Medical and Veterinary Entomology (2008) **22**, 248–257

An assessment of the effect of data partitioning on the performance of modelling algorithms for habitat suitability for ticks

A. ESTRADA-PEÑA[1] and W. THUILLER[2]

[1]Department of Parasitology, Veterinary Faculty, University of Zaragoza, Zaragoza, Spain and [2]Alpine Ecology Laboratory, University Joseph Fourier, Grenoble, France

Abstract. A comparison of the performance of five modelling methods using presence/absence (generalized additive models, discriminant analysis) or presence-only (genetic algorithm for rule-set prediction, ecological niche factor analysis, Gower distance) data for modelling the distribution of the tick species *Boophilus decoloratus* (Koch, 1844) (Acarina: Ixodidae) at a continental scale (Africa) using climate data was conducted. This work explicitly addressed the usefulness of clustering using the normalized difference vegetation index (NDVI) to split original records and build partial models for each region (cluster) as a method of improving model performance. Models without clustering have a consistently lower performance (as measured by sensitivity and area under the curve [AUC]), although presence/absence models perform better than presence-only models. Two cluster-related variables, namely, prevalence (commonness of tick records in the cluster) and marginality (the relative position of the climate niche occupied by the tick in relation to that available in the cluster) greatly affect the performance of each model ($P < 0.05$). Both sensitivity and AUC are better for NDVI-derived clusters where the tick is more prevalent or its marginality is low. However, the total size of the cluster or its fragmentation (measured by Shannon's evenness index) did not affect the performance of models. Models derived separately for each cluster produced the best output but resulted in a patchy distribution of predicted occurrence. The use of such a method together with weighting procedures based on prevalence and marginality as derived from populations at each cluster produced a slightly lower predictive performance but a better estimation of the continental distribution of the tick. Therefore, cluster-derived models are able to effectively capture restricting conditions for different tick populations at a regional level. It is concluded that data partitioning is a powerful method with which to describe the climate niche of populations of a tick species, as adapted to local conditions. The use of this methodology greatly improves the performance of climate suitability models.

Key words. Accuracy, climate niche modelling, clustering, marginality, predictive prevalence, ticks.

Introduction

The basic concept underlying species occurrence modelling is the definition of the ecological niche: each species is found within a specific range of environmental variables which enable individuals to survive and reproduce (Austin, 2002). Species occurrence can be predicted through the identification of appropriate environmental variables, commonly referred to as habitat suitability models (HSM; Guisan & Thuiller, 2005): the relationships are generalized from a sample of observations where

Correspondence: Professor A. Estrada-Peña, Department of Parasitology, Veterinary Faculty, Miguel Servet 177, 50013 Zaragoza, Spain. Tel.: +34 976 761558; Fax: + 34 976 761612; E-mail: aestrada@unizar.es

species presence is matched with specific values for the environmental variables. This concept has been increasingly applied to the predictive mapping and ecological determinants of the distribution of disease vectors such as insects (e.g. Rogers *et al.*, 1996) or ticks (e.g. Cumming, 2000). Concerns about the impact of forecast climate change on the distribution of disease vectors indicate a need to evaluate the influence of various techniques on the modelling process and the final output.

Data availability is a major constraint in building large-scale models of species distribution (Osborne *et al.*, 2001) as inductive modelling requires a large amount of optimally assessed information in order to predict species occurrence (Hirzel & Guisan, 2002). Although vast stores of presence-only data exist for health-threatening arthropods, absence data are rarely available, or they may be of questionable value in many situations. Algorithms that use presence/absence data better fit the expected distribution of a given organism. Alternatively, and at the cost of restrictive assumptions, absence data may be generated in the form of pseudo-absences which refer to areas for which no definite information regarding species occurrence is available and which are therefore assumed to be unsuitable in terms of provision of statistical data upon which analyses may be based (Wintle *et al.*, 2005; Guisan *et al.*, 2006).

As larger areas are modelled, it is highly likely that heterogeneity in the predictor variables will increase. There are indications that the performance of models is affected by several species-specific geographic attributes, such as latitudinal range, marginality, prevalence in the ecological sense and rarity (Brotóns *et al.*, 2004; Segurado & Araújo, 2004; Luoto *et al.*, 2005). Therefore, variables restricting the distribution of a tick in a given area may have different roles in geographically separate sites. Although different modelling techniques have been used for the prediction of tick distribution, none, to our knowledge, have investigated systematically how variation in the geographical distribution of the target species and the use of different portions of the environmental niche affect modelling outcomes.

The purpose of this study was to compare the performance of different models in predicting the distribution of the tick *Boophilus decoloratus* (Koch, 1844) in Africa based on presence/absence or presence-only data, as a baseline for further developments in modelling tick distribution. Particular attention was devoted to exploiting the distribution of the tick in ecological clusters (a technique known as data partitioning) in an attempt to build more accurate, partial models.

Materials and methods

The dataset

The distribution of *B. decoloratus* in Africa has been documented since the beginning of the 20th century. For this study, a dataset with relatively recent, geo-registered and accurate tick records which allow for comparison with contemporary climate was used. This study was based on the compilation of previously published records of economically important African ticks (International Consortium of Ticks and Tick-borne Diseases, 2004). A total of 1304 records of presence, recorded between 1970 and 2000, were selected as suitable for this study (Fig. 1). These records represent accurate determinations and unambiguously refer to pairs of co-ordinates. The basic geographical unit for this work was a 5 × 5-km cell. The tick was considered as present in the cell if at least one record fitted within the limits of that cell.

