Journal of Applied Ecology 2006
43, 424–432

Modelling ecological niches with support vector machines

JOHN M. DRAKE,* CHRISTOPHE RANDIN† and ANTOINE GUISAN†

**National Center for Ecological Analysis and Synthesis, 735 State St, Suite 300, Santa Barbara, CA 93101, USA; and*
†Department of Ecology and Evolution, University of Lausanne, Biology Building, CH-1015 Lausanne, Switzerland

Summary

1. The ecological niche is a fundamental biological concept. Modelling species' niches is central to numerous ecological applications, including predicting species invasions, identifying reservoirs for disease, nature reserve design and forecasting the effects of anthropogenic and natural climate change on species' ranges.
2. A computational analogue of Hutchinson's ecological niche concept (the multidimensional hyperspace of species' environmental requirements) is the support of the distribution of environments in which the species persist. Recently developed machine-learning algorithms can estimate the support of such high-dimensional distributions. We show how support vector machines can be used to map ecological niches using only observations of species presence to train distribution models for 106 species of woody plants and trees in a montane environment using up to nine environmental covariates.
3. We compared the accuracy of three methods that differ in their approaches to reducing model complexity. We tested models with independent observations of both species presence and species absence. We found that the simplest procedure, which uses all available variables and no pre-processing to reduce correlation, was best overall. Ecological niche models based on support vector machines are theoretically superior to models that rely on simulating pseudo-absence data and are comparable in empirical tests.
4. *Synthesis and applications*. Accurate species distribution models are crucial for effective environmental planning, management and conservation, and for unravelling the role of the environment in human health and welfare. Models based on distribution estimation rather than classification overcome theoretical and practical obstacles that pervade species distribution modelling. In particular, ecological niche models based on machine-learning algorithms for estimating the support of a statistical distribution provide a promising new approach to identifying species' potential distributions and to project changes in these distributions as a result of climate change, land use and landscape alteration.

Key-words: ecological modelling, habitat, presence-only data, species distribution

Journal of Applied Ecology (2006) **43**, 424–432
doi: 10.1111/j.1365-2664.2006.01141.x

Introduction

Methodological advances in species distribution modelling have been rapid (Guisan & Zimmermann 2000; Scott *et al*. 2002). While the practical and intellectual benefits of obtaining well-tested models for species' distributions are numerous, including forecasting species' range shifts from climate change (Thomas *et al*. 2004) and invasion by introduced species (Peterson 2003; Drake & Bossenbroek 2004), testing evolutionary hypotheses (Graham *et al*. 2004), identifying reservoirs for disease (Peterson *et al*. 2002), and planning for conservation in a dynamic landscape (Ferrier 2002), modelling species' niches is complicated by conceptual and technical difficulties and by data limitations (Guisan & Thuiller 2005). Recent advances in machine-learning techniques for statistical pattern recognition might be used to overcome many of these obstacles, which generally result from assumptions about the statistical distribution of data or restrictive parametric modelling paradigms. We studied the accuracy and reliability of ecological niche models built with support vector machines (SVM) for estimating the support of a statistical distribution (Schölkopf *et al*. 2001; Tax 2001; Tax

Correspondence: John M. Drake, National Center for Ecological Analysis and Synthesis, 735 State St, Suite 300, Santa Barbara, CA 93101, USA (e-mail drake@nceas.ucsb.edu).

& Duin 2004). We show that the SVM framework performs comparably or is superior to other methods with only moderate amounts of data while avoiding common problems and limitations.

The most common obstacles to conventional parametric and non-parametric statistical methods for modelling species' distributions are: (i) autocorrelated observations resulting from the inherent spatial distribution of ecological systems, spatial autocorrelation in species' actual distributions, and haphazard rather than designed sampling; and (ii) observations only of species' occurrences without complementary observations of species' absences. Autocorrelated observations result in inflated *P*-values for hypothesis testing when modelling techniques are based on parametric statistics, and have the potential to introduce bias in estimated models. One approach to this problem in a parametric setting is to add to a generalized linear model (GLM; e.g. logistic model) terms to model the spatial correlation (Augustin, Mugglestone & Buckland 1996; He, Zhou & Zhu 2003). Other studies have taken a similar approach with semi-parametric regression techniques, such as generalized additive models (GAM; Leathwick & Austin 2001). However, these methods place further demands on already sparse data and extrapolate poorly.

