Spatial and Spatio-temporal Epidemiology 2 (2011) 125–133

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Spatial modelling of disease using data- and knowledge-driven approaches

Kim B. Stevens *, Dirk U. Pfeiffer

Veterinary Epidemiology and Public Health Group, Department of Veterinary Clinical Sciences, Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire AL9 7TA, UK

ARTICLE INFO

Article history:
Available online 19 July 2011

Keywords:
Disease distribution
Spatial modelling
Mapping
Data-driven
Knowledge-driven

ABSTRACT

The purpose of spatial modelling in animal and public health is three-fold: describing existing spatial patterns of risk, attempting to understand the biological mechanisms that lead to disease occurrence and predicting what will happen in the medium to long-term future (temporal prediction) or in different geographical areas (spatial prediction). Traditional methods for temporal and spatial predictions include general and generalized linear models (GLM), generalized additive models (GAM) and Bayesian estimation methods. However, such models require both disease presence and absence data which are not always easy to obtain. Novel spatial modelling methods such as maximum entropy (MAXENT) and the genetic algorithm for rule set production (GARP) require only disease presence data and have been used extensively in the fields of ecology and conservation, to model species distribution and habitat suitability. Other methods, such as multicriteria decision analysis (MCDA), use knowledge of the causal factors of disease occurrence to identify areas potentially suitable for disease. In addition to their less restrictive data requirements, some of these novel methods have been shown to outperform traditional statistical methods in predictive ability (Elith et al., 2006). This review paper provides details of some of these novel methods for mapping disease distribution, highlights their advantages and limitations, and identifies studies which have used the methods to model various aspects of disease distribution.

1. Introduction

The purpose of spatial modelling in animal and public health is three-fold: describing existing spatial patterns of risk using techniques such as kernel smoothing, kriging or Bayesian smoothing (i.e. descriptive), attempting to understand biological mechanisms that lead to the occurrence of disease (i.e. explanatory) and attempting to predict what will happen in the medium to long-term future or in different geographical areas (i.e. predictive). The results of such models are used for a variety of purposes including targeting areas for surveillance, risk management, simulating different control scenarios, predicting what will happen under different environmental conditions such as those resulting from climate change (i.e. temporal prediction) and identifying new geographical areas suitable for the introduction of diseases (i.e. spatial prediction).

Descriptive risk mapping aims to illustrate the spatial variation in disease risk while simultaneously removing excess noise or outliers using interpolation techniques such as kernel smoothing, kriging or Bayesian methods (Pfeiffer et al., 2008b). The data used in such maps are frequently obtained via surveys or surveillance. These data are only really

Abbreviations: AHP, analytical hierarchy process; AUC, area under the curve; BRT, boosted regression trees; CART, classification and regression trees; DDM, disease distribution model; ENFA, ecological niche factor analysis; ENM, ecological niche model; GARP, genetic algorithm for rule set production; GAM, generalized additive model; GLM, generalized linear model; MAXENT, maximum entropy; MCDA, multicriteria decision analysis; MPA, minimum predicted area; ROC, receiver operating characteristic; SDM, species distribution model.

* Corresponding author. Tel.: +44 (0) 1707666595; fax: +44 (0) 1707666374.
E-mail address: kstevens@rvc.ac.uk (K.B. Stevens).

1877-5845/$ - see front matter © 2011 Published by Elsevier Ltd.
doi:10.1016/j.sste.2011.07.007

useful when accompanied by denominator data so that the distribution of the population can be accounted for by calculating and mapping the disease risk or rate. However, descriptive risk maps can be incomplete or even misleading as they are frequently based on data which may have a sampling bias. In contrast, predictive risk mapping aims to identify new, unsampled geographical areas suitable for disease by extrapolating beyond the boundary of the data points used in the model (Peterson et al., 2004). However, extrapolation generally leads to greater uncertainty associated with the risk estimate. Such models can be thought of as disease distribution models (DDM), similar to the species distribution models (SDM) widely used in the fields of ecology and conservation to describe the species niche and to identify habitats suitable for supporting a species. SDM have been described as 'geographical modeling of biospatial patterns in relation to environmental gradients' (Franklin, 1995) but unlike SDM which are generally concerned with macro-organisms, DDM are more concerned with modelling the distribution and habitat suitability of disease micro-organisms or their vectors. Although the same methods and principles apply independent of the physical size of the species being modeled or whether the purpose is conservation of endangered species or identification of areas suitable for the introduction of a disease vector or micro-organism, modelling potential disease distribution introduces additional layers of complexity to methods which are generally best suited to modelling the relationship between a single species and its environment. Thus, certain SDM, while useful tools for modelling potential disease distribution, may not always be the most appropriate method for modelling complex disease patterns.

Traditional methods of temporal and spatial prediction include general and generalized linear models (GLM), generalized additive models (GAM) and Bayesian estimation methods (Lawson, 2006). These models comprise well-established algorithms and can effectively account for spatial dependence within the data. However, such models require both disease presence and disease absence points. Absence data are generally collected through observational studies which can be costly and, due to logistical constraints, may only cover a small geographical area. Yet DDM results are frequently applied to large areas (e.g. south-east Asia or Africa), even though the only available inputs may be disease presence data obtained through surveillance or knowledge of the causal factors leading to disease occurrence.

In the past decade the ecological and conservation fields have seen an increase in the application of SDM which require only presence data. These models, which are static and probabilistic, relate the species distribution to their current environment through a range of predictor variables believed to limit the species distribution (Guisan and Zimmermann, 2000). In addition to only requiring presence data, these models can also be extrapolated beyond the current distribution of the species or disease to show the potential distribution further afield. Some of these models have been shown to have predictive ability equal to or greater than that of traditional statistical methods (Elith et al., 2006) and have begun to be used in the animal and public health fields to produce maps showing the potential distribution of a variety of diseases (Peterson, 2006). This review paper provides details of some of these novel methods for modelling the spatial distribution of disease, – which can be divided into data- and knowledge-driven methods – highlights their advantages and limitations, and identifies studies which have used the methods to model various aspects of disease distribution.