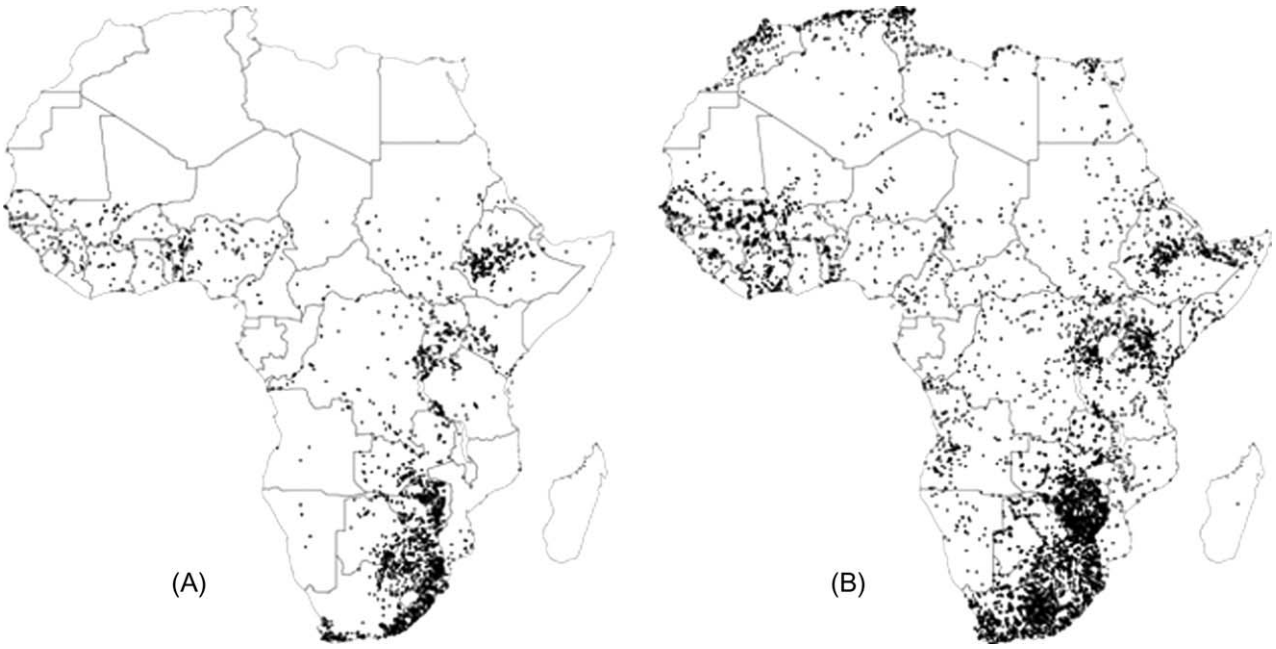

Fig. 1. Map of Africa, displaying (A) presence records (point localities) for *Boophilus decoloratus* and (B) presence records for other tick species of domestic animals over the region of study, used as negative sites for the target species.

A set of absence data was obtained using the following approach. The complete set of tick records for Africa including all species of veterinary interest was used. The tick was considered to be absent in a given cell if records existed for other tick species in the cell. Figure 1 shows the spatial distribution of records used in the current study.

Clustering

The main goal of this study was to determine if the building of partial models for geographically separate populations of tick species could have a role in improving model performance. Thus, the complete set of records was used as a whole (herein called 'complete models') or split into subsets of tick records collected within separate regions ('cluster models').

A multivariate clustering method was used to classify the habitat into categories. Multivariate clustering based on maps of abiotic variables has been used previously to produce a spectrum of quantitative eco-regions for vegetation (e.g. Lobo *et al.*, 1997; Hargrove & Hoffman, 2005). These statistically generated eco-regions capture regional environmental differences summarizing local conditions in terms of gradients and clines. The environmental data selected to produce the eco-regions referred to monthly normalized difference vegetation index (NDVI) values, at 1-km resolution, obtained between 1992 and 2002. Principal components analysis (PCA) was performed on the image composed of the 12 monthly NDVI values to reduce variability and correlation between monthly variables. Then, an unsupervised classification was performed on the values of the three first principal components (reported as significant in the matrix of loadings) to produce a set of clusters. Mahalanobis distance was used as a measure of dissimilarity and the weighted pair-group average was selected as an amalgamation method to produce the clusters, maximizing the distance between cluster centroids and minimizing the distance within points of the same cluster. Tick records collected 'within' each cluster were considered to pertain to that particular cluster and were modelled separately. ERDAS IMAGINE software (Erdas, Inc., Norcross, GA, U.S.A.) was used for all statistical procedures.

Predictor variables for tick habitat suitability

A grid-based dataset at a resolution of 5 km was used, in accordance with the grid for presence/absence tick data. The following 19 variables, obtained from the WorldClim dataset (Museum of Vertebrate Zoology, Berkeley, CA, U.S.A.), were used for predictive mapping:

(1) annual mean temperature;
(2) mean diurnal range (mean of monthly [maximum temperature−minimum temperature]);
(3) isothermality (2/7*100);
(4) temperature seasonality (standard deviation*100);
(5) maximum temperature in the warmest month;
(6) minimum temperature in the coldest month;
(7) temperature annual range (item 5 minus item 6);
(8) mean temperature for the wettest quarter;
(9) mean temperature for the driest quarter;
(10) mean temperature for the warmest quarter;
(11) mean temperature for the coldest quarter;
(12) annual precipitation;
(13) precipitation for the wettest month;
(14) precipitation for the driest month;
(15) precipitation seasonality (coefficient of variation);
(16) precipitation for the wettest quarter;
(17) precipitation for the driest quarter;
(18) precipitation for the warmest quarter, and
(19) precipitation for the coldest quarter.

All 19 variables were initially entered for every model. The best subset of variables was then selected by backwards and forwards substitution and elimination of every combination of variables. The influence of these climate conditions on the geographical range of the studied tick has been reported previously (Estrada-Peña *et al.*, 2006a).

