Ecological Modelling 220 (2009) 3512–3520

Contents lists available at ScienceDirect

Ecological Modelling

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

The performance of state-of-the-art modelling techniques depends on geographical distribution of species

Mathieu Marmion[a,b,*], Miska Luoto[a,b], Risto K. Heikkinen[c], Wilfried Thuiller[d]

[a] Department of Geography, University of Oulu, P.O. Box 3000, 90014 Oulu, Finland
[b] Thule Institute, University of Oulu, P.O. Box 7300, 90014 Oulu, Finland
[c] Finnish Environment Institute, Research Department, P.O. Box 140, Helsinki, Finland
[d] Laboratoire d'Ecologie Alpine, UMR CNRS 5553, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France

ARTICLE INFO

Article history:
Available online 13 December 2008

Keywords:
Bioclimatic modelling
Latitudinal range
Machine learning
Predictive accuracy
Prevalence
Spatial autocorrelation

ABSTRACT

We explored the effects of prevalence, latitudinal range and clumping (spatial autocorrelation) of species distribution patterns on the predictive accuracy of eight state-of-the-art modelling techniques: Generalized Linear Models (GLMs), Generalized Boosting Method (GBM), Generalized Additive Models (GAMs), Classification Tree Analysis (CTA), Artificial Neural Network (ANN), Multivariate Adaptive Regression Splines (MARS), Mixture Discriminant Analysis (MDA) and Random Forest (RF). One hundred species of Lepidoptera, selected from the Distribution Atlas of European Butterflies, and three climate variables were used to determine the bioclimatic envelope for each butterfly species. The data set consisting of 2620 grid squares $30' \times 60'$ in size all over Europe was randomly split into the calibration and the evaluation data sets. The performance of different models was assessed using the area under the curve (AUC) of a receiver operating characteristic (ROC) plot. Observed differences in modelling accuracy among species were then related to the geographical attributes of the species using GAM. The modelling performance was negatively related to the latitudinal range and prevalence, whereas the effect of spatial autocorrelation on prediction accuracy depended on the modelling technique. These three geographical attributes accounted for 19–61% of the variation in the modelling accuracy. Predictive accuracy of GAM, GLM and MDA was highly influenced by the three geographical attributes, whereas RF, ANN and GBM were moderately, and MARS and CTA only slightly affected. The contrasting effects of geographical distribution of species on predictive performance of different modelling techniques represent one source of uncertainty in species spatial distribution models. This should be taken into account in biogeographical modelling studies and assessments of climate change impacts.

1. Introduction

During recent years, a variety of modelling approaches have been developed and used to convert point information of species distribution into predictive maps. One increasingly employed class of models is bioclimatic envelope models, which can be considered as a special case of niche-based models or species distribution models (Guisan and Zimmermann, 2000; Austin, 2002; Guisan and Thuiller, 2005; Heikkinen et al., 2006). Bioclimatic envelope models correlate current species distributions with climate variables, and may then be used to project spatial shifts in species climatic envelopes according to selected climate change scenarios (Bakkenes et al., 2002; Beaumont and Hughes, 2002; Berry et al., 2002; Pearson and Dawson, 2003; Thuiller, 2003; Huntley et al., 2004; Thuiller et al., 2004a,b).

However, developing useful and reliable applications of bioclimatic models requires a considerable amount of knowledge concerning the factors influencing the accuracy of model predictions (Heikkinen et al., 2006). One potential source of uncertainty in models is the fact that the performance of bioclimatic models is affected by geographical attributes of species, e.g. latitudinal range/marginality (Araújo and Williams, 2000; Segurado and Araújo, 2004), prevalence (Manel et al., 2001; Brotons et al., 2004; McPherson et al., 2004), spatial autocorrelation (Boone and Krohn, 1999) and rarity (Karl et al., 2000, 2002). However, to our knowledge the effects of these factors on the performance of different state-of-the-art bioclimatic modelling techniques have not been analyzed systematically. Our understanding of whether some modelling techniques are more sensitive than others to the effects of geographical attributes of species distribution patterns, or whether some of the techniques are more buffered against such effects, is thus rather limited. Improved knowledge of the potential sources of uncertainties stemming from species geographical characteristics is essential for developing better understanding of the performance of bioclimatic models (Heikkinen et al., 2006)

* *Corresponding author at*: Department of Geography, University of Oulu, P.O. Box 3000, 90014 Oulu, Finland. Tel.: +358 8 553 1714; fax: +358 8 553 1693.
E-mail address: mathieu.marmion@oulu.fi (M. Marmion).

0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2008.10.019

and for interpreting the accuracy assessments (Fielding and Bell, 1997).

In order to produce reliable estimates for species distributions, it is important to know how different modelling techniques behave, particularly when modelling species with different ecological and geographical characteristics. A number of studies (Kadmon et al., 2003; Brotons et al., 2004; McPherson et al., 2004; Segurado and Araújo, 2004; Luoto et al., 2005) have shown that these factors may affect the modelling accuracy. However, the results of these studies have been contradictory. For example, Luoto et al. (2005) showed that the prevalence and the latitudinal range of species were negatively and the spatial autocorrelation was positively related to the modelling accuracy. By contrast, Manel et al. (2001) reported that model accuracy was independent of species prevalence. One possible reason for these contrasting results may be the fact that the two studies employed different modelling techniques (Generalized Additive Model (GAM) in Luoto et al. (2005) and logistic regression in Manel et al. (2001)), which may lead to divergent interpretations. Furthermore, as highlighted by Austin (2007), even models which belong to the same class (e.g. GAM) but employ different settings (e.g. degree of freedom of the smoothers) may have different behaviours, indicating that results from different studies should be compared carefully. Nevertheless, the overall message emerging from these studies, as well as from other complementary studies (e.g. Kadmon et al., 2003; Brotons et al., 2004; McPherson et al., 2004; Segurado and Araújo, 2004), is that species geographical attributes can significantly influence the behaviour and uncertainty of species climate modelling techniques. This should be taken into account in applications such as assessment of climate change impacts.

In this study we provide a relatively comprehensive evaluation of the effects of species geographical attributes on modelling performance using atlas data on butterfly distribution for the whole of Europe (Kudrna, 2002). We explore simultaneously the effects of three geographical attributes on the accuracy of 100 climate–butterfly models using eight state-of-the-art modelling techniques that are implemented in the BIOMOD modelling framework (see Thuiller, 2003). BIOMOD contains conventional and new modelling methods: Generalized Linear Models (GLMs), Generalized Boosting Method (GBM), Generalized Additive Models, Classification Tree Analysis (CTA), Artificial Neural Network (ANN), Multivariate Adaptive Regression Splines (MARS), Mixture Discriminant Analysis (MDA) and Random Forest (RF). The predictive accuracy of the models was studied with a particular focus on two questions: (i) How are the different modelling techniques influenced by the prevalence, spatial autocovariate and the latitudinal range of the species? and (ii) What are the relative roles of different geographical attributes in the uncertainty of different modelling techniques?