Strictly speaking, the second obstacle, lack of data confirming species' absences, renders modelling approaches based on classification/discrimination impossible (Robertson, Caithness & Villet 2001; Hirzel *et al*. 2002). Previous studies have sought to overcome this problem by simulating observations of species' absences (sometimes called pseudo-absences) from data domains in which there are no observations of species' occurrences (Engler, Guisan & Rechsteiner 2004). While remarkably robust models have been developed using this approach (Anderson, Lew & Peterson 2003), a method that does not rely on such heuristics would be useful. Further, it is not clear that these procedures can be used in a setting that is not already information rich, where background knowledge of species' ecologies can guide modelling heuristics (Anderson, Lew & Peterson 2003), although these are precisely the cases where species distribution models are most useful, for instance for forecasting species invasions or range shifts from climate change. Finally, classification models fitted to simulated data are generally ecologically uninformative or cumbersome to interpret (Keating & Cherry 2004). The aim of this study was to introduce a technique that overcomes these obstacles.

A promising alternative to conventional classification-based species distribution models is to use methods designed for modelling one type of data only (Robertson, Caithness & Villet 2001; Hirzel *et al*. 2002; Brotons *et al*. 2004; Phillips, Dudík & Schapire 2004). Many such techniques may be found in the literature on statistical pattern recognition, where a frequent goal is to separate statistical outliers from observations drawn from a high-dimensional distribution (Schölkopf *et al*. 2001; Tax 2001; Tax & Duin 2004). Indeed, rather than estimating the full probability distribution, in such situations it may be simpler (and more robust) to model just the support of the distribution, the set of points where the (unknown) probability density is greater than zero (Schölkopf *et al*. 2001). Sometimes support estimation is called one-class classification (Tax 2001). While many different methods for estimating statistical distributions might be optimized for one-class classification (Tax 2001; Tax & Duin 2004), methods based on SVM have been particularly successful in applications where data represent a large set of variables (Tax 2001, table 4·2; Tax & Duin 2004). SVM use a functional relationship known as a kernel to map data onto a new hyperspace in which complicated patterns can be more simply represented (Müller *et al*. 2001). The choice of kernel is typically based on theoretical properties, while any kernel parameters are optimized using computational techniques such as cross-validation. Because SVM are not based on characteristics of statistical distributions there is no theoretical requirement for observed data to be independent, thereby overcoming the problem of autocorrelated observations, although model performance will be affected by how well the observed data represent the range of environmental variables. Further, SVM are more stable, require less model tuning, and have fewer parameters than other computational optimization methods such as neural networks (Lusk, Guthery & DeMaso 2002). Finally, computational complexity is minimal and standard algorithms can be used for optimization. Thus, implementation is straightforward in familiar scientific computing environments such as R (http://www.r-project.org/, accessed 16 February 2006) and MATLAB (Mathworks Inc., Natick, MA). In contrast to genetic algorithms (Stockwell & Peters 1999; Drake & Bossenbroek 2004), the solution is deterministic, resulting in both faster computation and repeatable results. Thus, the potential gains from using support vector machines for ecological niche modelling are great, including reliable and accurate forecasting, feasible computation and a high level of ecological interpretability (Guo, Kelly & Graham 2005).

Methods

STUDY AREA

The study area and data in our analysis were derived from a project to generate forecasts of the effects of climate change on the distribution and diversity of plant species in alpine areas (MODIPLANT project; http://ecospat.unil.ch, accessed 16 February 2006). The study area included all mountains of the pre-Alps of the Canton de Vaud, a Swiss state (6°60′–7°10′E and 46°10′–46°30′N), with a total area of 564 km^2 (Fig. 1). Altitude ranged from 400 m to 3200 m a.s.l. The bedrock in the area is mainly calcareous. The climate is temperate, and generally wet with abundant rain and snowfalls. The sequence of vegetation along the altitudinal gradient is typical of the calcareous Alps, with deciduous forests at the lowest altitudes, mixed forests at middle altitudes, and coniferous