Models

It is outwith the scope of this paper to evaluate every model capable of producing HS-predictive maps. Five different models commonly used to derive HS maps were tested. Two approaches that rely on presence/absence data (the generalized additive model [GAM], non-linear discriminant analysis [DA]) and three that use only presence data (the genetic algorithm for rule-set prediction [GARP], ecological niche factor analysis [ENFA], the Gower metric) were selected. A summarized description of each model follows. Readers should refer to the bibliography for exhaustive information about these procedures.

Generalized additive models implemented in the generalized regression analysis and spatial prediction (GRASP) framework (Lehmann *et al.*, 2003) were tested. The GAM is a non-parametric extension of the commonly used generalized linear model and makes no assumption on the form of the species' response curves. It has been successfully applied for different species and contexts (Guisan *et al.*, 2002; Araújo *et al.*, 2005; Thuiller *et al.*, 2006).

Discriminant analysis is a broad class of methods concerned with the development of rules for assigning unclassified objects/specimens to previously defined groups. A linear or non-linear discriminant function is used to assign an observation to one of a set of groups taking a vector of observations from a specimen and multiplying it by a vector of coefficients to produce a score which is used to classify the specimen as belonging to a group. Non-linear DA, as used here, has been used previously to predict the distribution of tsetse flies (Rogers *et al.*, 1996; Robinson *et al.*, 1997).

The GARP (Stockwell & Peters, 1999) was tested as a presence-only model. The GARP evaluates non-random associations between environmental characteristics of localities of known occurrence vs. those of the overall study region to produce a heterogeneous rule-set characterizing the species' ecological requirements. The GARP is designed to work with presence-only data; absence information is included via sampling of

pseudo-absence points. The change in predictive accuracy from one iteration to the next is used to evaluate whether a particular rule should be incorporated into the model, and the algorithm runs until convergence. The GARP has been extensively used for predicting the colonization success of invasive species and HS maps of disease vectors (Peterson, 2003).

Ecological niche factor analysis is an approach that uses presence-only data in a multivariate space of environmental variables (Hirzel *et al.*, 2002). This technique is based upon computation of the factors explaining the major part of species environmental distribution. An HS index for each cell is produced as a value inversely proportional to the weighted mean distance of the cell to the median of each ENFA factor, normalized in such a way that the suitability index ranges from 0 to 1. This type of analysis has been successfully applied for habitat assessment of some species (Brotóns *et al.*, 2004; Engler *et al.*, 2004). One of the niche measures derived from ENFA in subsequent calculations (see below) was used. The first component of ENFA factorial evaluation (called the marginality factor) explains how far the species' optimum environment is from the average environmental conditions (hereafter called 'global data') defined by all cells not previously excluded.

The Gower metric procedure assigns each cell in the output layer an average multidimensional similarity index pertaining to that cell and the closest presence cell in the training set (Carpenter *et al.*, 1993) using presence-only data. The higher the output (in the range of 0–1), the higher the similitude between the point and the set of actual captures in the training set, and hence the higher the suitability of the range of climate values for the tick species. Gower metric has not been extensively used recently, but it remains a useful approach for presence-only data (Miles *et al.*, 2005; Pearson *et al.*, 2006).

Development and evaluation of the models

Models were developed with a random training set of tick records (cells) and checked against an evaluation set (50% of records each). The whole set of climate layers was initially entered to train the models. In cluster-derived models, training and evaluation sets were derived for each cluster. Cluster models were used to compute the HS for the target tick species in three different ways. In the first, HS was computed separately for each single cluster and then predictive maps were 'patched' together. The second approach evaluated the HS for the whole territory with every partial model derived from each cluster and then averaged the final HS as the mean of the output of every cluster model ('cluster-averaged' models). The third was intended to produce a weighted output according to the size of each cluster (R = percentage of the total area in the zone of study), the prevalence (P) of the tick in the cluster (percentage of cells in the grid with confirmed tick presence within the cluster) and the marginality (M) of tick distribution in the cluster (obtained from ENFA, as described above). Thus, the weighting factor (Estrada-Peña *et al.*, 2006b) has the form:

$$W = (R \times P)/M.$$

The HS in this case was built according to the equation:

$$HS = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i H_i$$

where HS is the habitat suitability for the focal cell, H_i is the value obtained for that cell from the partial coefficient as obtained from *i*th cluster, and w_i is the weight assigned to the *i*th cluster. This approach is referred to here as 'cluster-weighted'.

Model performance using both presences and absences (generated as described above) from the tick database was assessed, even if the algorithm used presence-only data. The evaluation of performance measures first required the derivation of matrices of confusion that identified true positive, true negative, false positive and false negative. Sensitivity is based on the concept of true-presences misclassification and is calculated as 100% of false negatives. From the confusion matrix we calculated the area under the curve (AUC) of a receiver operating characteristic (ROC) plot of sensitivity against (1−specificity) (Swets, 1988). Sensitivity is defined as the proportion of true positives correctly predicted, whereas specificity is the proportion of true negatives correctly predicted (Fielding & Bell, 1997). The AUC was obtained from a customized function in s-plus software (Insightful Corp., Seattle, WA, U.S.A.).

Influence of cluster features on modelling performance

The effects of tick prevalence and marginality at each individual cluster, together with the effects of cluster size and its fragmentation (number of geographically separate patches conforming to a single cluster) as defined using Shannon's evenness index (Turner, 1990), on the performance of each model were tested by means of Spearman's rank correlation at cluster level. Demonstration of the existence of different limiting factors in the climate suitability for the tick at each individual cluster was also important. Climate variables ranked by each model as most important in predicting suitability for the tick at cluster level were considered the most limiting. A canonical analysis was performed between clusters and the most limiting climate factor for the tick in each cluster. Each cluster was plotted according to its position within the two first main NDVI-derived PCA axes, together with the position of the limiting climate factors, obtained from the canonical analysis, to facilitate understanding of the causes affecting environmental suitability for the tick within each cluster.