2. Material and methods

2.1. Butterfly data

A random selection of butterfly species ($n = 100$, 22%) was extracted from the 451 Lepidoptera species included in the Distribution Atlas of European Butterflies (Kudrna, 2002). In order to reduce the error associated with biased samples or small sample size (Barry and Elith, 2006), species with less than 10 records and species for which distribution appeared to be insufficiently known were excluded from the analysis. The remaining 332 species were assigned to six broad categories according to their biogeographical distribution, based on information derived from Kudrna (2002) and Tshikolovets (2003). The six biogeographical categories of species distribution were (1) bimodal/sporadic, (2) Southern Europe, (3) Mountains of Middle and Southern Europe, (4) Central Europe (including species ranging from Central to Southern Europe), (5) Northern Europe, and (6) Whole Europe (Luoto and Heikkinen, 2008). A set of 100 species was selected, including species from each of these categories, and thus a representative sample of the European butterflies from different environments was obtained (Appendix 1). Species distribution data in Kudrna (2002) is given using 2620 grid squares of $30' \times 60'$ in size. However, only 1608 grid squares were included in the analysis. Most of the eastern European countries were excluded because of the obvious undersampling in these areas. In total, 26,615 presences among the 100 species were recorded over the 1608 grid squares.

2.2. Climate data

Climate data were obtained from the Climatic Research Unit (CRU) climatological database (New et al., 2002; Mitchell et al., 2004). In order to extend the spatial resolution from $0.5° \times 0.5°$, the averages for the time period 1961–1995 were interpolated from the original $30' \times 60'$ grid to match the species data. Following Hill et al. (2003), we used three climate variables that provide essential information about factors limiting butterfly growth and survival: (i) annual temperature sum above 5 °C, (ii) mean temperature of the coldest month, and (iii) the water balance index. Water balance was calculated as the monthly difference between precipitation and potential evapotranspiration and by summing the separate differences, as presented by Skov and Svenning (2004).

2.3. Model calibration and evaluation

All the different models were calibrated using the R environment software (R Development Core Team, 2004) and the BIOMOD user interface (Thuiller, 2003). From the original set of data containing 1608 grid squares, 70% (1125 grid squares) were randomly selected to the model calibration data set, and the remaining 30% (483 grid squares) were assigned into the model evaluation set used in assessing the predictive accuracy of each model.

2.3.1. Models

2.3.1.1. Generalized Linear Models. GLMs are mathematical extensions of linear models (McCullagh and Nelder, 1989). GLMs can handle nonlinear relationships and different types of statistical distributions characterizing spatial data, and are technically closely related to traditional practices used in linear modelling and analysis of variance (ANOVA). For each of the 100 butterfly species, linear, 2nd and 3rd order polynomial terms were computed to provide the probability of occurrence in each grid square, as a response to the three climatic variables. An automatic stepwise procedure is used by BIOMOD to compute the best model by minimizing the Akaike information criterion (AIC) value (Thuiller, 2003).

2.3.1.2. Generalized Additive Models. GAMs are non-parametric extensions of GLMs (Hastie and Tibshirani, 1990), and they are often used in biogeographical studies (Guisan et al., 2002; Araújo et al., 2004; Thuiller et al., 2006). They provide a flexible data-driven class of models based on a cubic-spline smoother with four degrees of freedom that permit both linear and complex additive response shapes, as well as combination of the two within the same model. The smooth functions are computed independently for each explanatory variable and added to build the final model. The model selection of GAM in BIOMOD is based on AIC (Thuiller, 2003).

2.3.1.3. Classification Tree Analysis. CTA is an alternative to regression techniques and has been used rather often in biogeographical and environmental studies (Franklin, 2002). CTA uses recursive partitioning to split the data into increasingly smaller, homogenous, subsets until a termination is reached (Venables and Ripley, 2002). The optimal length of the tree is selected by a 50-fold cross-validation. The advantage of CTA is that it allows capturing of non-additive behaviour and complex interactions. However, CTA has a tendency to produce overly complex models that lead to spurious interpretations (Breiman et al., 1984). CTA is used frequently for biogeographical and environmental studies (De'Ath and Fabricius, 2000; Vayssière et al., 2000; Franklin, 2002; Thuiller et al., 2004a,b).

2.3.1.4. General Boosting Method. GBMs were recently introduced in ecology. They are highly efficient in fitting the data, are non-parametric and combine the strengths of different modern statistical techniques (Ridgeway, 1999). Here, GBM was implemented into R (R Development Core Team, 2004) using the library GBM (Generalized Boosted Regression Modelling). GBM is based on the Gradient Boosting Machine developed by Friedman (2001). GBM proceeds via sequential improvements. Boosting is a numerical optimization technique for minimizing a loss function (such as deviance) by adding at each step a new tree that best reduces the loss function (Ridgeway, 1999; Elith et al., 2008). Environmental variables Π are input into a first regression tree, which maximally reduces the loss function. For each following step, the focus is on the residuals. For example, at the second step a tree is fitted to the residuals of the first tree. The model is then updated to contain two trees, and the residuals from these two trees are calculated. The sequence is repeated as long as necessary (Elith et al., 2008). The maximum number of trees was set to 3000, and ten-fold cross-validations were performed. GBM belongs to the class of learning methods.

2.3.1.5. Mixture Discriminant Analysis. MDA is an extension of linear discriminant analysis (LDA) (Venables and Ripley, 2002). MDA assumes that the distribution of the class of each environmental variable follows a Gaussian distribution. MDA enhances the LDA, allowing the classifier to handle different prototype classes such as a mixture of Gaussians. The environmental parameters form primal classes, which are divided into sub-classes. The classification results from these sub-classes, a mixture density, describe the distribution density of the primal classes of environmental variables. The number of sub-classes was deduced from the variation of the calibration (training) data. The characteristics of the used Gaussian density curves were deduced from the 1125 grid squares included in the training data. An independent observation was then classified into the class, maximizing its probability to belong to this particular class among the other ones (Ju et al., 2003; Bashir and Carter, 2005). It should be noted that different regression methods can be used in the optimal scaling process. BIOMOD uses MARS (see below) to increase the predictive power of the model.