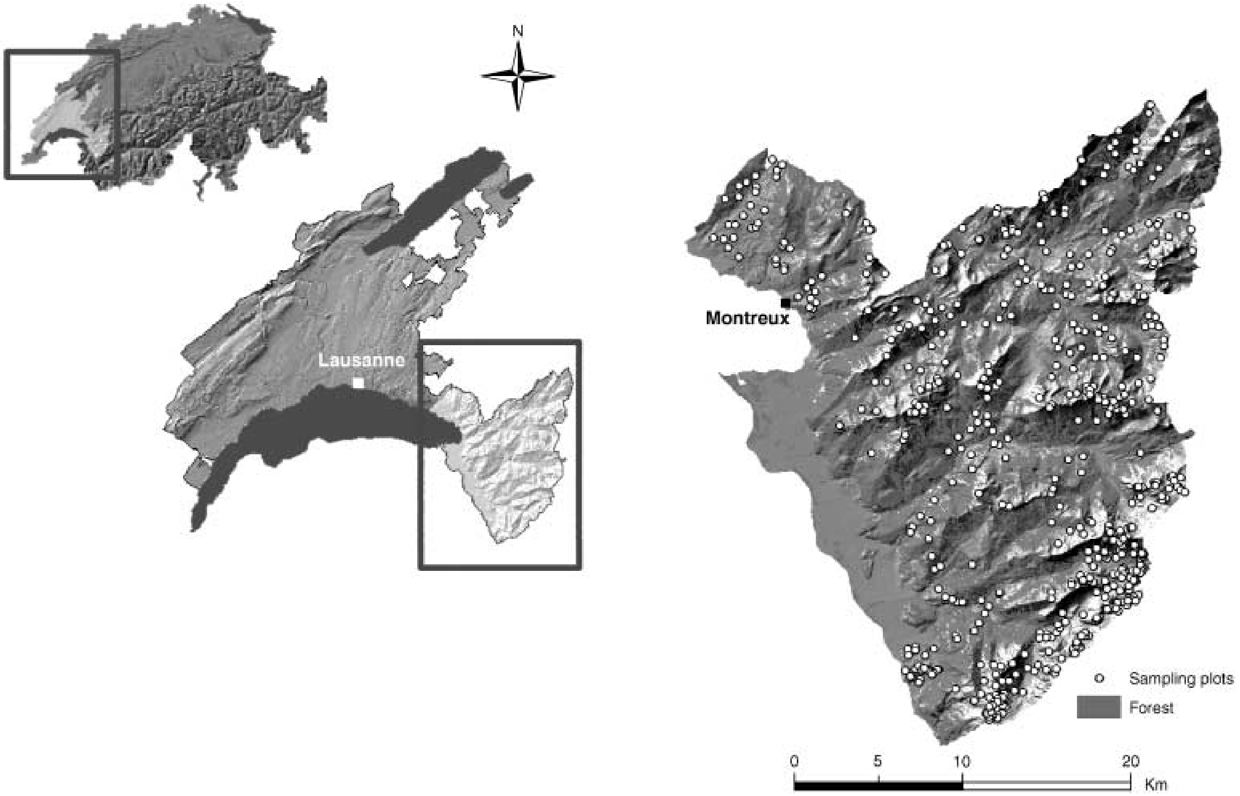

Fig. 1. Location of sampling plots in the pre-Alps of the Canton de Vaud, Switzerland, for 106 species used to compare accuracy and reliability of ecological niche models.

forest at the highest altitudes below the tree line (subalpine belt). Past and current human practices (such as pasturing) have created large gaps in the previously continuous forest cover at all altitudes. Above the tree line, heaths, meadows and grasslands alternate to define the mosaic of alpine vegetation (Randin *et al.* in press).

SPECIES DATA

The data set of species observations comprised 550 vegetation plots of 64 m^2 (8 × 8 m) that were sampled in summer (May through to mid-September; vegetation season) during the period 2002–04. A stratified random design, restricted to open non-woody vegetation (grasslands, rock and scree vegetation) and considering an equal number of replicates per stratum (Hirzel & Guisan 2002), was applied to the entire study area. We selected 106 species for modelling (see Table S1 in the Supplementary material), according to the following criteria: species must (i) have been represented by 10 or more observations, (ii) be taxonomically stable to avoid errors of identification and (iii) be easily identifiable in the field (i.e. have a high detectability) to ensure a low level of omission errors, which would bias estimates of model accuracy.

ENVIRONMENTAL DATA

All environmental variables were measured or computed at the 25-m spatial resolution of the digital elevation model (DEM; MNT25 Level 2, © 2001 Swisstopo (DV83.4), Federal Office of Topography, Wabern, Switzerland) from which they were derived. These were mainly topographic and climate variables. Slope was calculated from the DEM to account for gravitational and disturbance processes acting upon vegetation. A topographic index was calculated using an ArcInfo AML (Arc Macro Language) code (N. E. Zimmermann, http://www.wsl.ch/staff/niklaus.zimmermann, accessed 16 February 2006). This variable reflected the relative position of plants between ridges and valleys, which exposes them to different microclimates (wind, temperature and radiation). Maps of cumulative monthly precipitation and average monthly temperature were calculated by interpolating measurements from the Swiss meteorological network (1961–90) on the DEM (Federal Office of Meteorology and Climatology, Zurich). Lapse rates along the altitudinal gradient were used to normalize the monthly average to 0 m a.s.l., where the interpolations were performed, and then projected back to their actual altitude using the same lapse rates. This method differs slightly from the approach by Zimmermann & Kienast (1999) in using inverse-weighted difference rather than thin-plate splines for interpolation. The resulting climate maps were then used to derive predictors with putative causal relations to individual fitness (see Table S2 in the Supplementary material). Additionally, the spatial hydrological model PREVAH (Gurtz *et al.* 2003) was used to obtain a predictor for snow cover duration at all study sites, based on interpolated daily values of five meteorological variables, precipitation, air temperature, relative sunshine duration, wind speed and water vapour pressure, measured during the period 1979–2000 (Randin *et al*. in press). All environmental variables were expected to have direct physiological effects on the plant species (Pearson *et al*. 2002; Dirnböck *et al*. 2003; Körner 2003).