Results

A total of 25 categories, including water, were obtained through PCA analysis of monthly NDVI values. Figure 2 shows the geographical range of each cluster over the African continent. The first principal axis was strongly related to total NDVI and the second principal axis to NDVI variability (seasonality). The third axis was loaded with NDVI values for the period April–September. Together, the first and second axes accounted for 82% of total variability and the three axes together accounted for 91% of variability. Figure 3 shows the prevalence and marginality of

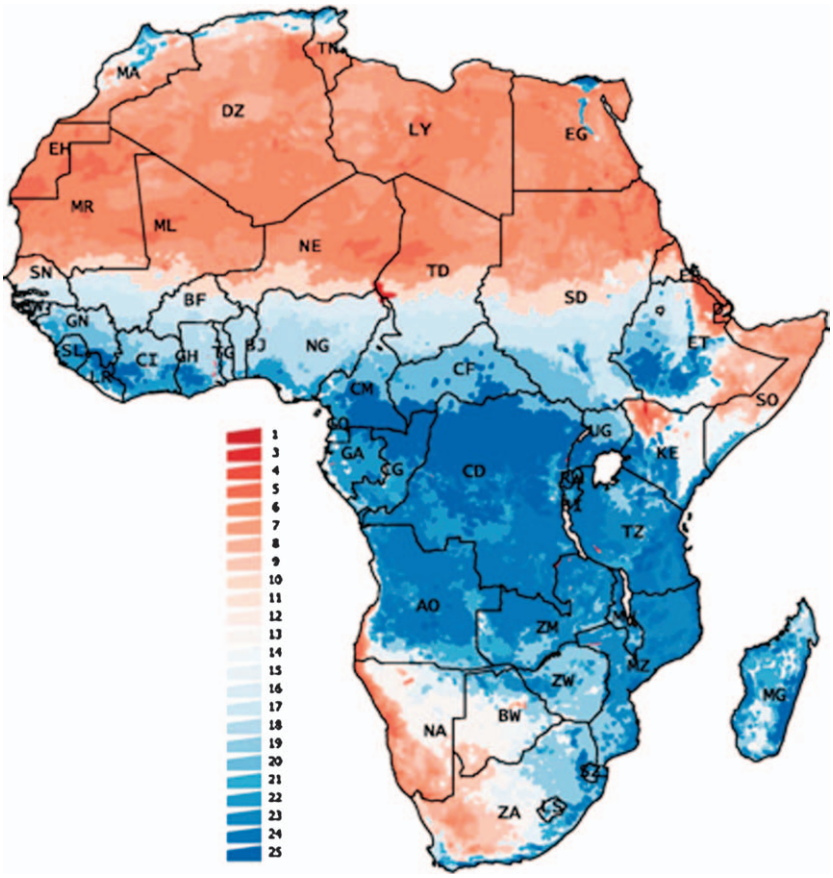

Fig. 2. Arbitrary colours showing the clusters detected over the region of study according to normalized difference vegetation index seasonal dynamics and classification of the three main axes of principal components analysis.

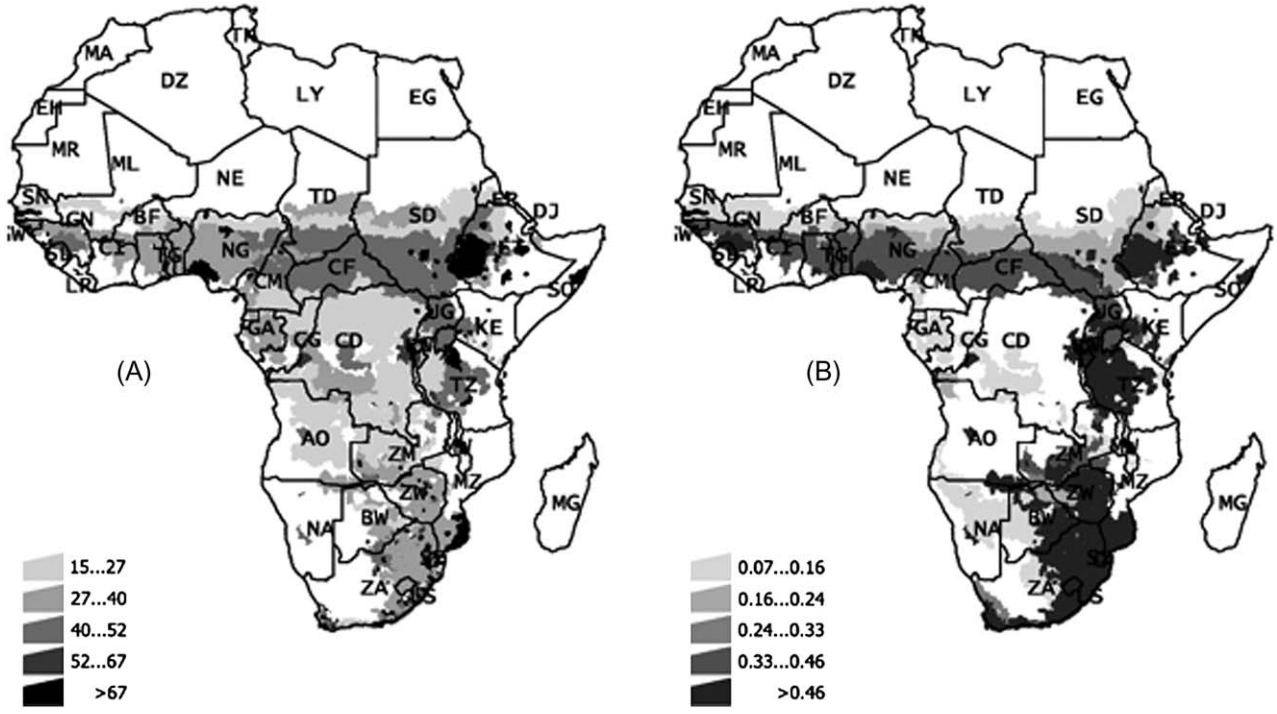

Fig. 3. (A) Prevalences of *Boophilus decoloratus* by cluster, obtained as the number of positive cells and total number of cells in the cluster. (B) Levels of marginality of the tick in each cluster, obtained from ecological niche factor analysis.