2.3.1.6. Random Forest. RF belongs to the machine learning methods (Breiman, 2001). Random Forest generates hundreds of random trees. A selective algorithm limits the number of implemented parameters in each tree. A training set for each tree is chosen as many times as there are observations, among the whole set of observations. For each node of trees, the decision is taken according to randomly selected environmental parameters. Trees thus constructed are not pruned and are as large as possible. After the trees have been built, data are entered into them and each grid square will be classified by all trees. At the end of the run, the classification given by each tree is considered as a "vote", and the classification of a grid square corresponds to the majority vote among all trees (Breiman, 2001). RF was used by Prasad et al. (2006) for vegetation mapping under current and future climate scenarios.

2.3.1.7. Multivariate Adaptive Regression Splines. MARS combines classical linear regression, mathematical construction of splines and binary recursive partitioning to produce a local model in which relationships between response and predictors are either linear or nonlinear (Friedman, 1991). A pre-processing algorithm of the explanatory variables uses the basic functions (BF) max $(0,X - c)$ and max $(0,c - X)$ to transform the environmental variables into a new set of variables. The main difficulty is to find appropriate "c" values, but a suitable choice makes it possible to approximate any functional shape (Briand et al., 2004). Then, MARS performs successive approximation of the system using different intervals of the transformed variables ranges, by a series of linear regressions. Examples of the use of MARS in biogeographical studies can be found in Muñoz and Felicísimo (2004), in climatology in Corte-Real et al. (1995) and in landscape ecology in Heikkinen et al. (2007).

2.3.1.8. Artificial Neural Networks. ANN is a powerful rule-based modelling technique (Lek and Guegan, 1999), which is increasingly used in bioclimatic envelope modelling (Thuiller, 2003; Heikkinen et al., 2006). We used feed forward neural networks, which belong to the machine learning methods and provide an alternative way to achieve generalized linear regression functions (Venables and Ripley, 2002). A network contains three different types of layers: the input layer (in which the environmental variables are input), the hidden (intermediate) layers and the output layer. Each layer is composed of independent neurons; each of them treats separately the outputs of all neurons from the previous layer as inputs of multivariate linear functions. The process is continued until processing of the output layer. To avoid overfitting in neural networks, a four fold cross-validation method was implemented to stop training of networks. Once the complete network is built, the different weighting factors of the multivariate linear functions are chosen by minimizing the quadratic error of estimate.

2.3.2. Estimation of the model performance

After the models were calibrated, they were transferred to the evaluation data set. In this process, climatic variables were used as input in the models, and the outputs of the models were then compared with the species binary presence/absence information from the evaluation data set. In the evaluation, the area under the curve (AUC) of a receiver operating characteristic (ROC) plot of each model was calculated. AUC is a graphical method assessing the ability of a model to predict the absence or presence of species on the basis of given criteria (e.g. climate variables), by representing the relationship between the false positive fraction and the true positive fraction of the related confusion matrix of the evaluated model (Fielding and Bell, 1997). The range of AUC is from 0 to 1. A model providing excellent prediction has an AUC higher than 0.9, a fair model has an AUC between 0.7 and 0.9, and a model is considered as poor if its AUC is below 0.7 (Swets, 1988).

2.3.3. Species distribution

The geographical patterns of the modelled species were measured by three variables: latitudinal range, spatial autocovariate (clumping of occurrences) and prevalence. The latitudinal range of butterflies was measured as the distance between the northernmost and southernmost distribution record in Europe. Our latitudinal range variable measures geographical distance to the range boundary. For example, a small distance from the northernmost distribution record to the southernmost point in Europe indicates that the species is close to the northern edge of its geographical distribution range (Thuiller et al., 2003). To measure the degree of clumping of occurrences, a spatial autocorrelation variable (i.e. autocovariate) for each individual species was calculated using presence–absence information of the species in order to reveal a patch-like autocorrelation structure in the butterfly data

Table 1
Pearson correlation coefficient between the AUC values of the modelling techniques based on the evaluation data set, and the geographical attributes of the butterfly species. The symbols* and ** indicate that the correlation is significant at the 0.05 and 0.01 levels, respectively.

	ANN	CTA	GAM	GBM	GLM	MARS	MDA	RF
All species								
Prevalence	−0.241*	0.193	−0.651**	−0.384**	−0.651**	−0.206**	−0.469**	−0.188
Spatial autoc.	0.041	0.336**	−0.051	0.142	−0.159	0.183	0.136	0.283**
Latitudinal range	−0.371**	0.205*	−0.696**	−0.387**	−0.680**	−0.164	−0.461**	−0.178

(Augustin et al., 1996). The autocovariate was based on Moran's index, following the method used by Luoto et al. (2005). Moran's index was calculated using the program Rookcase for irregular lattice data using a lag of 75 km (eight possible nearest-neighbour grid squares included) (Sawada, 1999). Prevalence, i.e. the ratio of presence squares to the total sample, was calculated for all the butterfly species studied (Manel et al., 2001). The performance of different modelling techniques for each species was related to the three geographical attributes of the species using multiple GAM. We acknowledge here that both prevalence and latitudinal range are not only functions of the species, but they also depend on how the study area is delimited and how the sampling has been performed (Albert and Thuiller, 2008). For example, some of the species studied here may have different prevalences on the regional scale than on the continental scale. However, because our study area covers a representative part of the whole of Europe, we consider that our data provide a good approximation of the study species prevalence on the continental scale.

3. Results

3.1. Effects of the geographical attributes

The species prevalence varied from 0.01 to 0.62, with a mean of 0.16. This variation in species prevalence values had different impacts on the performance of different modelling techniques. For all methods except CTA, a significant decrease in accuracy in response to increasing prevalence was revealed (Table 1). As examples, Figs. 1A and B illustrate the variation in model accuracy based on GAM and RF in relation to species prevalence. The predictive performance of both models is better for low prevalence species. Fig. 2 shows projections of the distribution of two butterfly species: *Aricias nicias* and *Apatum iris*.

The spatial autocovariate varied from 0.00 to 0.89, with a mean value of 0.42. The negative correlation between the AUC values of GLM and GAM, and the Moran index of the species indicate that both methods are more accurate when the spatial clumping is low (Table 1). For all other models, the correlation was positive, and significant at the 0.01 level only for RF and CTA. Figs. 1C and D show that AUC values based on RF increase with clumping of the modelled species, whereas the AUC values based on GAM hardly follow any trend.

The latitudinal range of species varied between 167 km and 3840 km, with a mean of 1787 km. Table 1 and Figs. 1E and F show a clear negative correlation between the latitudinal range and model performance, except for CTA. The accuracy of CTA models increased with increasing latitudinal range of the species.