MODEL TRAINING

Models were fit using only observations of species' presences. Prior to analysis, data were linearly rescaled to the interval between 0 and 1 to avoid potential problems with numerical stability. Most variables were symmetrically distributed and centred within the sample space, providing good coverage of the data domain (see Figure S1 in the Supplementary material). Therefore, no further transformations were performed. Because our goal was to estimate accuracy and reliability of ecological niche models fitted to observations of presence only, records of species' absences were excluded from data used for model training but were retained for model testing. Observations of species' presences were then randomly assigned to one of two subsets for model training (80%) and testing (20%). The final model testing data set was obtained by combining these observations (i.e. 20% of the total) with all records of species absence.

Several environmental variables were highly correlated (see Table S3 in the Supplementary material). Although our procedure did not assume uncorrelated predictors, we considered three different approaches to see how dimensionality and bias might be reduced by eliminating correlated variables. Our baseline, method 1, was to use all nine environmental predictors with no additional pre-processing.

Statistical learning exploits the variability in a set of observations and is often compromised by poorly scaled data. The initial rescaling we performed was expected to resolve problems concerning the stability of the numerical algorithms, but did not address the possibility that the data were clumped, reducing variation for the algorithm to exploit. Motivated by this problem, Tax & Juszczak (2003) suggest reducing the dimensionality of a data set through '*k*-whitening', by mapping the observed data onto the principal components of the training data. Accordingly, a computational analogue of principal components analysis (Kernel-PCA; Schölkopf & Smola 2001) was estimated and the principal components accounting for a specified fraction of the total variance were retained, rescaling the data to a feature space centred on zero and unit covariance. *k*-whitening may be thought of as a mapping that relates a full data set to a new, lower-dimensional space, which is retained and used in subsequent analyses. Thus, method 2 comprised constructing a new data set by *k*-whitening the full data set used in method 1, and then training the support vector machine on the new data set. Accounting for 99% of the total variance, the *k*-whitened data set resulted in reducing the number of dimensions used in model training from nine to five.

Finally, we considered whether or not we could reduce dimensionality by simply eliminating variables from the original data set altogether. As all of the variables related to solar radiation were highly correlated, we chose to eliminate all except SRADYY, which was relatively symmetrically distributed over the sample space. We further dropped MIND07 because it was highly correlated with PRECYY. Thus, method 3 consisted of using only the original (non-*k*-whitened) variables SLOPE, TOPO, SRADYY and PRECYY.

All analyses were conducted in MATLAB 7·0·1 (MathWorks Inc., Natick, MA). For model kernel we used the Gaussian radial basis function (RBF), which is a standard kernel for classification tasks and relies on tuning only one parameter. We optimized the models subject to the target false negative rate of 0·1 using the consistency criterion of Tax & Müller (2004), which minimizes overfitting by increasing the complexity of the model until the cross-validation error on the training data is greater than expected based on random sampling. Specifically, a model was determined to be inconsistent when the estimated false negative rate exceeded the target false negative rate by two standard deviations from the binomial distribution. Analyses were performed using a MATLAB toolbox for statistical pattern recognition (Duin *et al*. 2004) and the MATLAB Data Description Toolbox (Tax 2005).

MODEL TESTING

Models were tested using only data that were not used for model training. We were not able to obtain a consistent model for every species for each protocol. For species for which a consistent model could not be found we used as the best model the support vector machine defined by the Gaussian RBF with a tuning parameter equal to the maximum Euclidean distance between observations in the (rescaled) data set. Otherwise the model optimized by the consistency criterion of Tax & Müller (2004) was retained as the best model. This convention for 'best' model where consistency could not be obtained did not greatly affect our analysis as consistent models were almost always obtained for species represented by greater than 40 observations.

Having trained a best model for each species according to each method, we applied it to the data in the model-testing data set to estimate the following performance criteria: the false-positive rate, the false-negative rate, precision (the ratio of the correct positive predictions to the total number of positive predictions in the testing data set, also called positive predictive power), and recall (the ratio of the number of correct predictions to the total number of observations in the testing data set). We also calculated a composite measure of accuracy that accounted for both precision and recall (Tax 2005):

$$f_1 = \frac{2(\text{precision})(\text{recall})}{\text{precision} + \text{recall}}$$

Finally, in analyses such as this where maximizing the rate of true-positive classifications comes at the cost of false-positive classifications, it is customary to compute the receiver-operator curve (ROC), which represents the true-positive rate as a function of the false-positive rate (Fielding & Bell 1997). The ROC summarizes this fundamental trade-off and can be used to compare different modelling methods or data sets. However, as ROC for different models can intersect, the area under the curve (AUC) is commonly computed (Bradley 1997). This quantity permits comparison of different studies using a single number. AUC ranges from 0 (incorrect classification of all examples) to 1 (correct classification of all examples).