Table 1. Sensitivity and area under the curve of the modelling algorithms for complete models and for cluster-based models (single, averaged and weighted).

	Complete		Single cluster		Averaged cluster		Weighted cluster	
Model	Sens	AUC	Sens	AUC	Sens	AUC	Sens	AUC
GAM	0.74	0.61	0.73	0.92	0.74	0.83	0.67	0.89
DA	0.64	0.66	0.72	0.90	0.75	0.87	0.77	0.90
GARP	0.69	0.71	0.62	0.80	0.69	0.75	0.48	0.78
ENFA	0.59	0.58	0.71	0.84	0.69	0.76	0.55	0.78
Gower	0.61	0.69	0.69	0.90	0.70	0.83	0.62	0.86

Sens, sensitivity; AUC, area under the curve; GAM, generalized additive model; DA, discriminant analysis; GARP, genetic algorithm for rule-set prediction; ENFA, ecological niche factor analysis; Gower, Gower distance.

B. decoloratus in the clusters. The highest tick prevalence was detected in wide areas of central Africa, the mountains of Ethiopia and in western Africa, corresponding with vegetation categories 18, 20, 22 and 25. Marginality was negatively correlated with these areas of highest prevalence. Marginality was lowest in a wide strip across the continent in central Africa. The highest area of marginality corresponded to a narrow strip in the Sahelian transition zone.

In general terms, complete models produced a lower sensitivity and predictive performance (AUC) than cluster-derived models (Table 1). Best performance in complete models was obtained in the GARP and Gower metric models. Performance was greatly enhanced using cluster-derived models, but different results were obtained according to the clustering method. In single clusters, GAM performed better, whereas GAM and DA produced the highest AUCs in averaged and weighted clusters. In the whole set of cluster models, the single-cluster method provided the best modelling approach for all the algorithms and the weighted approach provided the next best.

In complete models (Fig. 4, upper row), both GAM and GARP clearly overestimated tick distribution, performing with high sensitivity but at the cost of many false positives. However, DA and ENFA underestimated tick distribution in some parts of its range. The Gower metric procedure overestimated tick distribution in its southern range, but produced an adequate result for central regions of Africa. There was a drastic change in the output of cluster-weighted models compared with other methodologies. Figure 4 (lower row) shows the visual output of cluster-weighted models, with better agreement between actual and predicted distribution as detected by AUC values for most of the algorithms.

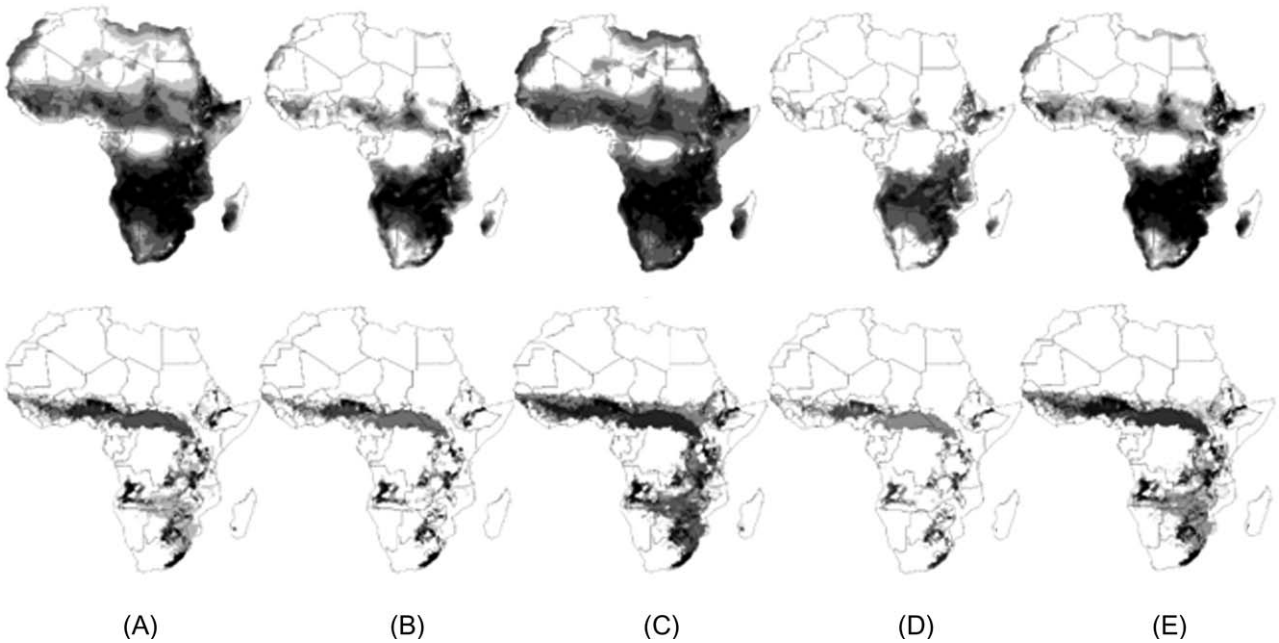

Fig. 4. Visual output of habitat suitability obtained from the different modelling algorithms in the prediction of the distribution of *Boophilus decoloratus* in Africa. (A) Generalized additive model. (B) Discriminant analysis. (C) Genetic algorithm for rule-set prediction. (D) Ecological niche factor analysis. (E) Gower distance. The upper row shows the complete models (developed, tested and performed on the whole set of positive and negative records). The lower row shows outputs obtained using the cluster-weighted approach.