2. Data-driven methods: presence–absence data

2.1. Classification and regression trees (CART)

Classification and regression trees (CART) is a non-parametric, decision-tree based method which explains the variation in a response variable with respect to one or more predictor variables (Breiman et al., 1984; Sutton et al., 2005). CART is a decision tree applied to qualitative and quantitative response variables respectively. Construction of a decision tree occurs in three stages; tree building, tree stopping and tree pruning (Olden et al., 2008). During the building phase a tree is constructed by repeatedly splitting the data into homogenous subgroups (usually binary) on the basis of a simple decision rule. For regression trees homogeneity of the groups is determined through calculation of the reduction in variance, while for classification trees it is determined using an information or entropy statistic, such as the Gini index (Franklin, 2009b). The split which results in the greatest homogeneity is used to divide the data and the process is repeated to produce a tree with nested decision thresholds (Franklin, 2009b).

Tree stopping occurs when a split no longer attains some predetermined level of increased homogeneity or when the resulting nodes would have less than a minimum number of observations. Tree pruning then involves removing those splits that contribute least to the overall subgroup homogeneity. The best tree size can be selected using the one standard error rule (i.e. the smallest tree within one standard error of the minimum) (De'ath and Fabricius, 2000). Free software for implementing decision trees is available at the following websites: http://www.cs.waikato.ac.nz/ml/weka/index.html and http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html.

The advantages of CART are discussed in detail in various studies (De'ath and Fabricius, 2000; Franklin, 2009b; Sutton et al., 2005) and include their flexibility with respect to a variety of response types (for example, numeric, categorical, ratings and survival data), outliers, missing data and spatial dependence. Decision trees are non-parametric, so can be used to analyse highly-skewed data (Ivanek et al., 2009), and are easy to construct and interpret; when displayed graphically, decision trees are highly intuitive and provide prediction rules which can be extrapolated to other areas (Elnaggar and Noller, 2010). They are particularly useful for describing patterns in large, messy and complex datasets. Limitations of this method include a lack of significance testing of the identified risk factors and the risk of model over-fitting. Trees can be unstable and have been shown to have lower predictive accuracy than statistical methods (Elith et al., 2006): they need to be based on large datasets to allow for drawing meaningful

and generalizable inferences. In addition, trees are not guaranteed to be optimal, especially if they incorporate correlated variables (Sutton et al., 2005).

CART have been used to identify risk factors for malaria in central Vietnam (Thang et al., 2008), determine the probability of *Listeria* species isolation from natural environments (Ivanek et al., 2009) and to investigate disease risk in Welsh cattle herds (Ortiz-Pelaez and Pfeiffer, 2008).

2.2. Bagging, boosting and random forests

Decision trees have been taken a step further through the introduction of bagging, boosting and random forest methods. Briefly, all three methods involve generating a large number of decision trees using different subsets of data, and then averaging the results to obtain final prediction values. Bagging entails repeatedly sampling the data, with replacement, to create a decision tree for each subset of data (Breiman, 1996; Sutton et al., 2005). In each instance two-thirds of the data are used to generate the model and one-third remains 'out-of-bag' for evaluating the model. The predictions from all the trees are then averaged to create a final prediction. Boosting (Freund and Schapire, 1996; Ridgeway, 1999; Sutton et al., 2005) is similar to bagging but instead of each observation having an equal probability of being selected, observations are weighted; those which were misclassified in previous models receive a higher weighting (De'ath, 2007; Elith et al., 2008). Standard errors for the predictive values can be generated using bootstrap procedures (Leathwick et al., 2006). Random forests (Breiman, 2001) are a type of bagging that produces a large number of decision trees (Hastie et al., 2009) which are then averaged.

Together with the usual advantages of decision trees, trees produced using these enhanced machine learning methods have been shown to have a predictive accuracy similar to that of statistical models (Elith et al., 2006, 2008). However, producing ensembles of trees may make them difficult to interpret in terms of the relevance of individual risk factors.

Boosted regression trees (BRT) have been used to map the distribution of *Anopheles* species in Africa, Europe and the Middle East (Sinka et al., 2010) and to predict the spatial distribution of wild water bird reservoirs of highly pathogenic avian influenza in the Inner Niger Delta in Mali (Cappelle et al., 2010).

3. Data-driven methods: presence-only data

Although statistical and machine learning methods require both disease presence and absence data, absence data can be problematic because it can be difficult to distinguish between the absence of a disease and the lack of observation or reporting of disease events in an area. Alternatively, the disease species may be absent – even though the habitat is suitable for its occurrence – due to a geographical or man-made barrier preventing its spread into the area (Hirzel et al., 2002). These situations can be considered 'false absences', biasing study results. However, if a disease has been confirmed at a location we can be reasonably certain that it occurs there. Thus, there are considerable advantages to using predictive mapping methods which use only presence data and preclude the use of absence data. Some 'presence-only methods' such as ecological niche models (ENM) use pseudo-absence or background data in place of actual absence data. An additional advantage of using presence-only methods versus more traditional methods is that data from international georeferenced disease reporting systems can be used to explore the large-scale spatial epidemiology of a disease and generate national or even global disease maps – something which would be impossible to do if we had to rely on presence-absence data generated through observational studies.

Unlike traditional statistical methods of disease mapping which show absolute risk, presence-only methods show the likelihood of an organism being present or the habitat suitability of a spatial unit relative to another unit within the study area (Franklin, 2009d). In addition, SDMs predict the probable presence or absence of a species and not its abundance.

3.1. Ecological niche models (ENMs)

The ecological niche is defined as the biological and physical conditions under which a species can maintain a population without immigration (Grinnell, 1917). This can be divided further into the fundamental and realized niches; the fundamental niche describes a species' potential distribution, while the realized niche is the subset of the fundamental niche that it actually occupies (Hutchinson, 1959). ENMs predict the fundamental niche of the species from a suite of predictor variables and project this onto geographical space (Peterson, 2007). Ecological niches are usually defined at coarse spatial scales to avoid 'complexities of biotic interactions' (Williams and Peterson, 2009).