3.2. Accuracy of the models in relation to the geographical attributes

The alone contribution (variable on its own) and the drop contribution (when the variable was dropped from the saturated model) of the three geographical attributes derived from GAM analysis with modelling accuracy as response variable are presented in Table 2. The explained deviance illustrates the degree to which the variance of the modelling techniques is influenced by the three geographical attributes in a multivariate setting based on GAM. The accuracies of GAM, GLM and MDA were highly influenced by the three geographical attributes of the species. The explained deviances varied from 43.8% to 61.5%. The machine learning methods RF, GBM and ANN were moderately influenced by the geographical attributes, whereas MARS and CTA were the least influenced techniques (18.5% and 25.1%, respectively; see Fig. 3).

4. Discussion

Recently, several novel modelling methods have been utilised in bioclimatic studies that have foundations in ecological, biogeographical and statistical research (Elith et al., 2006). Along with well-established modelling methods such as Generalized Additive Models and Artificial Neural Networks, we explored methods that have been developed more recently, e.g. the Random Forest and General Boosting Methods, or have rarely been applied to modelling species distributions, e.g. MARS and MDA. In addition to the inherent differences in the predictive capabilities of different techniques (Thuiller, 2003; Segurado and Araújo, 2004; Pearson et al., 2006; Heikkinen et al., 2007), a major problem in predictive modelling studies is to understand what attributes of species might affect model performance (McPherson et al., 2004; Luoto et al., 2005; Pöyry et al., 2008).

In general, research on the effects of the geographical distribution of species on the accuracy of models has focused on univariate analysis, e.g. the impact of prevalence on model performance (Fielding and Bell, 1997; Manel et al., 2001; McPherson et al., 2004). Studies on species distribution modelling have yielded contrasting inferences about the importance of various geographical distribution factors for the performance of distribution models. Statistical artefacts can confound results in comparative studies investigating the role of species geographical attributes in modelling performance (McPherson et al., 2004). In order to mitigate

Table 2
The effects of the geographical attributes on the performance of the eight modelling techniques based on GAM. "NS" attests not selected into the GAMs. "al" is the abbreviation for alone contribution and "drop" stands for drop contribution. The underlined value for a single row of the table emphasizes which modelling technique is most influenced by the attribute. The bold value underscores the model for which its variance is most independent of the attributes.

	ANN al–drop	CTA al–drop	GAM al–drop	GBM al–drop	GLM al–drop	MARS al–drop	MDA al–drop	RF al–drop
Prevalence	8.2–3.0	6.6–6.1	25.7–1.7	8.4–**1.4**	30.9–2.1	6.0–10.8	15.8–10.8	**3.1**–3.7
Spatial autoc.	17.3–21.4	14.6–14.0	7.2–6.7	8.9–11.0	18.5–**7.1**	7.9–7.2	9.4–7.2	**6.9**–11.1
Latitudinal range	23.0–18.6	5.4–**NS**	28.8–4.8	10.2–0.8	32.2–2.1	6.2–**NS**	17.7–**NS**	**4.0**–0.8
Expl. deviance	32.2	25.1	61.4	37.1	58.7	18.5	43.8	34.2

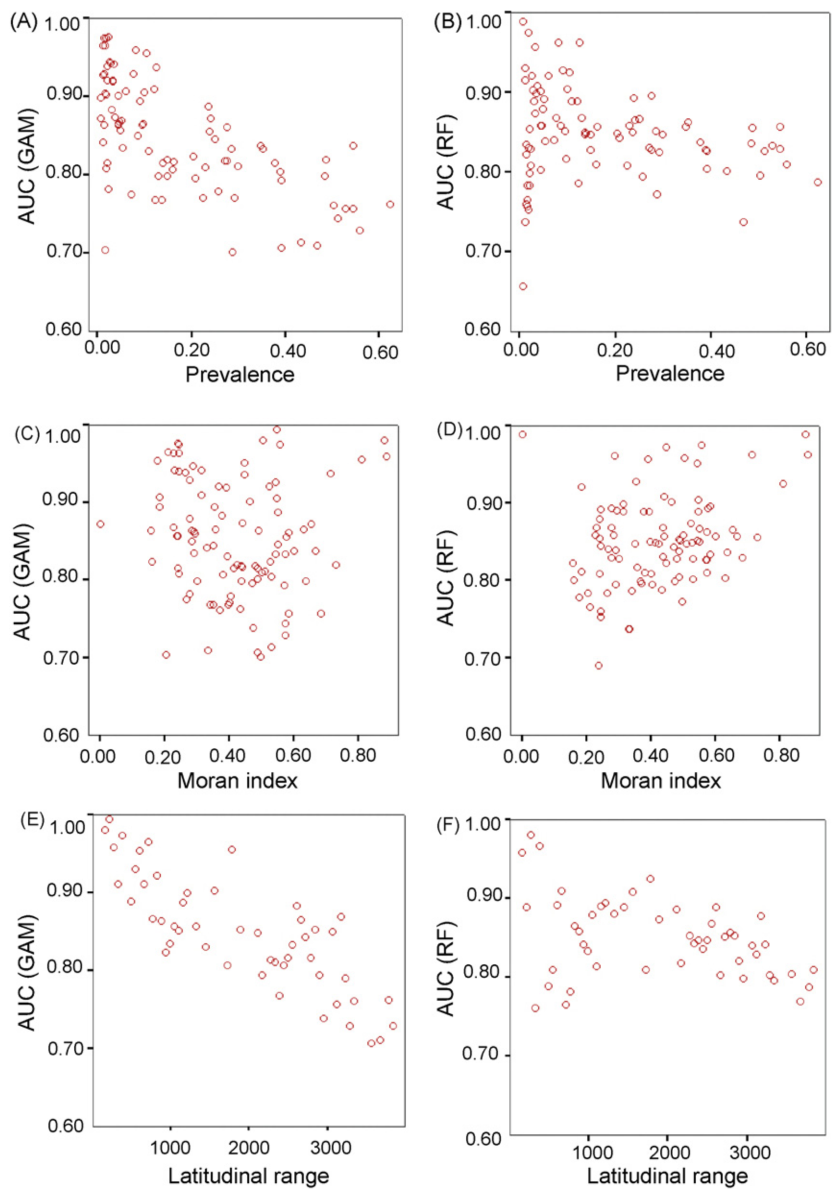

Fig. 1. The three geographical attributes of the species, prevalence (A and B), spatial autocovariate (C and D) and latitudinal range (E and F) showing relationships with the accuracy of climate-butterfly models based on GAM and RF.

these artificial effects, we based this study on the AUC derived from the receiver operating characteristic plots, which are practically immune to prevalence and errors related to sample size (Manel et al., 2001; McPherson et al., 2004). Most importantly, we estimated the relative importance of different geographical attributes of the species on different modelling techniques in order to deepen our understanding of the performance of the techniques with different species distribution patterns.