Results

Overall, we found that methods 1 and 2 performed similarly over the different measures and were superior to method 3. However, consistent models were more often obtained with methods 1 and 3 than with method 2, so that method 1 most reliably provided the best results. We obtained consistent models with method 2 for 58 out of 106 species (54·7%). In contrast, consistent models were obtained with method 1 for 87 (82·0%) species and with method 3 for 80 (75·5%) species. Logistic regression showed that, for all three methods, the likelihood of obtaining a consistent model for any given species increased significantly with the number of observations in the data set (method 1, $P < 0{\cdot}0001$; method 2, $P < 0{\cdot}0001$; method 3, $P = 0{\cdot}0001$; see Figure S2 in the Supplementary material).

Error rates and summary performance criteria were also affected by the number of observations with which we trained the model (Fig. 2; see Figures S3–S8 in the Supplementary material). We computed Spearman rank-order correlation between each measure of accuracy except f_1 (which, in one case, was undefined) and sample size, by pre-processing method, using only species for which we obtained consistent models. We consistently found highly significant relationships (see Table S4 in the Supplementary material). Surprisingly, AUC was not significantly correlated with sample size.

To see if performance differed significantly among protocols, for each measure of performance we used two-way ANOVA with pre-processing method as a fixed effect and individual species' identities as a random effect, using only species for which we obtained consistent models. Using species identity as a factor accounted both for the effect of sample size (which was shown to significantly affect performance) and for differences among species in their ability to be modelled with the observed environmental variables. Not surprisingly, species identity had a significant effect on each measure of performance ($P < 0{\cdot}0001$). There was no evidence for an effect of modelling method on recall ($P = 0{\cdot}248$) or false-negative rate ($P = 0{\cdot}248$), which was the target of optimization and so was expected to be approximately the same across methods. Modelling method did have a significant effect on false-positive rate ($P < 0{\cdot}0001$), precision ($P < 0{\cdot}0001$), f_1 ($P < 0{\cdot}0001$) and AUC ($P < 0{\cdot}0001$). The group mean performances for each pre-processing method across species showed that, where consistent models could be obtained, methods 1 and 2 typically performed similarly and were superior to method 3 (Fig. 2).

To see if accuracy was driven by the idiosyncratic response of different sampling locations, i.e. if some locations were consistently unpredictable while others were consistently predictable, we compared the predictions (method 1) for each species at each sampling location in the testing data with known occurrence, and

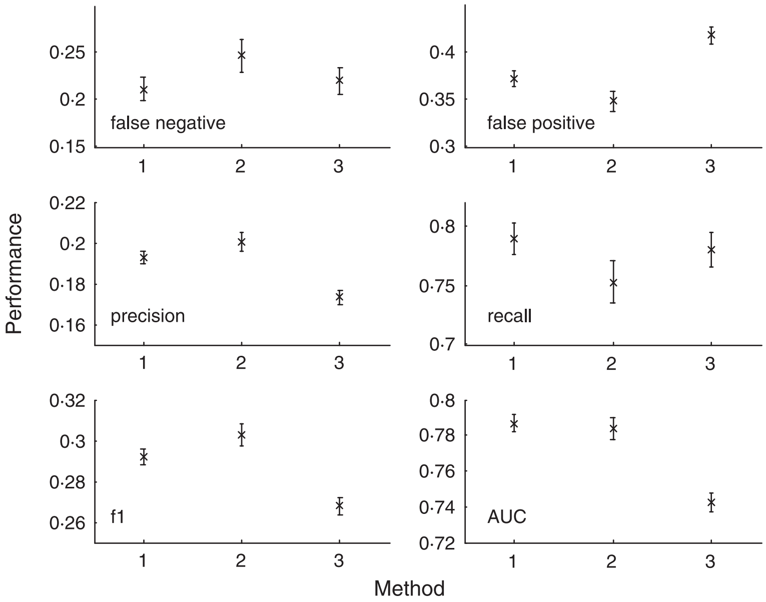

Fig. 2. Means and standard errors of performance criteria by modelling method. Method 1 is for models without pre-processing, using nine environmental variables. Method 2 is for models pre-processed with *k*-whitening, using nine environmental variables. Method 3 is for models using only four environmental variables.

performed two-way ANOVA with species and sampling site as factors. Both effects were highly significant ($P < 0{\cdot}0001$) but explained a small portion of the overall variation ($R^2 = 0{\cdot}203$, partial $R^2_{\text{species}} = 0{\cdot}154$, partial $R^2_{\text{location}} = 0{\cdot}049$).