3.1.1. Ecological niche factor analysis (ENFA)

Ecological niche factor analysis (ENFA) is a presence-only method used to map species distributions (Hirzel et al., 2002). The method is based on the theory that species occur non-randomly in space, defined by certain eco-geographical variables. For example, a species whose distribution is limited by temperature will occur in spatial cells that fall within its optimum range, which comprise only a small proportion of the total range of this eco-geographical variable. The species distribution may therefore comprise only part of the larger distribution of all the cells and may differ from the distribution of the background cells in both its mean and variance. This difference allows the species 'marginality' and 'specialization' to be determined. A marginality close to one indicates that the species lives in a very specific habitat relative to the reference cells while a marginality of zero implies that the species is ubiquitous. Alternatively, specialization describes the species' level of tolerance; a value greater than one indicates some form of specialization. However, these values will depend on the reference set of cells and the resolution at which the analysis occurs.

Having defined the species multivariate niche, factor analysis is then used to extract those variables which contribute the most to a species' marginality and specialization, with the first axis accounting for all the marginality while the subsequent axes maximize specialization (Hirzel et al., 2002). Having summarized the predictor variables into a few uncorrelated – yet ecologically relevant factors – ENFA constructs habitat suitability maps by combining the chosen factors using one of four algorithms: (1) median (Hirzel et al., 2002), (2) distance geometric mean (Hirzel and Arlettaz, 2003), (3) distance harmonic mean and (4) minimum distance. In addition, three algorithms have been developed for use when the study area is at the edge of the species niche (Braunisch et al., 2008). ENFA is implemented in the free software Biomapper (http://www2.unil.ch/biomapper/) (Hirzel et al., 2007).

The main advantage of ENFA as a predictive mapping tool is that it does not require absence data to define areas suitable for disease distribution. By using factor analysis to summarise the ecological predictors into a few uncorrelated factors – which still provide ecological information – the method is able to cope with the correlation between factors frequently encountered in habitat suitability modelling (Hirzel and Arlettaz, 2003). In addition, the relationship between habitat suitability and environmental conditions is seldom linear or monotonic (assumptions of discriminant functions and first-order regressions). Therefore ENFA, which theorizes that suitability decreases from either side of an optimum, allows for a better representation of ecological conditions than traditional statistical methods (Hirzel et al., 2002).

There are four limitations of ENFA, as discussed by the developer of the method (Hirzel et al., 2002). The distribution maps currently do not include confidence intervals (although these could be obtained through bootstrapping the presence data) and the method can only include linear dependencies within the species niche. Some variables may be constant, particularly if variables are measured at a coarse scale or when using small species datasets. Although the Biomapper software identifies the constant variable, allowing the user to remove it from the model, a more effective solution would be to measure variables on a finer scale. In addition, the ecological niche is characterized relative to a particular reference set, and therefore using a different reference set could lead to different results.

ENFA has not been widely used to create disease maps, possibly because it has been superseded by other presence-only methods such as the genetic algorithm for rule set production (GARP) and maximum entropy (MAXENT). However, ENFA has been used to define the climatic niches of six species of tick involved in disease transmission in the Mediterranean region and forecast changes in habitat suitability as a result of climate change (Estrada-Pena and Venzal, 2007), and to construct habitat suitability maps for the five *Anopheles* species responsible for the majority of malaria transmission in Cameroon (Ayala et al., 2009).

3.1.2. Genetic algorithm for rule set production (GARP)

A more recent presence-only method which has been widely used to create disease maps is GARP (Stockwell and Noble, 1992; Stockwell and Peters, 1999); a genetic algorithm which uses an iterative process in which rules are selected, evaluated, tested, and either incorporated or rejected from the final model. GARP uses four types of rules to identify a species' ecological niche – atomic, envelope, range and logit rules (Stockwell et al., 2006). The model starts with a set of rules which are modified in various ways leading to their improvement. The predictive accuracy of the rule is assessed using a χ^2 test of the difference in the probability of the predicted value before and after the rule is applied. The difference in predictive accuracy between iterations is used to decide whether or not a particular rule should be included in the model. As rule modification may lead to improved rules, better models are achieved with longer run times (Stockwell and Noble, 1992). The rule set with the highest predictive accuracy is used to create the predictive distribution map which is projected onto geographical space to identify areas of suitable habitat for the species. Each run of the model results in a slightly different map because of the randomness included in the iterative process. Averaging the results from a series of runs produces a model with values ranging from 0% to 100%, which is interpreted as the probability of occurrence. GARP can be implemented using either the free software Desktop GARP (http://www.nhm.ku.edu/desktopgarp/) or the more recent and extensive version of the software available on the openModeller (de Souza Muñoz et al., 2009) website (http://openmodeller.sourceforge.net/index.php?option=com_frontpage&Itemid=1).

The advantages of GARP have been discussed in detail elsewhere (Stockwell and Peters, 1999). GARP requires few data points to achieve a high level of accuracy (Stockwell and Peterson, 2002b) and the resulting models can be successfully extrapolated beyond the area or environmental conditions used in the model (Soberon and Peterson, 2005). GARP predicts well using a range of environmental types, not just continuous predictor variables (Stockwell and Peterson, 2002b) and has also been shown to perform better than other models with respect to the problem of over-fitting (Stockwell and Peterson, 2002a). It has been suggested that genetic algorithms such as GARP are most useful when the modeller has little background knowledge of the species being modelled (DeJong, 1988) or when the dataset being modelled is large, complex and noisy (Fitzpatrick and Grefenstette, 1988). However, GARP tends to over-predict species distributions (Lawler et al., 2006), especially when using small data sets (Wisz et al., 2008).

GARP is the algorithm used in Lifemapper (http://www.lifemapper.org/), a distributed computing project harnessing the computing power of hundreds of thousands of personal computers to create a 'predictive electronic atlas of the earth's biological biodiversity' (Stockwell et al., 2006). On a smaller scale GARP has been used to model the persistence of highly pathogenic avian influenza H5N1 virus worldwide (Hogerwerf et al., 2010) and more specifically in the Middle East and northeastern Africa (Williams and Peterson, 2009) and Nigeria and West Africa (Williams et al., 2008). The distribution of *Bacillus anthracis*, the causative agent of anthrax, has been modelled in the United States (Blackburn et al., 2007) and Kazakhstan (Joyner et al., 2010) while the distribution of leishmaniasis

vector species has been modelled in São Paulo (Peterson et al., 2004), the state of Bahia, Brazil (Nieto et al., 2006) and southern Brazil (Peterson and Shaw, 2003). GARP has also been used to predict the global risk of spread of the mosquito *Aedes albopictus* (Benedict et al., 2007) and to model the distribution of the flea vectors of the plague (*Yersinia pestis*) (Adjemian et al., 2006) and the disease itself (Maher et al., 2010; Neerinckx et al., 2008), the distribution of *Cryptococcus gattii* (Mak et al., 2009), malarial vectors in northern Australia (Sweeney et al., 2007), Africa (Levine et al., 2004a) and the United States (Levine et al., 2004b) and the tick vectors of Lyme disease in British Columbia, Canada (Mak et al., 2010).