Table 2. Values of total area (in sq km), prevalence and marginality of *Boophilus decoloratus*, and Shannon's evenness index for each normalized difference vegetation index-derived cluster. Included are the sensitivity and area under the curve of each model as obtained separately for each cluster.

Cluster	Area	Prevalence	Marginality	Shannon's index	Sens-GAM	Sens-DA	Sens-GARP	Sens-ENFA	Sens-GOWER	AUC-GAM	AUC-DA	AUC-GARP	AUC-ENFA	AUC-GOWER
1	338 624	22.96	0.50	0.40	0.73	0.64	0.83	0.78	0.79	0.60	0.74	0.62	0.81	0.69
3	61 440	30.82	0.35	0.17	0.70	0.78	0.77	0.78	0.88	0.68	0.73	0.71	0.70	0.75
4	75 584	15.11	0.72	0.22	0.64	0.67	0.72	0.69	0.61	0.71	0.66	0.57	0.53	0.75
5	1 355 776	12.35	0.98	0.40	0.71	0.77	0.80	0.38	0.73	0.66	0.73	0.75	0.71	0.74
6	5 682 880	3.08	4.14	0.57	0.66	0.54	0.71	0.82	0.66	0.62	0.70	0.65	0.57	0.56
7	3 559 488	6.43	1.95	0.74	0.70	0.70	0.65	0.78	0.88	0.65	0.64	0.64	0.69	0.50
8	1 121 344	3.40	3.54	0.55	0.67	0.51	0.58	0.70	0.60	0.92	0.63	0.58	0.54	0.53
9	555 072	5.93	1.98	0.28	0.62	0.63	0.67	0.76	0.73	0.57	0.60	0.49	0.61	0.57
10	1 128 064	17.03	0.71	0.18	0.82	0.72	0.81	0.93	0.87	0.74	0.75	0.72	0.80	0.62
11	13 184	30.77	0.33	0.5	0.70	0.74	0.71	0.81	0.63	0.67	0.77	0.70	0.67	0.57
12	206 528	26.27	0.43	0.28	0.74	0.70	0.86	0.87	0.85	0.66	0.76	0.65	0.74	0.74
13	886 208	13.98	0.85	0.31	0.77	0.77	0.95	0.87	0.79	0.62	0.65	0.70	0.64	0.59
14	500 608	22.88	0.51	0.15	0.83	0.86	0.77	0.93	0.88	0.82	0.57	0.80	0.81	0.62
15	1 015 744	28.32	0.42	0.71	0.73	0.85	0.61	0.86	0.83	0.69	0.84	0.77	0.78	0.80
16	1 086 400	28.47	0.42	0.22	0.94	0.78	0.54	0.89	0.76	0.90	0.86	0.93	0.65	0.81
17	1 254 272	30.00	0.40	0.20	0.83	0.95	0.91	0.97	0.88	0.92	0.73	0.83	0.76	0.77
18	1 333 056	35.54	0.34	0.29	0.92	0.88	0.85	0.93	0.91	0.84	0.98	0.91	0.83	0.65
19	1 462 912	30.15	0.40	0.47	0.80	0.84	0.99	0.94	0.93	0.77	0.88	0.81	0.79	0.65
20	1 446 400	44.23	0.27	0.24	0.97	0.99	0.95	0.94	0.75	0.84	0.96	0.85	0.95	0.87
21	1 081 088	29.46	0.41	0.76	0.79	0.97	0.86	0.92	0.88	0.78	0.89	0.73	0.72	0.58
22	1 213 376	32.84	0.37	0.44	0.81	0.96	0.80	0.91	0.68	0.70	0.89	0.67	0.77	0.59
23	1 447 360	23.28	0.52	0.48	0.82	0.87	0.86	0.94	0.95	0.85	0.82	0.74	0.73	0.75
24	1 892 288	29.70	0.41	0.52	0.95	0.83	0.85	0.98	0.74	0.74	0.93	0.84	0.86	0.62
25	1 575 488	39.66	0.31	0.33	0.89	0.98	0.98	0.97	0.96	0.85	0.93	0.86	0.71	0.99

Sens, sensitivity; AUC, area under the curve; GAM, generalized additive model; DA, discriminant analysis; GARP, genetic algorithm for rule-set prediction; ENFA, ecological niche factor analysis; Gower, Gower distance.

A relationship was observed between the performance of the modelling algorithms as individually obtained from the clusters of NDVI values and different variables associated with clusters (Table 2). All the algorithms performed less well when applied to areas of low tick prevalence or high marginality. Spearman's rank correlation (Table 3) shows that both sensitivity and AUC were significantly affected by the prevalence and marginality of the tick in a given cluster. However, model performance did not correlate with total area of the cluster or Shannon's index.

Discussion

Predictive modelling of the geographic distribution of health-threatening arthropods, based on environmental conditions, constitutes an important technique in analytical epidemiology. Considerable concern exists about the possible spread in response to forecast climate change of some arthropods and the diseases they carry (e.g. Sutherst, 2001). However, the weakest points in predicting distribution are data availability and the application of algorithms to increasingly larger regions (e.g. entire continents). Although vast stores of presence-only data exist, absence data are rarely available, especially for poorly sampled regions. Occurrence data for most tick species were recorded, in the best cases, through local sampling schemes and the great majority of these data consist of presence-only records from museum collections. The main problem with such occurrence data is that the intent and methods of collecting are rarely known, so that absences cannot be inferred with certainty. These data also have associated errors and biases, reflecting the frequently unsystematic manner in which samples were accumulated. Even when absence data are available, they may be of questionable value in many situations because sampling intensity in space and time may be non-random. All these reasons indicate the need to evaluate algorithms based on presence-only data because of the inherent problems in obtaining accurate data for absences for large regions.

