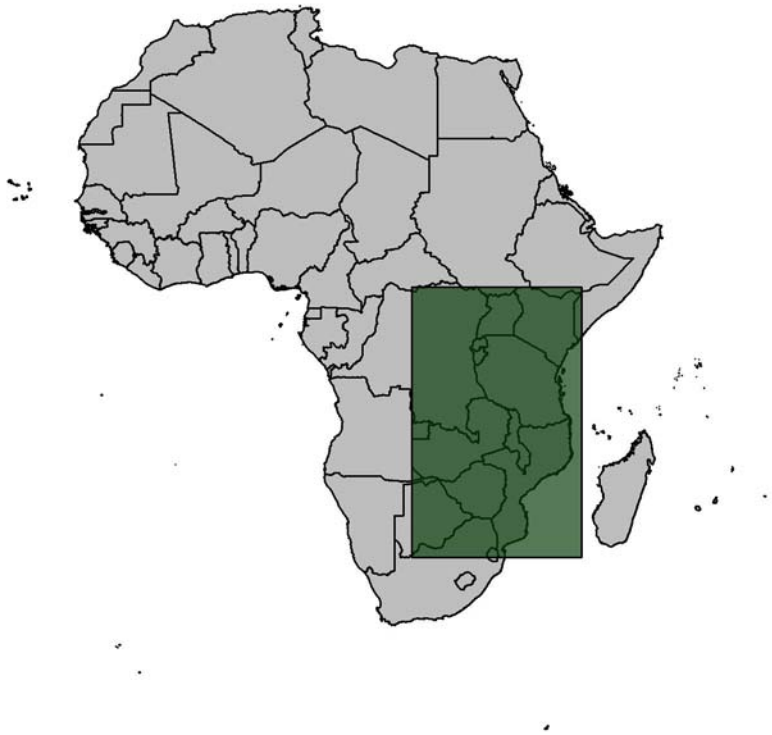


Figure 2

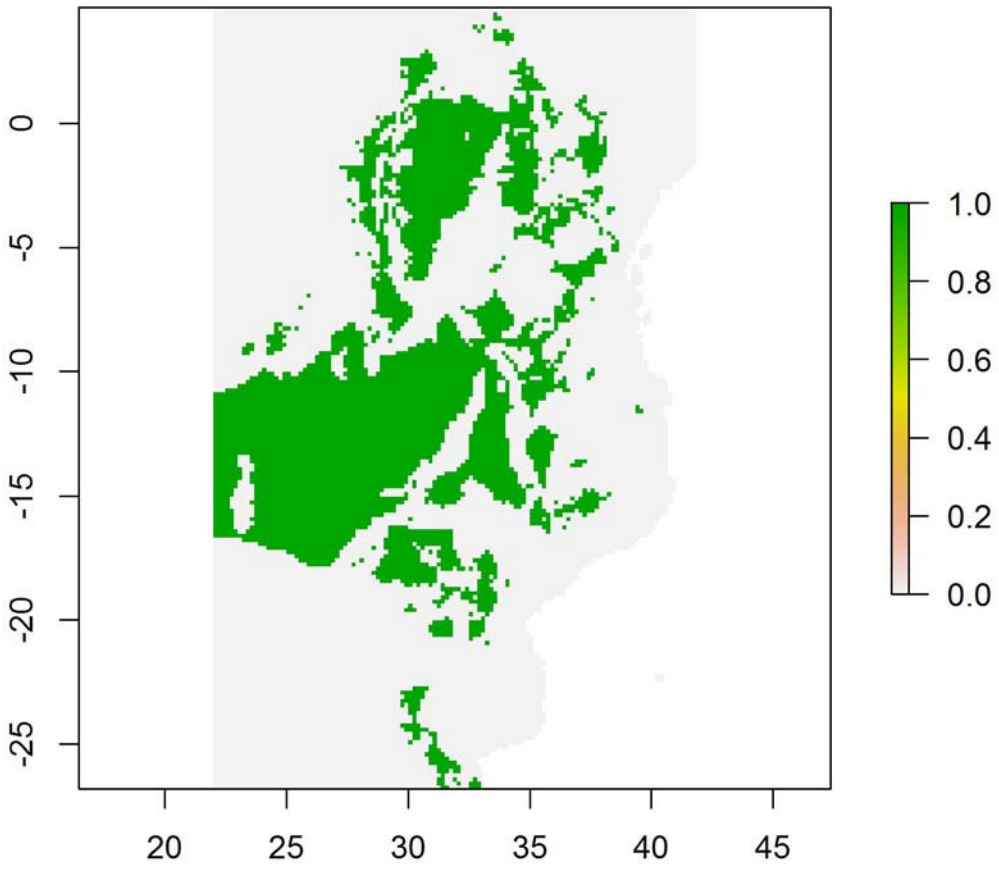


Figure 3

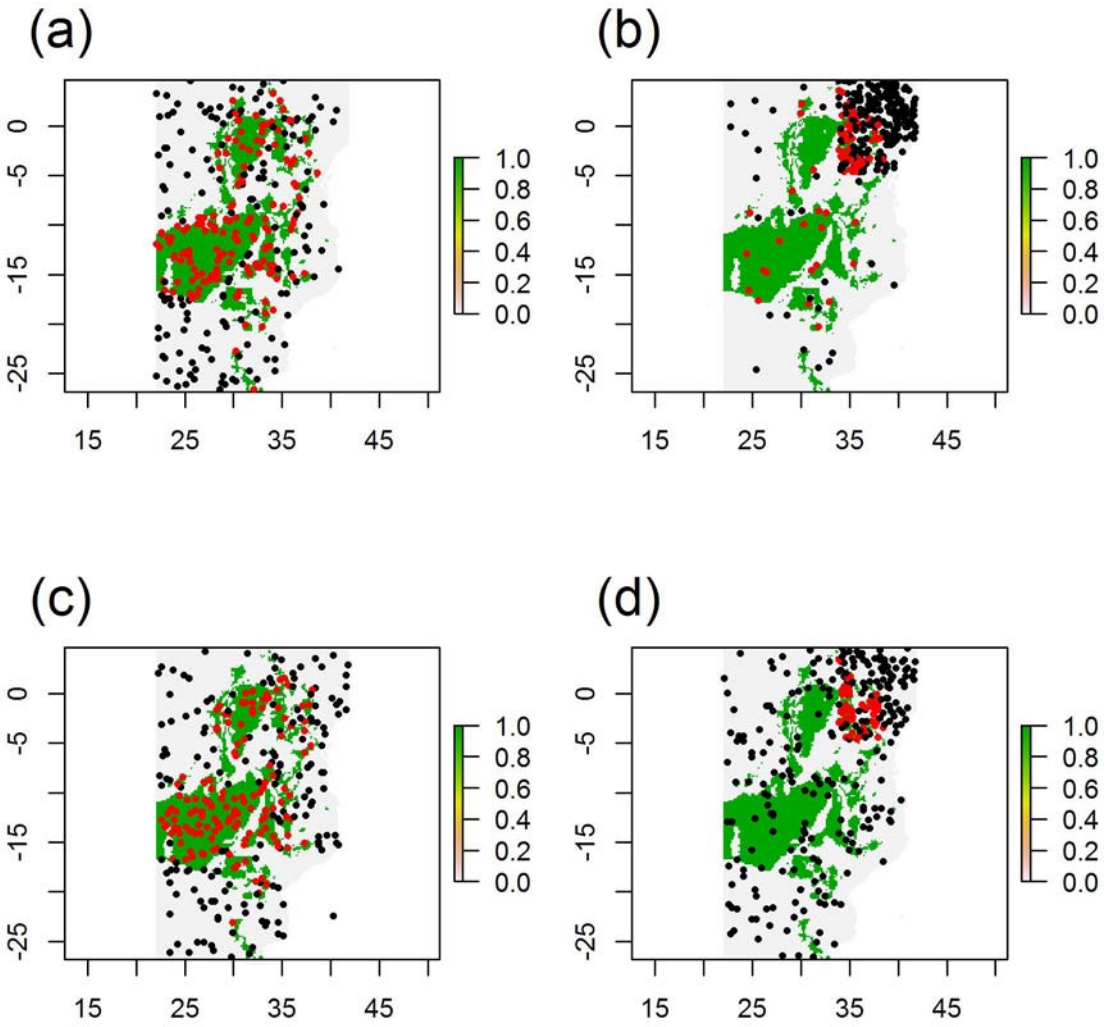