3.1.3. Maximum entropy (MAXENT)

MAXENT is a machine learning method which originated from statistical mechanics and information theory but which has since been successfully used in a wide range of fields, including that of species distribution modelling (Phillips et al., 2006). Presence data, together with maps of relevant predictor variables are used to estimate the target distribution by identifying the distribution closest to uniform (i.e. maximum entropy). This is subject to the constraint that the expected value of each predictor variable under this estimated distribution matches its empirical average (Phillips et al., 2004).

The contribution of each predictor variable to the model is calculated using a jackknife procedure in which the area under the receiver operating characteristic (ROC) curve (AUC) is determined for the model incorporating each variable individually, and again after removing the variables one by one (Phillips et al., 2006). The difference between these values is calculated and those predictor variables for which the difference is greatest are considered to contribute the most information. From this, the contribution of each predictor to the final model is represented as a percentage (Phillips et al., 2006).

MAXENT incorporates a number of features which can be used to constrain the species geographic space (Phillips et al., 2006; Phillips and Dudík, 2008). These include linear (a continuous variable), quadratic (the square of a continuous variable), product (the product of two continuous variables), threshold (piecewise constant responses), binary and hinge (piecewise linear responses) features. Modellers should choose those features they think most likely constrain the species' geographic space. The resulting models have a natural probabilistic interpretation, exhibiting a smooth gradient from least to most suitable (Phillips et al., 2004). Recently a logistic output format, which estimates the probability of species presence dependent on the predictor variables, was added to the MAXENT freeware (downloadable from http://www.cs.princeton.edu/~schapire/maxent/).

The advantages and limitations of MAXENT have been discussed in detail elsewhere (Phillips et al., 2006). This method can be used successfully with very small sample sizes – less than 100 observations (Phillips et al., 2004) or even as few as 10 presence records (Wisz et al., 2008). The method does not assume observations are independent (Phillips et al., 2006), can incorporate interactions between variables and uses both continuous and categorical data (Phillips et al., 2006). In addition, the resulting models are easily interpreted (Phillips et al., 2004). Although models which include a large number of predictor variables tend to over-fit small training sets due to correlations between predictor variables, this results in more accurate models when large training sets are used (Phillips et al., 2004). As MAXENT is optimized to predict within the realized niche, it should be used with caution when predicting outside the realized distribution (Phillips and Dudík, 2008). It has been suggested that MAXENT models may be highly conservative, predicting false absences more often than GARP (Moffett et al., 2007). As such regions predicted as suitable for a species can be considered reliable but regions predicted as unsuitable should not be discounted.

MAXENT has been used to predict the distribution of a number of disease vectors including those for malaria in Africa (Moffett et al., 2007) and Northern Tanzania (Kulkarni et al., 2010), leishmaniasis in France (Chamaillé et al., 2010) and North America (Gonzalez et al., 2010), West Nile virus in Iowa (Larson et al., 2010), Chagas disease in Texas (Sarkar et al., 2010) and four equine piroplasms in Greece. The distribution of two Phlebotomus species (vectors of visceral and cutaneous leishmaniasis) has been modelled for the Middle East (Colacicco-Mayhugh et al., 2010).

4. Comparison of data-driven methods

A comprehensive comparison of 16 modelling methods over 226 animal species from six regions of the world found that some of the newer methods outperformed the traditional statistical methods with regards to predicting species distribution (Elith et al., 2006). The authors divided the models into three groups based on their performance as assessed by AUC and correlation. The group with the best performance included BRT and MAXENT; most of the standard regression methods (GLM and GAM), including the OpenModeller version of GARP, had intermediate performance; while the group with the worst performance included Desktop GARP. Although other comparative studies have focused on a single region or species, in general they have provided similar results (Elith and Graham, 2009; Hernandez et al., 2008; Prasad et al., 2006; Segurado and Araujo, 2004; Tsoar et al., 2007). However, most model comparisons have been based on AUC or kappa, whereas studies have shown that models can have very similar predictive ability but generate very different spatial distributions (Pearson et al., 2006). Thus, emphasis should be placed on both measures of predictive ability and the predicted distribution when comparing methods (Franklin, 2009c). Choice of modelling method depends not only on the type of data available but also whether the primary objective is the identification of variables which influence the distribution of the disease, interpolation of unmeasured geographical areas or extrapolation into new areas.

5. Knowledge-driven models

5.1. GIS-based multicriteria decision analysis (MCDA)

In data-sparse situations, maps identifying areas suitable for disease can be produced using GIS-based multicri-

teria decision analysis (MCDA) – also known as multicriteria decision modelling or making (MCDM). This method uses decision rules derived from existing knowledge to identify areas potentially suitable for disease (Pfeiffer et al., 2008a). Details of the modelling process have been provided by a number of authors (Eastman, 1997; Eastman et al., 1995; Malczewski, 1999, 2000, 2004; Pfeiffer et al., 2008a) and include defining the objective of the modelling exercise, defining the factors and constraints and identifying the necessary factor maps, defining the relationship between each factor and suitability, defining the relative importance of each factor in relation to the objective, standardizing the factors, combining factors, constraints and weights to produce a final weighted estimate of suitability for each cell in the study area and finally, ranking the cells in order of suitability. However, GIS-based MCDA is frequently implemented in a way which does not take into account the important assumptions underpinning the approach, or the issues surrounding the creation of factor weights or attribute maps. In an attempt to highlight these oversights a comprehensive comparison of the common and best practice approaches for GIS-based MCDA are detailed elsewhere (Malczewski, 2000).