Table 3. Spearman's rank correlation of sensitivity and area under the curve values of every model obtained individually for each normalized difference vegetation index-derived cluster, as affected by values of area of clusters, prevalence, and marginality of *Boophilus decoloratus,* and Shannon's evenness index in each cluster. Significant values are shown in bold.

	Model	Area	Prevalence	Marginality	Shannon's index
Sensitivity	GAM	0.064	**0.001**	**0.003**	0.514
	DA	0.191	**0.000**	**0.000**	0.945
	GARP	0.209	**0.013**	**0.021**	0.933
	ENFA	0.021	**0.002**	**0.005**	0.900
	Gower	0.238	0.064	0.086	0.831
AUC	GAM	0.189	**0.031**	**0.042**	0.651
	DA	0.105	**0.000**	**0.000**	0.478
	GARP	0.066	**0.001**	**0.001**	0.530
	ENFA	0.326	**0.005**	**0.009**	0.824
	Gower	0.923	**0.011**	**0.014**	0.153

AUC, area under the curve; GAM, generalized additive model; DA, discriminant analysis; GARP, genetic algorithm for rule-set prediction; ENFA, ecological niche factor analysis; Gower, Gower distance.

Models that work with both presence and absence data are better than those that use presence-only data. Brotóns *et al.* (2004) showed that predictions derived from generalized linear models (GLM) are more accurate than those from ENFA when accurate absence data are available. Zaniewski *et al.* (2002) argued that pure presence-only methods are more likely to predict potential distributions that more closely resemble the fundamental climate preferences of the species, whereas presence/absence modelling is more likely to reflect the present natural distributions derived from the realized niche. As presence-only methods do not take into account the areas from which the species might be absent, they are less conservative in estimating the niche. Methods based on presence-only data appear to fully cover habitat modelling when the main objective is to identify overall suitable areas for a given species (Pearson *et al.*, 2006). If a complete and accurate set of absences is available, and assumptions of equilibrium are not violated, presence/absence methods should be prioritized because they better capture the ecological relationships of the species with the niche (Brotóns *et al.*, 2004; Guisan & Thuiller, 2005).

This study argues that the dividing of tick records from a large region (a continent) into clusters is a useful tool to enhance model performance. Although the concept of intraspecific niche differentiation in the field of predictive mapping has been proposed previously (Osborne & Suárez-Seoane, 2002; Peterson & Holt, 2003; Guisan *et al.*, 2006; Murphy & Lovett-Doust, 2007), this study demonstrates its usefulness in assessing climate suitability for ticks with the detection of clusters based on ecological grounds. The use of NDVI as a proxy to build ecosystem units is increasingly applicable. Remote sensing is a valuable tool that can be used to describe the spatial heterogeneity of ecosystems functioning at regional and global scales (Lobo *et al.*, 1997). Biomes (clusters) are not predefined but emerge as distinct combinations of plant dominant types, which are governed by climatic parameters that, in turn, also regulate the lifecycle of ticks. This method has been used to reappraise ecoclimatic regions in the U.S.A. (Loveland *et al.*, 1991), Africa (Mayaux *et al.*, 2004) and Europe (Metzger *et al.*, 2005). It is most important that definitions based on remotely sensed data are objective and repeatable, and that these NDVI-derived values define features of the tick lifecycle (Estrada-Peña *et al.*, 2006a, 2006b). Reducing the number of variables while retaining the original variability of the whole dataset using PCA has also been reported as a useful tool in the management of NDVI values (Duchateau *et al.*, 1997; Metzger *et al.*, 2005) for incorporating variation into adequate environmental structures. Thus, NDVI-derived clusters of habitat are useful for dividing up habitat over wide regions and for describing the functional features of the respective clusters, while retaining an understanding of the habitat's overall impact on the tick lifecycle.

Some cluster features clearly affect model performance. Both sensitivity and AUC clearly depend on the prevalence and marginality of the tick within each cluster. It is obvious that prediction will have poor sensitivity in areas where the tick is uncommon. However, even if the tick is well represented in a cluster, the distance between the species' optimum conditions and prevailing environmental conditions in the cluster (as measured by marginality) will bias the algorithm's response. It is

interesting to note that algorithm performance depends on neither the size of the cluster nor its fragmentation. Furthermore, the most limiting climate variables for building the best predictive output for each cluster were defined differently by the various models. This both indicates the influence of different limiting variables in different parts of the geographical range of the tick and supports the use of cluster-derived models to capture the best set of local predictive conditions.

Clusters as defined here should be considered as dynamic entities that can shift in geographic space according to ecological forces operating on the tick population and longterm changes in climate. The effect of the ecological plasticity of the tick population to prevailing climate conditions requires critical assessment in terms of both tick ecology and predictive mapping and is considerably important in the invasive behaviour of some tick species (Estrada-Peña *et al.*, 2007). This methodology is also applicable to the tracking of longterm ecological changes caused by climate conditions. When absence data are available, both DA and GAM produce a better inference than algorithms based on presence-only data. In any case, spatial partitioning of the data is clearly necessary to improve predictions of models where regional niche variation occurs or for wide-ranging species. The inclusion of prevalence and marginality values in cluster-weighted models, although it reduces the final statistical output of the model, allows for deeper ecological significance.

Acknowledgements

The authors thank Sarah E. Randolph (Oxford University, U.K.), Petr Zeman and Milan Daniel (Institute of Health, Prague, Czech Republic) for their valuable comments on preliminary versions of this paper. This work was facilitated by the Integrated Consortium on Ticks and Tick-borne Diseases (ICTTD-3) and financed by the International Co-operation Programme of the European Union through Co-ordination Action Project no. 510561. WT acknowledges support from the EU FP6 MACIS species-targeted project (Minimization of and Adaptation to Climate Change: Impacts on Biodiversity, no: 044399) and the EU FP6 ECOCHANGE integrated project (Challenges in Assessing and Forecasting Biodiversity and Ecosystem Changes in Europe).