Discussion

We used a class of recently developed machine-learning algorithms (support vector machines) to model species' distributions using only data concerning species' occurrences. This method assumes only that observations reasonably represent the range of habitable environments; in particular, independence is not assumed. Thus, where data are available only concerning species presence these methods are theoretically superior to classification/discrimination techniques (Hirzel *et al*. 2002). We note that SVM can also be used when there are observations of both habitat and non-habitat (i.e. confirmed absence of species) and indeed can be optimized to make use of as many observations of absence and presence as are available, without requiring balanced observations. We emphasize that the SVM models the support of the statistical distribution of environments from which the species presence observations are drawn, an environmental hyperspace. Thus, the interpretation of the SVM model as an ecological niche is consistent with the classical definition of a niche as a multidimensional environmental space (Hutchinson 1957). Logistic regression (Keating & Cherry 2004), MAXENT (Phillips, Dudík & Schapire 2004), ENFA (Hirzel *et al*. 2002) and other models based on probability densities (Robertson, Caithness & Villet 2001) represent the relative frequency of habitat use and are therefore more closely related to the idea of resource utilization or resource selection (Schoener 1989; Boyce *et al*. 2002; Keating & Cherry 2004).

Of course, theoretical warrant for using support vector machines to model habitat would be unimportant if models performed poorly in independent validation. We used independent observations of species' presences and absences to estimate model accuracy. Summary measures of model performance were generally high. For instance, using our best procedure (method 1) the average AUC obtained was 0·79. For comparison, an analysis of 30 bird species in the Catalan region (Brotons *et al*. 2004) obtained an average AUC of 0·74 on independent data for a model fit using ENFA with only observations of species' presences (Hirzel *et al*. 2002), and 0·82 for logistic regression fit to both species' presences and absences. Zaniewski, Lehmann & Overton (2002) used generalized additive models (GAM) with a logistic link function and binomial distribution fit to both presences and absences of 43 fern species sampled at 19 875 plots in New Zealand to obtain an average AUC of 0·86. Thus, our results are comparable with published results obtained with data for only species presence and data comprising both presence and absence.

We also studied pre-processing approaches that might be taken to increase model performance. Method 1 used no pre-processing or data reduction. Method 2 pre-processed training data using the technique of *k*-whitening. Method 3 relied on a restricted data set in which highly correlated variables were removed from the model training data set. We found that when consistent models could be obtained, method 1 resulted in models with the highest recall and lowest false-negative rate. In contrast, method 2 resulted in models with the highest precision and lowest false-positive rate. Methods 1 and 2 performed similarly as evaluated according to the summary measures f_1 and AUC. In comparison, method 3 performed poorly overall. Consistent models were obtained using method 1 much more frequently than using method 2. Thus, method 1 appears to be the most reliable method in general. Finally, we observed that the relative performance of method 3 compared with methods 1 and 2 indicates that useful information can be obtained by the addition of more environmental variables, even if they are highly correlated.

Finally, we studied how model performance depends on the sample size of the training data set. For 106 species with all three methods we were almost always able to identify a consistent model when the model training data set contained at least 40 observations, which we suggest is the minimum number of observations with which models should be trained in practice. Not unexpectedly, measures of accuracy, such as error rates and precision, were also related to sample size (see Table S4 in the Supplementary material). Minimum sample sizes for modelling and heuristics about how sample size should scale with the number of environmental variables are important topics for research. Curiously, when all species were considered together, AUC was not significantly related to sample size, although the lowest observed AUC scores were always obtained for species represented by fewer than 30 observations (see Figures S6–S8 in the Supplementary material). These results are promising and indicate that often the most accurate models can be obtained with relatively modest data sets. Indeed, models obtained for species with only 40–50 observations routinely performed as well as models for species represented by more than 100 observations. Only precision seemed to increase continuously over the entire range (see Figures S3–S5 in the Supplementary material). These results are about the same as for GARP, which is another machine-learning algorithm and on average obtains near maximal accuracy with 50 observations (Stockwell & Peterson 2002). In contrast, to obtain similar accuracy with logistic regression required 100 observations (Stockwell & Peterson 2002).