The properties that the risk factor maps should have – both individually and as a set – include that they be comprehensive (i.e. 'its level for a particular decision problem clearly indicates the degree to which the associated objective is achieved (Malczewski, 2000), measurable (i.e. able to assign numeric values to the attribute), complete (i.e. cover all aspects of the problem), operational (i.e. can be meaningfully used in the analysis), decomposable (i.e. the performance of one attribute can be evaluated independently of its performance on another attribute), non-redundant (i.e. situations in which double-counting occur should be avoided) and minimal (i.e. the number of factors included should be kept to a minimum) (Malczewski, 1999, 2000). Yet these attributes are frequently ignored in GIS-based approaches because it is often difficult to meet them all in spatial modelling situations (Malczewski, 2000).

There is a large participatory element to MCDA as expert or local opinion can be used to identify risk factors, determine the relationships between risk factors and the objective, and define weights for the factors. Although there are a number of different ways of defining factor weights (Malczewski, 2000; Steele et al., 2009) the most commonly used method is Saaty's analytical hierarchy process (AHP) (Saaty, 1980, 1990), mainly because it is easily understood by both modellers and decision makers and is easily implemented. However, the need for caution when using AHP has been highlighted (Steele et al., 2009) because it is frequently implemented without taking into account the different scales of the various factors. Before combining the factor maps they need to be standardized; a linear transformation is most frequently used (Malczewski, 2000), yet different transformations can produce very different results. The transformation used to standardize the maps has been shown to have a greater effect on the final map than the choice of weights (Clements et al., 2006). Although factors are most frequently combined using weighted linear combination, other methods such as ordered weighted averaging (Malczewski, 2004, 2006; Pfeiffer et al., 2008a) can also be used to produce a final weighted estimate of suitability for each cell in the study area. To convert the continuous suitability scale into a binary one a simple choice heuristic is used to rank all cells and the resulting map is divided by the maximum rank to produce a map of ranked suitability (Eastman, 1997; Malczewski, 2000). MCDA maps describe the likelihood of disease or habitat suitability in one spatial unit relative to the likelihood in another unit within the study area.

Estimates of uncertainty can be incorporated into MCDA maps using either fuzzy logic, Dempster-Shafer theory or both. Fuzzy logic can be used to model uncertainty regarding the accuracy of the map layers or to define the relationship between different factors and the outcome using fuzzy membership functions (Pfeiffer et al., 2008a). Dempster-Shafer theory, on the other hand, (Dempster, 1966, 1967) provides a visual representation of the level of uncertainty surrounding the probability estimate (Eastman, 2006; Pfeiffer et al., 2008a).

As the disease suitability maps generated using MCDA require no data they are relatively quick and inexpensive to produce. Together with their large participatory element, they provide a framework for interaction between formal statistical methods and human intuition. By using non-linear transformations when standardizing factor maps MCDA is able to accommodate the non-linear relationships generally encountered between disease organisms, vectors or reservoirs and their environment. Both quantitative and qualitative variables can be included in the model. It is also possible to explicitly incorporate uncertainty associated with variables as well as causal relationships, and this can be different between different outcome categories (Pfeiffer et al., 2008a). Limitations of MCDA include the subjectivity associated with the identification of risk factors, membership functions and weights, and the risk of error propagation. In addition, because the modelling rules often assume additivity issues arise when using correlated data. As with disease maps generated using ENM, only factors for which georeferenced data are available can be modelled and therefore the final suitability map may not be comprehensive.

Studies that have used MCDA to map disease include predicting the suitability of both Africa (Clements et al., 2006) and Senegal (Clements et al., 2007) for the occurrence of endemic or epidemic Rift Valley fever, and prioritizing areas in Madagascar for malarial vector control (Rakotomanana et al., 2007).

6. Model uncertainty and validation

Disease maps should always be used in conjunction with explicit information regarding the source and magnitude of any uncertainties incorporated in the modelling process (Barry and Elith, 2006; Pfeiffer et al., 2008a). Model uncertainty can be either epistemic (e.g. measurement error, natural variability, model uncertainty) or stochastic (i.e. uncertainty due to inherent variability in the underlying biological processes) (Refsgaard et al., 2007). Although uncertainty resulting from data errors can be substantial,

of equal importance is the uncertainty resulting from missing predictors due to incomplete knowledge of the organism being modelled or because the spatial data for the predictor is unavailable (Franklin, 2009a). Estimates of uncertainty are generally only generated or considered at the end of the modelling process despite that, whenever possible, the concept should be incorporated at all stages (Refsgaard et al., 2007). A detailed framework for assessing the predictive uncertainties of environmental models is provided elsewhere (Refsgaard et al., 2006).

Most modellers are familiar with Box's (Box and Draper, 1987) quotation that 'Essentially, all models are wrong, but some are useful'. As all models are simplifications of reality they will contain prediction errors, and model validation allows us to quantify the amount of error in the model. Validation techniques generally assess the predictive ability of the model. However, this is only one aspect of validation; the model also needs to be ecologically realistic.

Although researchers prefer to work with a continuous scale so that even the smallest change in risk levels can be identified, decision makers usually require a binary map showing areas which are suitable or unsuitable for disease. A probabilistic output therefore needs to be converted to a binary one using a threshold value and a range of methods exist for selecting such a threshold (Liu et al., 2005). Owing to the binary nature of the output, the usual indicators of model validation such as R^2 are not applicable. Instead distributional models make use of threshold-dependent or threshold independent means of validation. Most threshold-dependent measures are based on the confusion matrix; a 2×2 table of the observed versus the predicted presence and absence. From this table accuracy measures such as sensitivity, specificity, false positive (negative) rate and kappa can be calculated, as discussed elsewhere (Franklin, 2009a). Threshold-independent measures such as AUC are favoured by many modellers, although some (Lobo et al., 2008) caution against its use for a number of reasons including that the AUC ignores the predicted probability values and goodness-of-fit of the model.

All these validation measures assume the availability of both presence and absence data, yet in situations which use presence-only data a uniform, random sample of the background data (known as pseudo-absences) can be used in place of true absence data. However, in such instances the measures need to be interpreted slightly differently. For example, a model AUC calculated using pseudo-absence data shows the probability that a random presence site scores higher than a random background site (Phillips et al., 2009). Some authors (Pearce and Boyce, 2006) insist that presence-only data should be validated using presence-only methods such as the minimum predicted area (MPA) (Engler et al., 2004; Peterson et al., 2008). MPA is similar to the ROC curve but instead of plotting sensitivity against 1 – specificity, sensitivity is plotted against the proportion of land predicted as suitable.