In general, our modelling showed a relatively close fit between the three climate variables and the distributions of the studied butterfly species in Europe, although the butterfly data were only binary (present/absent) and coarse-grained ($30' \times 60'$). The average level of discrimination in the models was 0.82, and varied between 0.75 (CTA) and 0.85 (GAM and RF). The rather high discrimination ability and low proportion of poor models suggest that butterfly distribution in Europe is clearly correlated with climate (see Luoto and Heikkinen, 2008), and that bioclimate envelope models can provide useful tools to identify the broad-scale relationships between these species and the environment (Pearson and Dawson, 2003). However, comparisons of the performance of the eight modelling techniques indicated certain clear and important differences between the techniques in relation to the three geographical attributes of the species. Thus geographical attributes, such as prevalence, latitudinal range and spatial autocorrelation, may have a notable influence on the accuracy of the models.

4.1. Species geography

Numerous studies have recently demonstrated that the performance of the bioclimatic models may depend on the characteristics of the species (e.g. Venier et al., 1999; Karl et al., 2002; Thuiller et al., 2003; Segurado and Araújo, 2004; Luoto et al., 2006). These studies have indicated that species with limited geographic ranges and specialist species with strict ecological requirements are generally modelled more accurately than species with wide geographic ranges and generalist species with wide ecological tolerance (Heikkinen et al., 2006). However, systematic com-

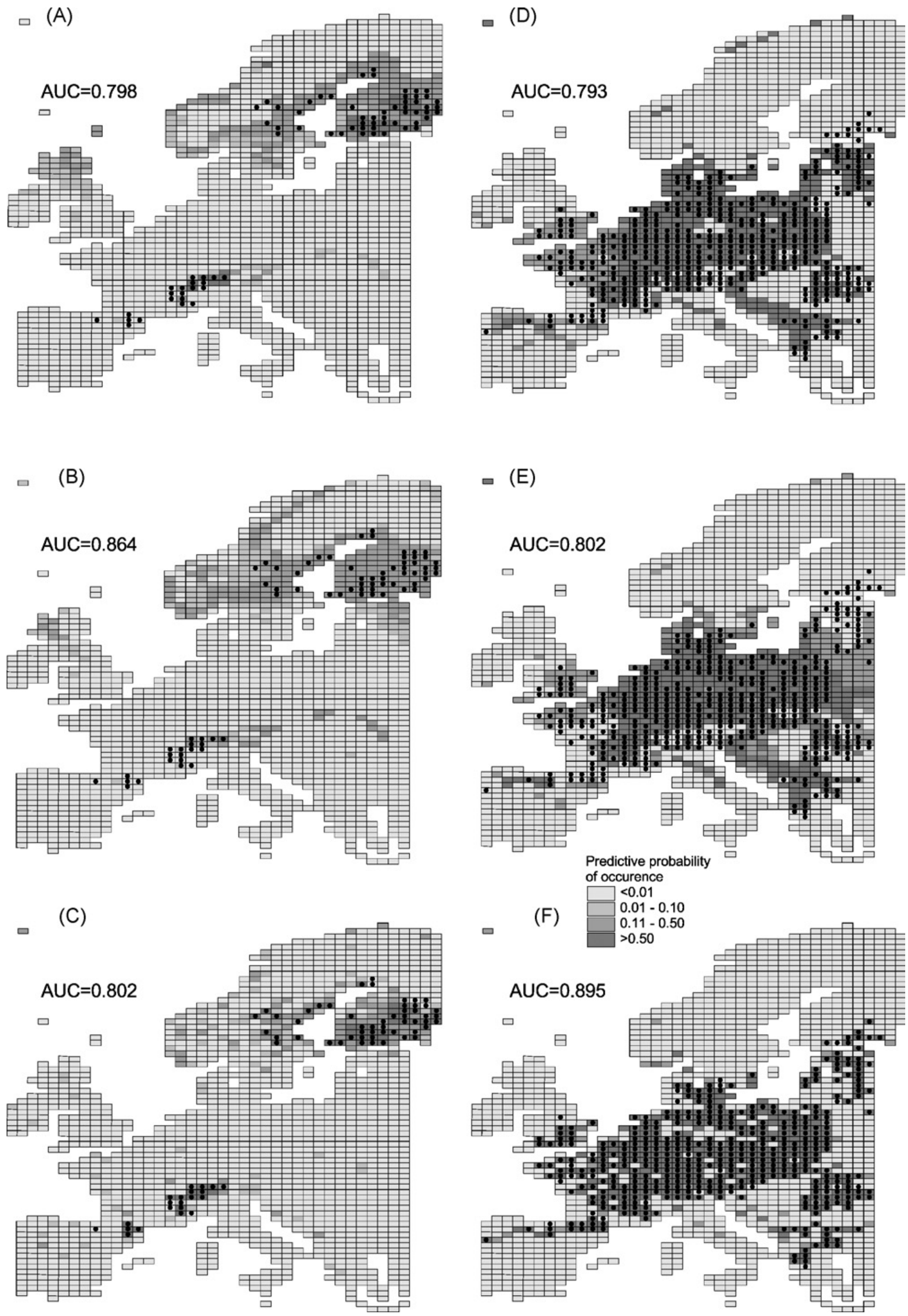

Fig. 2. Projected probability of occurrence of *Aricia nicias* (A, B and C) and *Apatum iris* (D, E, and F) provided by CTA (top row), GAM (middle row) and RF (bottom row). The grey scale represents the four different probability classes and the black dots are the observed occurrences.

parisons with large samples of species and several statistical techniques potentially contributing to model uncertainty are lacking. In order to take full advantage of the species–climate models and to identify critical sources of uncertainties in the models, we need to understand whether the variation in model performance reveals inherent biogeographical or ecological differences in the predictability of different species or whether it reflects statistical or spatial artifacts (Legendre et al., 2002; McPherson et al., 2004).