An important unanswered question is how many environmental variables are required to predict accurately species' potential distributions, whether with support vector machines or any other technique. In our study, the method with the greatest number of variables (method 1) and no pre-processing provided the best results. An underlying worry is that the higher dimensionality of this method leads to a model that is overfit and would generalize poorly. We overcame this obstacle by fixing a target error rate and tuning models using cross-validation, which estimates the generalization error directly. Thus, our comparison of the different methods was designed to create the fairest comparison: each method was optimized to achieve (approximately) the same generalization error. Three lines of evidence point to success at achieving this fair comparison. First, we failed to detect an effect of modelling method on the false-negative rate when the model was tested with independent data. Thus, the true generalization error was consistent across methods. Secondly, if correlation among environmental variables had led to overfitting, method 2 would have performed best as the algorithm would only have been trained on the information contained in the first few principal components of the data. Finally, in image-recognition experiments (classifying digital images of hand-written numerals) Tax & Müller (2004) found that the optimized model was sometimes not complex enough when non-target observations (i.e. species' absences) were too close to the training data. Therefore, if anything, there is reason to suspect that our models are underfit rather than overfit. Indeed, simply including more environmental variables, rather than developing more sophisticated ways of reducing dimensionality, might result in the greatest improvements to accuracy. Our analysis was limited by the availability of relevant systematically collected data that had been geo-referenced to the particular sampling sites where our species' distribution records were collected. Future studies could certainly include many more variables as the computational cost that would be imposed is minor. Indeed, we suggest that the computational complexity of the SVM approach is one of its primary features. This aspect could be exploited in many ways that await development. Some obvious possibilities are that different 'submodels' obtained from subsets of the data corresponding to classes of environmental variables (biotic vs. abiotic, chemical vs. physical, etc.) could be compared to explore how these differently affect species' distributions; modelling could be embedded in a non-parametric bootstrap to obtain confidence bounds on the estimated distributions; and resampling schemes could be devised to test hypotheses about niche differentiation, partitioning and competitive exclusion or facilitation. These possibilities, together with the relatively strong performance already shown by this approach, should motivate further research, resulting in both methodological improvements and applications in many areas.

Species distribution modelling is a part of many ecological applications, including forecasting species invasions, devising protocols for biodiversity monitoring, designing nature reserves and planning for habitat conservation, managing vector-borne and environmentally mediated disease, and cultivating renewable resources (e.g. aquaculture and timber). Finally, species distribution modelling is often a fundamental component of projects aimed at understanding the consequences of anthropogenic climate change, such as the MODIPLANT initiative that generated the data used in this study. As SVM are stable algorithms that can deal with large sets of predictors at once, they may prove particularly useful in this arena. In conclusion, these results support the continued use of SVM for ecological niche modelling. Where data are available concerning only species' presences and not species' absences, support vector machines are theoretically superior to classification techniques that rely on simulation of pseudo-absence data. We have shown that support vector machines are also comparable with such models when validated by independent observations of both species presence and absence.

Acknowledgements

We thank Niklaus E. Zimmermann, Felix Kienast and Massimiliano Zappa for making their climate maps available. We additionally thank Stéphanie Maire, Dario Martinoni, Pascal Vittoz, Simone Peverelli and many others for their help with the field sampling of species data. Comments by A. T. Peterson and an anonymous referee resulted in a considerably improved paper. This work was conducted while J. M. Drake was a Postdoctoral Associate at the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant No. DEB-94-21535), the University of California at Santa Barbara, and the State of California. The Swiss Office for the Environment (Climate Change Section) and the Conservation Centre of the Canton de Vaud both provided funding to support the field work.

References

Anderson, R.P., Lew, D. & Peterson, A.T. (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, **162**, 211–232.

Anonymous (2004) *ArcInfo*, Version 9·0. Environmental Systems Research Institute Inc., Redlands, CA.

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

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Bradley, A. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**, 1145–1159.

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

Dirnböck, T., Dullinger, S. & Grabherr, G. (2003) A regional impact assessment of climate and land-use change on alpine vegetation. *Journal of Biogeography*, **30**, 401–417.

Drake, J.M. & Bossenbroek, J.M. (2004) The potential distribution of zebra mussels in the United States. *Bioscience*, **54**, 931–941.

Duin, R.P.W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D. & Tax, D.M.J. (2004) *Prtools4, a Matlab Toolbox for Pattern*

Recognition. Delft University of Technology, Delft, the Netherlands.

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.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51**, 331–363.

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.

Graham, C.H., Ron, S.R., Santos, J.C., Schneider, C.J. & Moritz, C. (2004) Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. *Evolution*, **58**, 1781–1793.

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.

Guo, Q., Kelly, M. & Graham, C.H. (2005) Support vector machines for predicting distribution of sudden oak death in California. *Ecological Modelling*, **182**, 75–90.

Gurtz, J. *et al.* (2003) A comparative study in modelling runoff and its components in two mountainous catchments. *Hydrological Processes*, **17**, 297–311.

He, F., Zhou, J. & Zhu, H. (2003) Autologistic regression model for the distribution of vegetation. *Journal of Agricultural, Biological and Environmental Statistics*, **8**, 205–222.