7. Conclusion

Spatial modelling and disease mapping are seldom ends to themselves. Rather, they are a means to an end; tools for informing surveillance, decision making and disease risk management. However, there is a need to communicate the outputs of such models more effectively to gain the trust of decision makers and risk managers. In addition, modellers need to communicate any uncertainty and biases inherent in the model outputs in a way that makes them easily understood by decision makers so that models are transparent and decision makers can use them with confidence. Despite the wide range of methods available they are all only as good as their inputs, whether data collected via observational studies, surveillance data or knowledge regarding the epidemiology of a disease. Although the quantity of available georeferenced data has increased during the last decade, ensuring that only data of the highest quality is used for modelling purposes remains of paramount importance. This needs to be combined with an awareness of any possible limitations and biases associated with that data. The wide choice of available modelling methods means that it is no longer necessary to rely solely on traditional statistical methods such as GLMs or GAMs and greater use should be made of those novel methods which have generally been shown to outperform the more traditional methods. Unfortunately GARP, the method which has been most extensively used in disease distribution modelling, has also been shown to be one of the worst performing methods when compared with others, while BRT and MAXENT – two methods with the highest performance – have been infrequently used to model disease distribution. Future studies should therefore concentrate on those methods which have been shown to be superior. However, when using these newer methods it is imperative that the modeller be fully cognizant of the theoretical and biological processes underpinning the models, the assumptions inherent in their methodology, and correct interpretation of the output. Otherwise the models may be incorrect.

References

Adjemian JCZ, Girvetz EH, Beckett L, Foley JE. Analysis of genetic algorithm for rule-set production (GARP) modeling approach for predicting distributions of fleas implicated as vectors of plague, *Yersinia pestis*, in California. J Med Entomol 2006;43:93–103.

Ayala D, Costantini C, Ose K, Kamdem G, Antonio-Nkondjio C, Agbor J-P, et al. Habitat suitability, ecological niche profile of major malaria vectors in Cameroon. Malar J 2009;8:307.

Barry S, Elith J. Error and uncertainty in habitat models. J Appl Ecol 2006;43:413–23.

Benedict M, Levine R, Hawley W, Lounibos L. Spread of the tiger: global risk of invasion by the mosquito *Aedes albopictus*. Vector Borne Zoonotic Dis 2007;7:76–85.

Blackburn JK, McNyset KM, Curtis A, Hugh-Jones ME. Modeling the geographic distribution of *Bacillus anthracis*, the causative agent of anthrax disease, for the contiguous United States using predictive ecologic niche modeling. Am J Trop Med Hyg 2007;77:1103–10.

Box GEP, Draper NR. Empirical model-building and response surfaces. Wiley; 1987.

Braunisch V, Bollmann K, Graf R, Hirzel A. Living on the edge – modelling habitat suitability for species at the edge of their fundamental niche. Ecol Modell 2008;214:153–67.

Breiman L. Bagging predictors. Mach Learn 1996;26:123–40.

Breiman L. Random forests. Mach Learn 2001;45:15–32.

Breiman L, Friedman J, Stone C, Olshen RA. Classification and regression trees. Chapman & Hall/CRC; 1984.

Cappelle J, Girard O, Fofana B, Gaidet N, Gilbert M. Ecological modeling of the spatial distribution of wild waterbirds to identify the main areas