References

Araújo, M.B., Thuiller, W., Williams, P.H. & Reginster, I. (2005) Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecology and Biogeography*, **14**, 17–30.

Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Brotóns, L., Thuiller, W., Araújo, M.B. & Hirzel, A. (2004) Presence/absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants, animals. *Biodiversity and Conservation*, **2**, 667–680.

Cumming, G.S. (2000) Using between-model comparisons to fine-tune linear models of species ranges. *Journal of Biogeography*, **27**, 441–455.

Duchateau, L., Kruska, R.L. & Perry, B.D. (1997) Reducing a spatial database to its effective dimensionality for logistic-regression analysis of incidence of livestock disease. *Preventive Veterinary Medicine*, **32**, 207–218.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Estrada-Peña, A., Bouattour, A., Camicas, J.-L. et al. (2006a) The distribution and ecological preferences of ticks of the subgenus *Boophilus* in Africa and Latin America. *Experimental and Applied Acarology*, **38**, 219–235.

Estrada-Peña, A., Venzal, J.M. & Sánchez Acedo, C. (2006b) The tick *Ixodes ricinus*: distribution and climate preferences in the western Palearctic. *Medical and Veterinary Entomology*, **20**, 189–197.

Estrada-Peña, A., Pegram, R., Barré, N. & Venzal, J.M. (2007) Using invaded range data to model the climate suitability for *Amblyomma variegatum* (Acari: Ixodidae) in the New World. *Experimental and Applied Acarology*, **41**, 203–214.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Guisan, A., Edwards, T.J. & Hastie, T.J. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100.

Guisan, A., Broennimann, O., Engler, R., Yoccoz, N.G., Vust, M., Zimmermann, N.E. & Lehmann, A. (2006) Using niche-based models to improve the sampling of rare species. *Conservation Biology*, **20**, 501–511.

Hargrove, W.W. & Hoffman, F.M. (2005) Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environmental Management*, **34**(Suppl. 1), 39–60.

Hirzel, A.H. & Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling? *Ecological Modelling*, **157**, 331–341.

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

International Consortium of Ticks and Tick-borne Diseases (ICTTD) (2004) *Ticks of Veterinary Importance in Africa (CD-ROM)*. European Union Programme (contract no. ICA4-CT-2000-300069. ICTTD, Utrecht.

Lehmann, A., Overton, J.M. & Leathwick, J.R. (2003) GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling*, **160**, 165–183.

Lobo, A., Ibáñez Martí, J. & Carrera Giménez-Cassina, C. (1997) Regional scale hierarchical classification of temporal series of AVHRR vegetation series. *International Journal of Remote Sensing*, **18**, 3167–3193.

Loveland, T.R., Merchant, J.W., Brown, J.F., Ohlen, D.O., Reed, B.C., Olson, P. & Hutchinson, J. (2001) Seasonal land-cover regions of the United States. *Annals of the Association of American Geographers*, **85**, 339–355.

Luoto, M., Pöyry, J., Heikkinen, R.K. & Saarinen, K. (2005) Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, **14**, 575–584.

Mayaux, P., Bartholomé, E., Fritz, S. & Belward, A. (2004) A new land-cover map of Africa for the year 2000. *Journal of Biogeography*, **31**, 861–877.

Metzger, M.J., Bunce, R.G.H., Jongman, R.H.G., Mücher, C.A. & Watkins, J.W. (2005) A climatic stratification of the environment in Europe. *Global Ecology and Biogeography*, **14**, 549–563.

Miles, L., Grainger, A. & Phillips, O. (2005) The impact of global climate change on tropical forest biodiversity in Amazonia. *Global Ecology and Biogeography*, **13**, 553–565.

Murphy, H.T. & Lovett-Doust, J. (2007) Accounting for regional niche variation in habitat suitability models. *Oikos*, **116**, 99–110.

Osborne, P. & Suárez-Seoane, S. (2002) Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling*, **157**, 249–259.

Osborne, P., Alonso, J.C. & Bryant, R.G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, **38**, 458–471.

Pearson, R.G., Thuiller, W., Araújo, M.B. et al. (2006) Model-based uncertainty in species' range prediction. *Journal of Biogeography*, **33**, 1704–1711.

Peterson, A.T. (2003) Predicting the geography of species' invasions via ecological niche modelling. *Quarterly Review of Biology*, **78**, 419–433.

Peterson, A.T. & Holt, R.D. (2003) Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. *Ecological Letters*, **6**, 774–782.

Robinson, T., Rogers, D. & Williams, B. (1997) Univariate analysis of tsetse habitat in the common fly belt of southern Africa using climate and remotely sensed vegetation data. *Medical and Veterinary Entomology*, **11**, 223–234.

Rogers, D.J., Hay, S.I. & Packer, M.J. (1996) Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Annals of Tropical Medicine and Parasitology*, **90**, 225–241.

Segurado, P. & Araújo, M.B. (2004) Evaluation of methods for modelling species probabilities of occurrence. *Journal of Biogeography*, **31**, 1555–1568.

Stockwell, D.R.B. & Peters, D. (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, **13**, 143–158.

Sutherst, R.W. (2001) The vulnerability of animal and human health to parasites under global change. *International Journal of Parasitology*, **31**, 933–948.

Swets, K.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Thuiller, W., Broennimann, O., Hughes, G.O., Alkemade, J.R.M., Midgley, G.F. & Corsi, F. (2006) Vulnerability of African mammals to anthropogenic climate change under conservative land transformation assumptions. *Global Change Biology*, **12**, 424–440.

Turner, M.G. (1990) Spatial and temporal analysis of landscape patterns. *Landscape Ecology*, **4**, 21–30.

Wintle, B.A., Elith, J. & Potts, J.M. (2005) Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, **30**, 719–738.

Zaniewski, A.E., Lehman, A. & Overton, J. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

Accepted 20 May 2008