In our study, prevalence strongly influenced the accuracies of the modelling techniques. This corresponds with observations made by Segurado and Araújo (2004) and Luoto et al. (2005). Their results indicate a trend towards increasing model performance for restricted-range species and decreasing performance for widespread species. One of the main arguments explaining the negative correlation between modelling accuracy and prevalence is the biological niche complexity. A low prevalence indicates a narrow biological niche of the species, which is often rather straightforward to define in a multivariate setting. Segurado and Araújo (2004) noted that model performance is higher for species with high environmental marginality and low niche breadth than for generalist species. By contrast, a species with high prevalence can adapt over

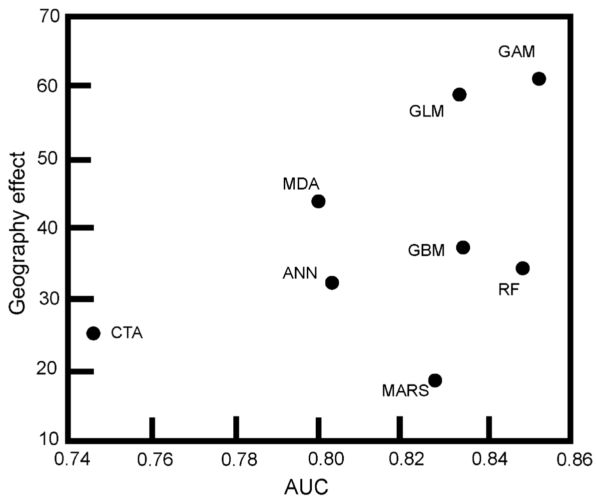

Fig. 3. The performance of the modelling techniques with respect to their predictive accuracy (AUC) and sensitivity to the geographical attributes (geography effects).

a wide range of different climatic environments and its distribution is more difficult to model.

However, we note here that a number of previous studies contradict these arguments. For example, Seoane et al. (2005) showed that the predictive power of regression trees (RT) was highest when modelling species with high prevalence, and Dormann (2007) and Reineking and Schröder (2006) showed that different autologistic regression methods were severely impacted by the prevalence of species, such models only being accurate for species with high prevalence. In the study by Meynard and Quinn (2007), Genetic Algorithm for Rule-Set Prediction (GARP) also provided most accurate models for species with high prevalence. However, models based on GARP use presence only data, which may in part explain this behaviour. McPherson et al. (2004) analyzed the accuracy of bird distribution and concluded that the models were more accurate for intermediate prevalence. By contrast to these studies, Pöyry et al. (2008) were not able to detect any effect of the niche width on the performance of climate-butterfly models in Finland. However, they noticed that the accuracy of climate-butterfly models decreases with increasing mobility and the length of the flight period. The mobility index was significantly positively correlated with prevalence, which is in agreement with our results. Finally, Manel et al. (2001) reported that AUC measures based on large invertebrate data from Himalayan streams were independent of prevalence. Potential reasons for these opposing outcomes may arise from the fact that in the current study a wider range of modelling techniques and their sensitivities to species geographical attributes were examined, which might lead us to results which could be applied more generally than previous studies.

In this study, latitudinal range also affected the model performance: the accuracy of the models decreased with increasing latitudinal range. The climatic environment varies considerably with the latitude. Thus, a species that occurs over a wide latitudinal range is obviously adapted to various types of climates. By contrast, when the latitudinal range of a given species is low, the climatic and environmental space of the species is restricted. Concerning the spatial autocorrelation, our results do not show any general trend which can be clearly linked with the different methods. However, spatial autocovariate was statistically the most significant factor of the three geographical attributes in explaining the variation in the performance of ANN, CTA and GLM. Our results correspond with those of de Frutos et al. (2007) and Dormann (2007), which highlighted the high sensitivity of logistic regression and other GLM methods to spatial autocorrelation. In comparison with spatial autocovariate, latitudinal range strongly influences ANN, GAM, GLM and MDA, whereas prevalence influences almost at the same level in all models.

The eight modelling techniques can be assigned to three categories on the basis of their sensitivities to the geographical attributes of the species. GAM, GLM and MDA are most severely influenced by the geographical attributes (explained deviance higher than 40%). GAM and GLM are rather similar techniques and computationally relatively simple. Machine learning methods (ANN, GBM, RF) are characterized by moderate effect of the geographical attributes on the modelling accuracy (explained deviance between 30 and 40%). By contrast, it appears that CTA and MARS are less controlled by the geographical attributes of the species than the other methods. This can be partly explained by the fact that there are probably other major sources of uncertainty that affect the predictive accuracy of these methods. However, we acknowledge that the outcomes of comparative analysis of different modelling techniques, including the present one, may also be influenced by the different setting of the algorithms. For example, Leathwick et al. (2006) compared GAM models with four different types of MARS models, including multiresponse and interaction settings. Their results suggested that (i) the deviance explained by the models and their predictive accuracy was highly influenced by the chosen setting and (ii) projections based on GAM and MARS had similar accuracy, which partly contradicts our results.

5. Conclusions

The results of this study indicate that novel modelling methods provide various prediction accuracies, which are notably influenced by geographical attributes of species. The modelling performance was related negatively to the latitudinal range and prevalence, whereas the effect of spatial autocorrelation on prediction accuracy depended on the modelling technique. Predictive accuracy of certain modelling techniques, particularly GAM, GLM and MDA, appears to be highly influenced by the three geographical attributes, whereas other techniques are less affected. These results draw attention to the importance of geographical attributes for bioclimatic envelope models, as well as for species spatial distribution models in general. Most importantly, geographical attributes have contrasting effects on the performance of different state-of-the-art modelling techniques. Such uncertainties should be taken into account by down-weighting or excluding species or statistical techniques in studies applying bioclimatic modelling and in assessments of climate change impacts.

Acknowledgments

Different parts of this research were funded by the EC FP6 Integrated Project ALARM (GOCE-CT-2003-506675). WT was partly funded by the EU FP6 MACIS species targeted project (Minimisation of and Adaptation to Climate change: Impacts on biodiversity, contract No.: 044399) and EU FP6 ECOCHANGE integrated project (Challenges in assessing and forecasting biodiversity and ecosystem changes in Europe). MM was funded by the Academy of Finland (project grant 116544). We thank Otakar Kudrna for permission to use the European butterfly distribution data via the ALARM coordinator. M. Bailey helped with correction of the English text. The comments raised by two anonymous referees helped in improving this paper.

Appendix A. Selected 100 butterfly species classified according their biogeographical distribution.