where avian influenza viruses are circulating in the Inner Niger Delta, Mali. EcoHealth 2010:1–11.
Chamaillé L, Tran A, Meunier A, Bourdoiseau G, Ready P, Dedet J-P. Environmental risk mapping of canine leishmaniasis in France. Parasit Vectors 2010;3:31.
Clements ACA, Pfeiffer DU, Martin V. Application of knowledge-driven spatial modelling approaches and uncertainty management to a study of Rift Valley fever in Africa. Int J Health Geog 2006;5:57–69.
Clements ACA, Pfeiffer DU, Martin V, Pittliglio C, Best N, Thiongane Y. Spatial risk assessment of Rift Valley fever in Senegal. Vector Borne Zoonotic Dis 2007;7:203–16.
Colacicco-Mayhugh MG, Masuoka PM, Grieco JP. Ecological niche model of *Phlebotomus alexandri* and *P. papatasi* (Diptera: Psychodidae) in the Middle East. Int J Health Geog 2010;9:2.
de Souza Muñoz M, De Giovanni R, de Siqueira M, Sutton T, Brewer P, Pereira R, et al. OpenModeller: a generic approach to species' potential distribution modelling. GeoInformatica 2009;15:111–35.
De'ath G. Boosted trees for ecological modeling and prediction. Ecology 2007;88:243–51.
De'ath G, Fabricius K. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 2000;81:3178–92.
DeJong K. Learning with genetic algorithms: an overview. Mach Learn 1988;3:121–38.
Dempster AP. New methods for reasoning towards posterior distributions based on sample data. Ann Math Stat 1966;37:355–74.
Dempster AP. Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 1967;38:325–39.
Eastman JR. GIS and uncertainty management: new directions in software development. Lurralde 1997;20:53–66.
Eastman JR. IDRISI andes: guide to GIS and image processing. Worcester, MA: Clark Labs, Clark University; 2006.
Eastman JR, Jin W, Kyem PAK, Toledano J. Raster procedures for multi-criteria/multi-objective decisions. Photogramm Eng Remote Sensing 1995;61:539–47.
Elith J, Graham CH. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 2009;32:66–77.
Elith J, Graham C, Anderson R, Dudik M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography 2006;29:129–51.
Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol 2008;77:802–13.
Elnaggar AA, Noller JS. Application of remote-sensing data and decision-tree analysis to mapping salt-affected soils over large areas. Remote Sens 2010;2:151–65.
Engler R, Guisan A, Rechsteiner L. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J Appl Ecol 2004;41:263–74.
Estrada-Pena A, Venzal JM. Climate niches of tick species in the Mediterranean region: modeling of occurrence data, distributional constraints and impact of climate change. J Med Entomol 2007;44:1130–8.
Fitzpatrick JM, Grefenstette JJ. Genetic algorithms in noisy environments. Mach Learn 1988;3:101–20.
Franklin J. Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. Prog Phys Geog 1995;19:474–99.
Franklin J. Model evaluation. In: Mapping species distributions: spatial inference and prediction. New York: Cambridge University Press; 2009a. p. 209–34.
Franklin J. Machine learning methods. In: Mapping species distributions: spatial inference and prediction. New York: Cambridge University Press; 2009b. p. 155–79.
Franklin J. Implementation of species distribution models. In: Mapping species distributions: spatial inference and prediction. New York: Oxford University Press; 2009c. p. 235–61.
Franklin J. Classification, similarity, and other methods for presence-only data. In: Mapping species distribution: spatial inference and prediction. New York: Cambridge University Press; 2009d. p. 180–205.
Freund Y, Schapire R. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the 13th International Conference; San Francisco. 1996.
Gonzalez C, Wang O, Strutz SE, Gonzalez-Salazar C, Sanchez-Cordero V, Sarkar S. Climate change and risk of leishmaniasis in North America: predictions from ecological niche models of vector and reservoir species. PLoS Negl Trop Dis 2010;4:e585.
Grinnell J. Field tests of theories concerning distributional control. Am Nat 1917;51:115–28.
Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecol Modell 2000;135:147–86.
Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. New York: Springer-Verlag; 2009.
Hernandez P, Franke I, Herzog S, Pacheco V, Paniagua L, Quintana H, et al. Predicting species distributions in poorly-studied landscapes. Biodivers Conserv 2008;17:1353–66.
Hirzel A, Arlettaz R. Modeling habitat suitability for complex species distributions by environmental-distance geometric mean. Environ Manage 2003;32:614–23.
Hirzel A, Hausser J, Chessel D, Perrin N. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 2002;83:2027–36.
Hirzel A, Hausser J, Perrin N. Biomapper 1.0-4.0. Lab of Conservation Biology 2007.
Hogerwerf L, Wallace R, Ottaviani D, Slingenbergh J, Prosser D, Bergmann L, et al. Persistence of highly pathogenic avian influenza H5N1 virus defined by agro-ecological niche. EcoHealth 2010:213–25.
Hutchinson G. Homage to Santa Rosalia, or why are there so many kinds of animals. Am Nat 1959;93:145–59.
Ivanek R, Grohn YT, Wells MT, Lembo Jr AJ, Sauders BD, Wiedmann M. Modeling of spatially referenced environmental and meteorological factors influencing the probability of *Listeria* species isolation from natural environments. Appl Environ Microbiol 2009;75:5893–909.
Joyner TA, Lukhnova L, Pazilov Y, Temiralyeva G, Hugh-Jones ME, Aikimbayev A, et al. Modeling the potential distribution of *Bacillus anthracis* under multiple climate change scenarios for Kazakhstan. PLoS ONE 2010;5:e9596.
Kulkarni MA, Desrochers RE, Kerr JT. High resolution niche models of malaria vectors in Northern Tanzania: a new capacity to predict malaria risk? PLoS ONE 2010;5:e9396.
Larson SR, DeGroote JP, Bartholomay LC, Sugumaran R. Ecological niche modeling of potential West Nile virus vector mosquito species in Iowa. J Insect Sci 2010;10:1–17.
Lawler JJ, White D, Neilson RP, Blaustein AR. Predicting climate-induced range shifts: model differences and model reliability. Glob Change Biol 2006;12:1568–84.
Lawson AB. Statistical methods in spatial epidemiology. Chichester, England: John Wiley & Sons, Ltd; 2006.
Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Mar Ecol Prog Ser 2006;321:267–81.
Levine R, Peterson A, Benedict M. Geographic and ecologic distributions of the *Anopheles gambiae* complex predicted using a genetic algorithm. Am J Trop Med Hyg 2004a;70:105–9.
Levine R, Peterson A, Benedict M. Distribution of members of *Anopheles quadrimaculatus* say s.l. (Diptera: Culicidae) and implications for their roles in malaria transmission in the United States. J Med Entomol 2004b;41:607–13.
Liu C, Berry PM, Dawson TP, Pearson RG. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 2005;28:385–93.
Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeog 2008;17:145–51.
Maher SP, Ellis C, Gage KL, Enscore RE, Peterson AT. Range-wide determinants of plague distribution in North America. Am J Trop Med Hyg 2010;83:736–42.
Mak S, Klinkenberg B, Bartlett K, Fyfe M. Ecological niche modeling of *Cryptococcus gattii* in British Columbia, Canada. Environ Health Perspect 2009;118:653–8.
Mak S, Morshed M, Henry B. Ecological niche modeling of Lyme Disease in British Columbia, Canada. J Med Entomol 2010;47:99–105.
Malczewski J. GIS and multicriteria decision analysis. New York: John Wiley & Sons, Inc; 1999.
Malczewski J. On the use of weighted linear combination method in GIS: common and best practice approaches. Trans GIS 2000;4:5–22.
Malczewski J. GIS-based land-use suitability analysis: a critical overview. Prog Plann 2004;62:3–65.
Malczewski J. Ordered weighted averaging with fuzzy quantifiers: GIS-based multicriteria evaluation for land-use suitability analysis. Int J App Earth Obs Geoinf 2006;8:270–7.
Moffett A, Shackelford N, Sarkar S. Malaria in Africa: vector species' niche models and relative risk maps. PLoS ONE 2007;2:e824.