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

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415–427.

Keating, K.A. & Cherry, S. (2004) Use and interpretation of logistic regression in habitat-selection studies. *Journal of Wildlife Management*, **68**, 774–789.

Körner, C. (2003) *Alpine Plant Life*. Springer, Berlin, Germany.

Kumar, L., Skidmore, A.K. & Knowles, E. (1997) Modelling topographic variation in solar radiation in a GIS environment. *International Journal for Geographical Information Science*, **11**, 475–497.

Leathwick, J.R. & Austin, M.P. (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology*, **82**, 2560–2573.

Lusk, J.J., Guthery, F.S. & DeMaso, S.J. (2002) A neural network model for predicting bobwhite quail abundance in the Rolling Red Plains of Oklahoma. *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds J.M. Scott, P.J. Heglund, M. Morrison, M. Raphael, J. Haufler, B. Wall & F.B. Samson), pp. 345–355. Island Press, Covello, CA.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K. & Schölkopf, B. (2001) An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, **12**, 181–202.

Pearson, R.G., Dawson, T.P., Berry, P.M. & Harrison, P.A. (2002) SPECIES: a spatial evaluation of climate impact on the envelope of species. *Ecological Modelling*, **154**, 289–300.

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

Peterson, A.T., Sanchez-Cordero, V., Beard, C.B. & Ramsey, J.M. (2002) Ecologic niche modeling and potential reservoirs for Chagas disease, Mexico. *Emerging Infectious Diseases*, **8**, 662–667.

Phillips, S.J., Dudík, M. & Schapire, R.E. (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*. ACM Press, New York, NY.

Randin, C., Dirnböck, T., Dullinger, S., Zappa, M., Zimmermann, N.E. & Guisan, A. (in press) Are niche-based species distribution models transferable in space? *Journal of Biogeography*.

Robertson, M.P., Caithness, N. & Villet, M.H. (2001) A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*, **7**, 15–27.

Schoener, T.W. (1989) The ecological niche. *Ecological Concepts* (ed. J.M. Cherrett), pp. 79–113. Blackwell Scientific Publications, Oxford, UK.

Schölkopf, B. & Smola, A. (2001) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. & Williamson, R.C. (2001) Estimating the support of a high-dimensional distribution. *Neural Computation*, **13**, 1443–1471.

Scott, J.M., Heglund, P.J., Morrison, M., Raphael, M., Haufler, J., Wall, B. & Samson, F.B. (2002) *Predicting Species Occurrences: Issues of Accuracy and Scale*. Island Press, Covello, CA.

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

Stockwell, D. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Tax, D.M.J. (2001) *One-class classification: concept-learning in the absence of counter-examples*. Thesis. Delft University of Technology, Delft, the Netherlands.

Tax, D.M.J. (2005) *Data Description Toolbox (Manual) V·1·1·2*. Delft University of Technology, Delft, the Netherlands.

Tax, D.M.J. & Duin, R.P.W. (2004) Support vector data description. *Machine Learning*, **54**, 45–66.

Tax, D.M.J. & Juszczak, P. (2003) Kernel whitening for one-class classification. *International Journal of Pattern Recognition and Artificial Intelligence*, **17**, 333–347.

Tax, D.M.J. & Müller, K.-R. (2004) A consistency-based model selection for one-class classification. *Proceedings 17th International Conference on Pattern Recognition (22–26 August 2004, Cambridge UK)* (eds J. Kittler, M. Petrou & M. Nixon), pp. 363–366. IEEE Computer Society, Los Alamitos, CA.

Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., Ferreira de Siqueira, M., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Van Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L. & Williams, S.E. (2004) Extinction risk from climate change. *Nature*, **427**, 145–147.

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

Zimmermann, N.E. & Kienast, F. (1999) Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *Journal of Vegetation Science*, **10**, 469–482.

Received 17 June 2005; final copy received 14 November 2005
Editor: Rob Freckleton

Supplementary material

The following supplementary material is available as part of the online article (full text) from http://w.w.w.blackwell-synergy.com.

Table S1. Species for which ecological niche models were trained using support vector machines ($n = 106$).

Table S2. Predictor variables used to model ecological niches.

Table S3. Pearson correlations for environmental variables used to model ecological niches ranked by the absolute value of the correlation coefficient.

Table S4. Spearman rank-order correlations between performance criteria and sample size.

Figure S1. Rescaled histograms of nine predictor variables used to model ecological niches.

Figure S2. Frequency with which consistent models could be obtained.

Figure S3. Performance of method 1.

Figure S4. Performance of method 2.

Figure S5. Performance of method 3.

Figure S6. Summary measures of performance for method 1.

Figure S7. Summary measures of performance for method 2.

Figure S8. Summary measures of performance for method 3.