Figure 4

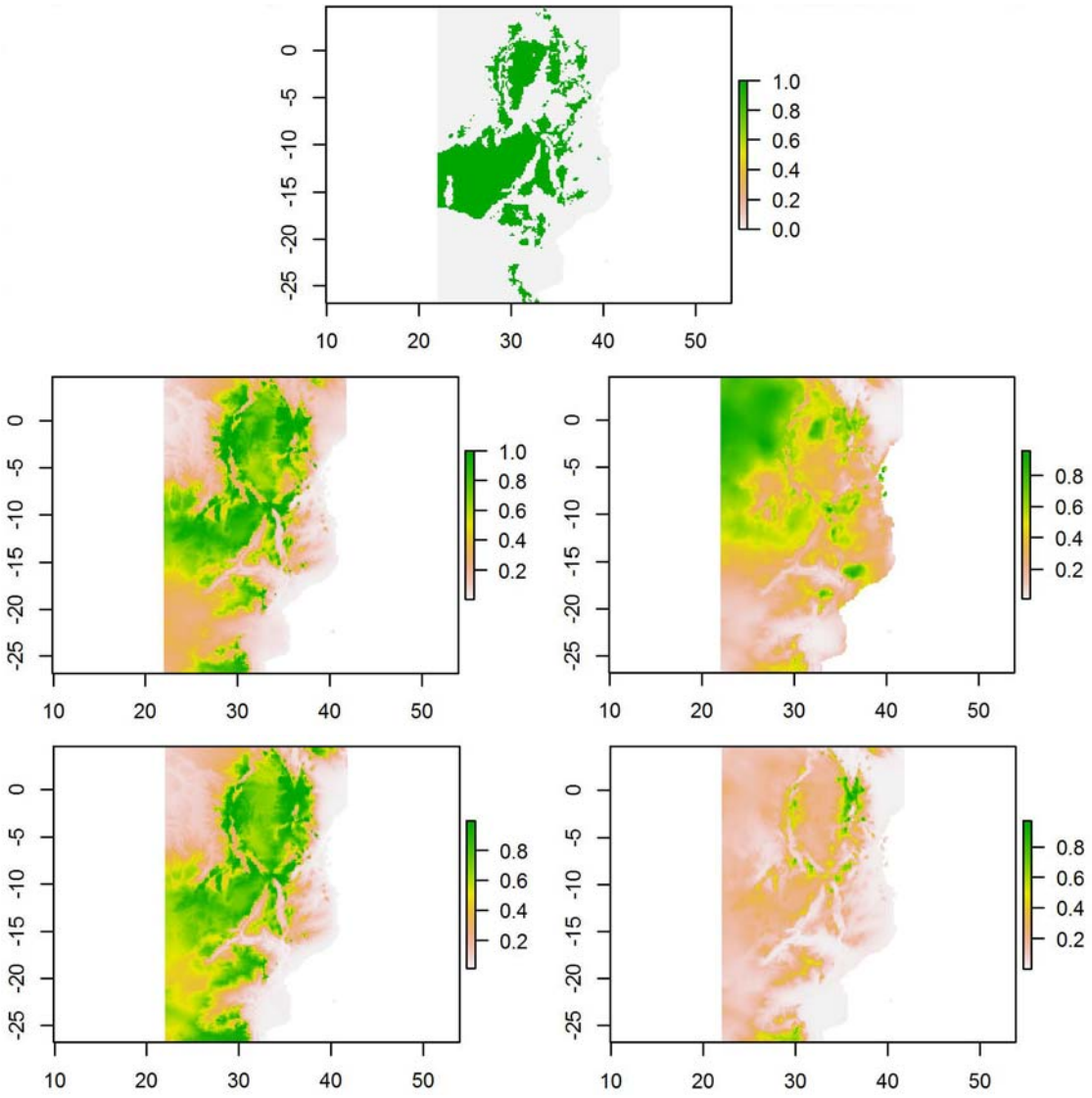


Figure 5

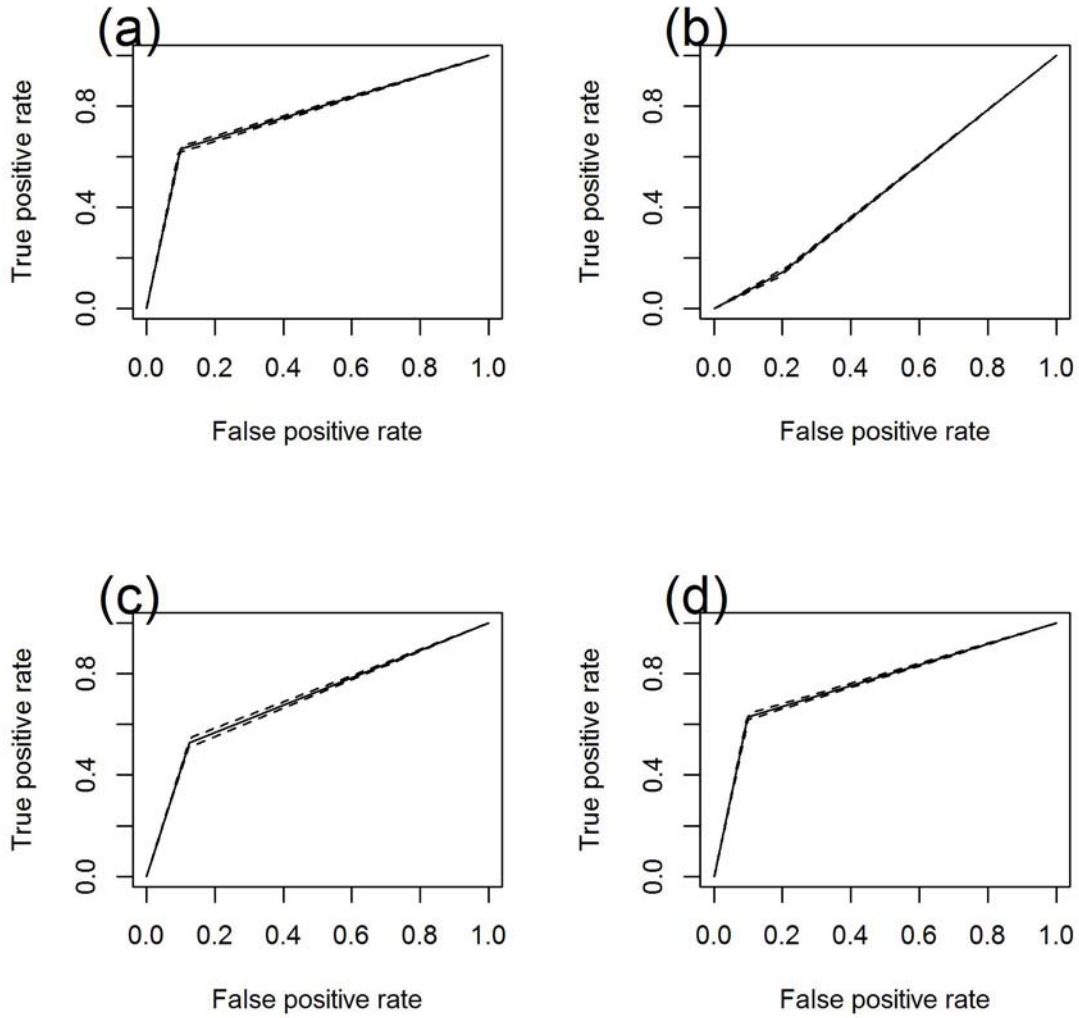
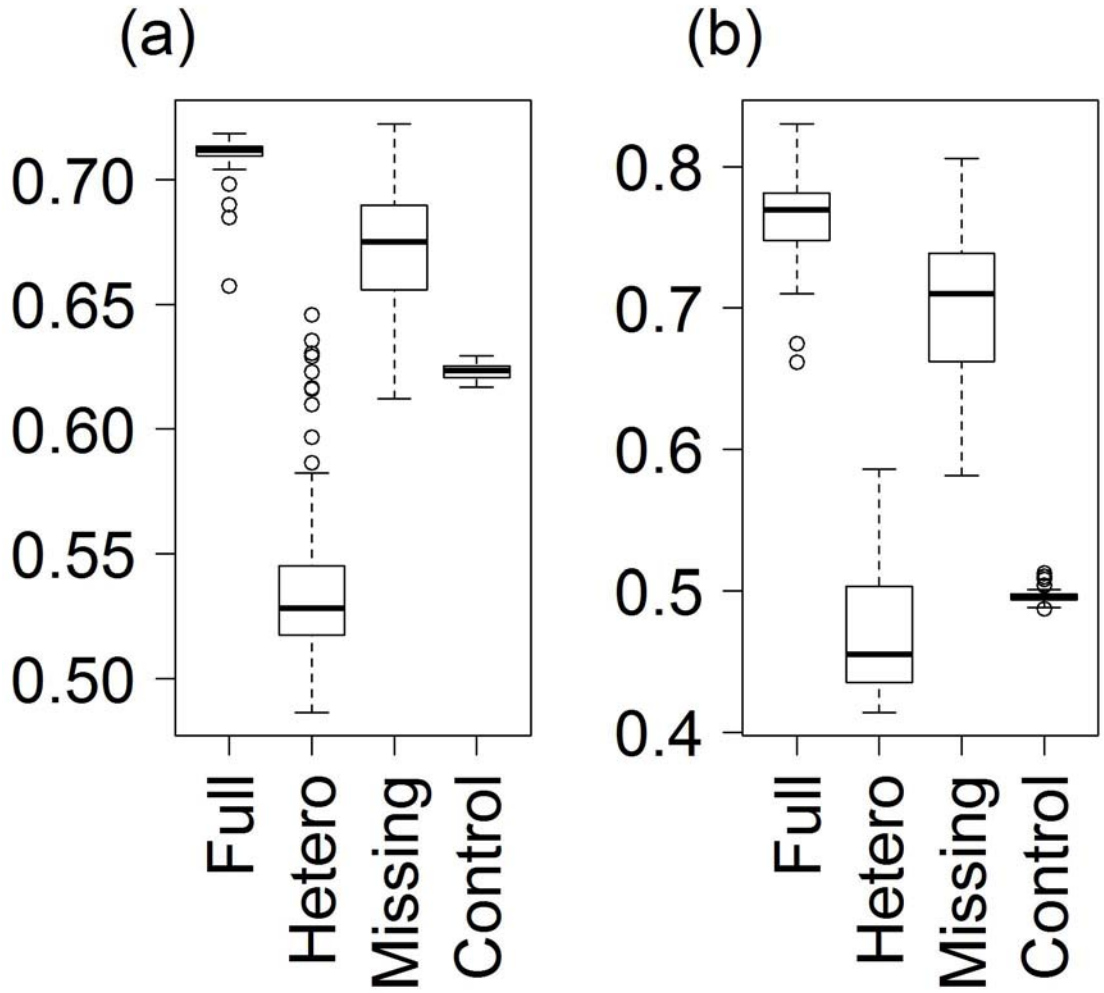

(a)
True positive rate
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
(b)
True positive rate
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
(c)
True positive rate
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
(d)
True positive rate
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

Figure 6


(a)
0.70
0.65
0.60
0.55
0.50
Full
Hetero
Missing
Control
(b)
0.8
0.7
0.6
0.5
0.4
Full
Hetero
Missing
Control