Neerinckx S, Peterson A, Gulinck H, Deckers J, Leirs H. Geographic distribution and ecological niche of plague in sub-Saharan Africa. Int J Health Geog 2008;7:54.
Nieto P, Malone JB, Bavia ME. Ecological niche modeling for visceral leishmaniasis in the state of Bahia, Brazil, using genetic algorithm for rule-set prediction and growing degree day-water budget analysis. Geospat Health 2006;1:115–26.
Olden JD, Lawler JJ, Poff NL. Machine learning methods without tears: a primer for ecologists. Quart Rev Biol 2008;83:171–93.
Ortiz-Pelaez A, Pfeiffer D. Use of data mining techniques to investigate disease risk classification as a proxy for compromised biosecurity of cattle herds in Wales. BMC Vet Res 2008;4:24.
Pearce JL, Boyce MS. Modelling distribution and abundance with presence-only data. J Appl Ecol 2006;43:405–12.
Pearson R, Thuiller W, Arajo M, Martinez-Meyer E, Brotons L, McClean C, et al. Model-based uncertainty in species range prediction. J Biogeog 2006;33:1704–11.
Peterson A. Ecological niche modeling and spatial patterns of disease transmission. Emerg Infect Dis 2006;12:1822–6.
Peterson AT. Ecological niche modelling and understanding the geography of disease transmission. Vet Ital 2007;43:393–400.
Peterson A, Shaw J. Lutzomyia vectors for cutaneous leishmaniasis in Southern Brazil: ecological niche models, predicted geographic distributions, and climate change effects. Int J Parasitol 2003;33:919–31.
Peterson AT, Pereira RS, Neves VFDC. Using epidemiological survey data to infer geographic distributions of leishmaniasis vector species. Rev Soc Bras Med Trop 2004;37:10–4.
Peterson A, Papes M, Soberon J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol Modell 2008;213:63–72.
Pfeiffer DU, Robinson TP, Stevenson M, Stevens KB, Rogers DJ, Clements ACA. Spatial risk assessment and management of disease. In: Spatial analysis in epidemiology. Oxford University Press; 2008a. p. 110–9.
Pfeiffer DU, Robinson TP, Stevenson M, Stevens KB, Rogers DJ, Clements ACA. Spatial variation in risk. In: Spatial analysis in epidemiology. Oxford: Oxford University Press; 2008b. p. 67–80.
Phillips SJ, Dudík M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 2008;31:161–75.
Phillips SJ, Dudik M, Schapire RE. A maximum entropy approach to species distribution modeling. Proceedings of the 21st international conference on machine learning; New York. 2004.
Phillips S, Anderson R, Schapier R. Maximum entropy modeling of species geographic distributions. Ecol Modell 2006;190:231–59.
Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol App 2009;19:181–97.
Prasad A, Iverson L, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 2006;9:181–99.
Rakotomanana F, Randremanana R, Rabarijaona L, Duchemin J, Ratovonjato J, Ariey F, et al. Determining areas that require indoor insecticide spraying using multi criteria evaluation, a decision-support tool for malaria vector control programmes in the Central Highlands of Madagascar. Int J Health Geog 2007;6:2.
Refsgaard JC, van der Sluijs JP, Brown J, van der Keur P. A framework for dealing with uncertainty due to model structure error. Adv Water Resour 2006;29:1586–97.
Refsgaard JC, van der Sluijs JP, Hojberg AL, Vanrolleghem PA. Uncertainty in the environmental modelling process – a framework and guidance. Environ Modell Softw 2007;22:1543–56.
Ridgeway G. The state of boosting. Comput Sci Stat 1999;31:172–81.
Saaty TL. The analytic hierarchy process. New York, NY: McGraw-Hill; 1980.
Saaty TL. Multicriteria decision making: the analytic hierarchy process. Pittsburgh, PA: RWS Publications; 1990.
Sarkar S, Strutz SE, Frank DM, Rivaldi C-L, Sissel B, Sanchez-Cordero V. Chagas disease risk in Texas. PLoS Negl Trop Dis 2010;4:e836.
Segurado P, Araujo M. An evaluation of methods for modelling species distributions. J Biogeog 2004;31:1555–68.
Sinka M, Bangs M, Manguin S, Coetzee M, Mbogo C, Hemingway J, et al. The dominant Anopheles vectors of human malaria in Africa, Europe, the Middle East: occurrence data, distribution maps, bionomic precis. Parasit Vectors 2010;3:117.
Soberon J, Peterson A. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiv Inform 2005;2:1–10.
Steele K, Carmel Y, Cross J, Wilcox C. Uses and misuses of multicriteria decision analysis (MCDA) in environmental decision making. Risk Anal 2009;29:26–33.
Stockwell D, Noble I. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Math Comput Simul 1992;33:385–90.
Stockwell D, Peters D. The GARP modelling system: problems and solutions to automated spatial prediction. Int J Geog Inf Sci 1999;13:143–58.
Stockwell DRB, Peterson AT. Controlling bias during predictive modelling with museum data. In: Scott JM, Heglund PJ, Morrison M, Raphael M, Haufler J, Wall B, editors. Predicting species occurrences: issues of scale and accuracy. Covello, CA: Island Press; 2002a. p. 537–46.
Stockwell DRB, Peterson AT. Effects of sample size on accuracy of species distribution models. Ecol Modell 2002b;148:1–13.
Stockwell DRB, Beach JH, Stewart A, Vorontsov G, Vieglais D, Pereira RS. The use of the GARP genetic algorithm and internet grid computing in the Lifemapper world atlas of species biodiversity. Ecol Modell 2006;195:139–45.
Sutton CD, Rao EJW, Solka JL. Classification and regression trees, bagging, and boosting. In: Handbook of statistics. Elsevier; 2005. p. 303–29.
Sweeney AW, Beebe NW, Cooper RD. Analysis of environmental factors influencing the range of anopheline mosquitoes in northern Australia using a genetic algorithm and data mining methods. Ecol Modell 2007;203:375–86.
Thang N, Erhart A, Speybroeck N, Hung L, Thuan L, Hung C, et al. Malaria in central Vietnam: analysis of risk factors by multivariate analysis, classification tree models. Malar J 2008;7:28.
Tsoar A, Allouche O, Steinitz O, Rotem D, Kadmon R. A comparative evaluation of presence-only methods for modelling species distribution. Divers Distrib 2007;13:397–405.
Williams RA, Peterson AT. Ecology and geography of avian influenza (HPAI H5N1) transmission in the Middle East and northeastern Africa. Int J Health Geogr 2009;8:47.
Williams R, Fasina F, Peterson A. Predictable ecology and geography of avian influenza (H5N1) transmission in Nigeria and West Africa. Trans R Soc Trop Med Hyg 2008;102:471–9.
Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, et al. Effects of sample size on the performance of species distribution models. Divers Distrib 2008;14:763–73.