Species	European distribution
Aricia nicias	Bimodal/sporadic
Boloria thore	Bimodal/sporadic
Erebia pandrose	Bimodal/sporadic
Plebejus orbitulus	Bimodal/sporadic
Archon apollinus	Southern European
Aricia anteros	Southern European
Aricia morronensis	Southern European
Carcharodus orientalis	Southern European
Charaxes jasius	Southern European
Coenonympha dorus	Southern European
Colias aurorina	Southern European
Cupido osiris	Southern European
Erebia melas	Southern European
Erebia ottomana	Southern European
Erynnis marloyi	Southern European
Euchloe belemia	Southern European
Glaucopsyche melanops	Southern European
Hipparchia briseis	Southern European
Hipparchia fatua	Southern European
Hipparchia fidia	Southern European
Hipparchia sentles	Southern European
Hipparchia volgensis	Southern European
Leptotes pirithous	Southern European
Lycaena ottomana	Southern European
Maniola bathseba	Southern European
Melanargia arge	Southern European
Melanargia occitanica	Southern European
Melitaea parthenoides	Southern European
Nymphalis egea	Southern European
Papilio alexanor	Southern European
Pararge roxelana	Southern European
Polyommatus albicans	Southern European
Polyommatus dolus	Southern European
Polyommatus escheri	Southern European
Polyommatus nivescens	Southern European
Pyrgus onopordi	Southern European
Scolitantides bavius	Southern European
Zerynthia cerisyi	Southern European
Boloria graeca	Mountains of middle and southern Europe
Boloria pales	Mountains of middle and southern Europe
Colias phicomone	Mountains of middle and southern Europe
Erebia epistygne	Mountains of middle and southern Europe
Erebia eriphyle	Mountains of middle and southern Europe
Erebia pronoe	Mountains of middle and southern Europe
Melitea varia	Mountains of middle and southern Europe
Oeneis glacialis	Mountains of middle and southern Europe
Parnassius phoebus	Mountains of middle and southern Europe
Plebejus glandon	Mountains of middle and southern Europe
Apatura ilia	Central Europe
Apatura iris	Central Europe
Boloria dia	Central Europe
Coenonympha arcania	Central Europe
Colias myrmidone	Central Europe
Cupido argiades	Central Europe
Hipparchia semele	Central Europe
Pararge achine	Central Europe
Satyrium pruni	Central Europe
Satyrium w-album	Central Europe
Carcharodus flocciferus	Central Europe
Favonius quercus	Central Europe
Hamearis lucina	Central Europe
Heteropterus morpheus	Central Europe
Melitaea didyma	Central Europe
Nymphalis polychloros	Central Europe
Pararge megera	Central Europe
Parnassius mnemosyne	Central Europe
Plebejus argyrognomon	Central Europe
Pyrgus armoricanus	Central Europe
Pyrgus carthami	Central Europe
Pyrgus serratulae	Central Europe
Satyrium ilicis	Central Europe
Satyrium spini	Central Europe
Scolitantides baton	Central Europe
Scolitantides vicrama	Central Europe
Thecla betulae	Central Europe
Aporia crataegi	Whole Europe
Argynnis niobe	Whole Europe
Argynnis paphia	Whole Europe
Boloria euphrosyne	Whole Europe
Boloria selene	Whole Europe
Brenthis ino	Whole Europe
Coenonympha glycerion	Whole Europe
Erebia ligea	Whole Europe
Euphydryas aurinia	Whole Europe
Euphydryas maturna	Whole Europe
Glaucopsyche alexis	Whole Europe
Nymphalis c-album	Whole Europe
Nymphalis io	Whole Europe
Nymphalis urticae	Whole Europe
Ochlodes sylvanus	Whole Europe
Papilio machaon	Whole Europe
Pararge aegeria	Whole Europe
Pararge maera	Whole Europe
Boloria aquilonaris	North Europe
Boloria chariclea	North Europe
Carterocephalus silvicolus	North Europe
Erebia embla	North Europe
Erebia polaris	North Europe
Oeneis bore	North Europe
Oeneis jutta	North Europe

References

Albert, C., Thuiller, W., 2008. Favourability functions versus probability of presence: advantages and misuses. Ecography 31, 417–422.

Araújo, M.B., Cabeza, M., Thuiller, W., Hannah, L., Williams, P.H., 2004. Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. Global Change Biology 10, 1618–1626.

Araújo, M.B., Williams, P.H., 2000. Selecting areas for species persistence using occurrence data. Biological Conservation 96, 331–345.

Augustin, N., Mugglestone, M.A., Buckland, S.T., 1996. An autologistic model for spatial distribution of wildlife. Journal of Applied Ecology 33, 339–347.

Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological Modelling 200, 1–19.

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

Bakkenes, M., Alkemade, J., Ihle, F., Leemans, R., Latour, J., 2002. Assessing the effects of forecasted climate change on the diversity and distribution of European higher plants for 2050. Global Change Biology 8, 390–407.

Barry, S., Elith, J., 2006. Error and uncertainty and habitat models. Journal of Applied Ecology 43, 413–423.

Bashir, S., Carter, E.M., 2005. High breakdown mixture discriminant analysis. Journal of Multivariate Analysis 93, 102–111.

Beaumont, L.J., Hughes, L., 2002. Potential changes in the distributions of latitudinally restricted Australian butterfly species in response to climate change. Global Change Biology 8, 954–971.

Berry, P., Dawson, T., Harrison, P., Pearson, R., 2002. Modelling potential impacts of climate change on the bioclimatic envelope of species in Britain and Ireland. Global Ecology and Biogeography 11, 453–462.

Boone, R.B., Krohn, W.B., 1999. Modeling the occurrence of bird species: are the errors predictable? Ecological Applications 9, 835–848.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, F., Olshen, F., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Pacific Grove, USA.

Briand, L.C., Freimut, B., Vollei, F., 2004. Using multiple adaptive regression splines to support decision making in code inspections. The Journal of Systems and Software 73, 205–217.

Brotons, L., Thuiller, W., Araujo, M.B., Hirzel, A.H., 2004. Presence-absence versus presence-only habitat suitability models: the role of species ecology and prevalence. Ecography 27, 165–172.

Corte-Real, J., Zhang, X., Wang, X., 1995. Downscaling GCM information to regional scales: a non-parametric multivariate regression approach. Climate Dynamics 11, 413–424.

De'Ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.

de Frutos, Á., Olea, P.P., Vera, R., 2007. Analyzing and modelling spatial distribution of summering lesser kestrel: the role of spatial autocorrelation. Ecological Modelling 200, 33–44.

Dormann, C.F., 2007. Assessing the validity of autologistic regression. Ecological Modelling 207, 234–242.

Elith, J., Graham, C., Anderson, R., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehman, A., Li, J., Lohman, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. Overton, J., Peterson, T.A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77, 802–813.
Fielding, A., Bell, J., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24, 38–49.
Franklin, J., 2002. Enhancing a regional vegetation map with predictive models of dominant plant species in chaparral. Applied Vegetation Science 5, 135–146.
Friedman, J., 1991. Multivariate adaptive regression splines. Annals of Statistics 19, 1–141.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29, 1189–1232.
Guisan, A., Edwards, T.C.J., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157, 89–100.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8, 993–1009.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.
Heikkinen, R.K., Luoto, M., Araújo, M.B., Virkkala, R., Thuiller, W., Sykes, M.T., 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography 30, 751–777.
Heikkinen, R.K., Luoto, M., Toivonen, T., Kuussaari, M., 2007. Effects of model complexity, spatial resolution and modelling technique on distribution modelling of a threatened butterfly. Landscape and Urban Planning 79, 347–357.
Hill, J.K., Thomas, C.D., Huntley, B., 2003. Modeling present and potential future ranges of European butterflies using climate response surfaces. In: Bogs, C., Watt, W., Ehrlich, P., Chicago (Eds.), Butterflies. Ecology and Evolution Taking Flight. The University of Chicago Press, pp. 149–167.
Huntley, B., Green, R.E., Collingham, Y.C., Hill, J.K., Willis, S.G., Bartlein, P.J., Cramer, W., Hagemeijer, W.J.M., Thomas, C.D., 2004. The performance of models relating species geographical distributions to climate is independent of trophic level. Ecology Letters 7, 417–426.
Ju, J., Kolaczyk, E.D., Gopal, S., 2003. Gaussian mixture discriminant analysis and sub-pixel land cover characterization in remote sensing. Remote Sensing of Environment 84, 550–560.
Kadmon, R., Farber, O., Danin, A., 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications 13, 853–867.
Karl, J., Svancara, L., Heglund, P., Wright, N.M., Scott, J.M., 2002. Species commeness and the accuracy of habitat-relationships models. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B. (Eds.), Predicting Species Occurrences. Issues of Accuracy and Scale. Island Press, Washington, pp. 573–580.
Karl, J.W., Heglund, P.J., Garton, E.O., Scott, J.M., Wright, N.M., Hutto, R.L., 2000. Sensitivity of species habitat-relationship model performance to factors of scale. Ecological Applications 10, 1690–1705.
Kudrna, O., 2002. The distribution atlas of European butterflies. Oedippus 20, 1–342.
Leathwick, J.R., Elith, J., Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199, 188–196.
Legendre, P., Dale, M.R.T., Fortin, M.-J., Gurevitch, J., Hohn, M., Myers, D., 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25, 601–615.
Lek, S., Guegan, J., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling 120, 65–73.
Luoto, M., Heikkinen, R.K., 2008. Disregarding topographical heterogeneity biases species turnover assessments based on bioclimatic models. Global Change Biology 14, 483–494.
Luoto, M., Heikkinen, R.K., Pöyry, J., Saarinen, K., 2006. Determinants of biogeographical distribution of butterflies in boreal regions. Journal of Biogeography 33, 1764–1778.
Luoto, M., Pöyry, J., Heikkinen, R.K., Saarinen, K., 2005. Uncertainty of bioclimate envelope models based on geographical distribution of species. Global Ecology and Biogeography 14, 575–584.
Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921–931.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, New York.
McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? Journal of Applied Ecology 41, 811–823.
Meynard, C.N., Quinn, J.F., 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. Journal of Biogeography 34, 1455–1469.
Mitchell, T.D., Carter, T.R., Jones, P.D., Hulme, M., New, M., 2004. A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901–2000) and 16 scenarios (2001–2100). Tyndall Centre Working Paper 55. pp. 1–30.
Muñoz, J., Felicísimo, A., 2004. Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science 15, 285–292.
New, M., Lister, D., Hulme, M., Makin, I., 2002. A high-resolution data set of surface climate over global land areas. Climate Research 21, 1–25.
Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12, 361–371.
Pearson, R.G., Thuiller, W., Araujo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.P., Lees, D.C., 2006. Model-based uncertainty in species range prediction. Journal of Biogeography 33, 1704–1711.
Pöyry, J., Luoto, M., Heikkinen, R.K., Saarinen, K., 2008. Species traits are associated with the quality of bioclimatic models. Global Ecology and Biogeography 17, 403–414.
Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.
R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Reineking, B., Schröder, B., 2006. Constrain to perform: regularization of habitat models. Ecological Modelling 193, 675–690.
Ridgeway, G., 1999. The state of boosting. Computing Sciences and Statistics 31, 172–181.
Sawada, M., 1999. ROOKCASE: an excel 97/2000 visual basic (VB) add-in for exploring global and local spatial autocorrelation. Bulletin of Ecological Society of America 80, 231–234.
Segurado, P., Araújo, M., 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography 31, 1555–1569.
Seoane, J., Carrascal, L.M., Alonso, C.L., Palomino, D., 2005. Species-specific traits associated to prediction errors in bird habitat suitability modelling. Ecological Modelling 185, 299–308.
Skov, F., Svenning, J.-C., 2004. Potential impact of climatic change on the distribution of forest herbs in Europe. Ecography 27, 366–380.
Swets, K., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.
Thuiller, W., 2003. BIOMOD—optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology 9, 1353–1362.
Thuiller, W., Araújo, M.B., Lavorel, S., 2004a. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. Journal of Vegetation Science 14, 669–680.
Thuiller, W., Araújo, M.B., Pearson, R.G., Whittaker, R.J., Brotons, L., Lavorel, S., 2004b. Uncertainty in predictions of extinction risk. Nature 430, 34.
Thuiller, W., Broennimann, O., Hughes, G.O., Alkemade, J.M.R., Midgley, G.F., Corsi, F., 2006. Vulnerability of African mammals to anthropogenc climate change under conservative land transformation assumptions. Global Change Biology 12, 424–440.
Thuiller, W., Vayreda, J., Pino, J., Sabate, S., Lavorel, S., Gracia, C., 2003. Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). Global Ecology and Biogeography 12, 313–325.
Tshikolovets, V., 2003. Butterflies of Eastern Europe, Urals ans Caucasus. An Illustrated Guide. National Academy of Sciences of Ukraine, National Museum of Natural History, Zoological Museum Kyiv - Brno.
Vayssière, M.P., Plant, R.E., Allen-Diaz, B.H., 2000. Classification trees: an alternative non-parametric approach for predicting species distributions. Journal of Vegetation Science 11, 679–694.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics. S. Springer-Verlag, Berlin.
Venier, L., McKenny, D., Wang, Y., McKee, J., 1999. Models of large-scale breeding-bird distribution as a function of macro-climate in Ontario, Canada. Journal of Biogeography 26, 315–328.