Received Date: 01-Aug-2013

Accepted Date: 09-Jun-2014

Article Type: Research Article

Editor : Pedro Peres-Neto

PATTERN RECOGNITION ECOLOGICAL NICHE MODELS FIT TO PRESENCE-ONLY AND PRESENCE-ABSENCE DATA

Running Title: Pattern recognition niche models

Sean P. Maher[1,2], Christophe F. Randin[3], Antoine Guisan[4,5], and John M. Drake[1]

1. Odum School of Ecology, University of Georgia, Athens, GA 30602-2202
2. Current address: Department of Environmental Science, Policy, and Management & Museum of Vertebrate Zoology, University of California Berkeley, Berkeley, CA 94720-3114
3. Institute of Botany, University of Basel, Schönbeinstrasse 6, CH - 4056 Basel, Switzerland
4. Department of Ecology and Evolution, University of Lausanne, Biophore, CH-1015 Lausanne, Switzerland
5. Institute of Earth Sciences, University of Lausanne, Geopolis, CH-1015 Lausanne, Switzerland

Corresponding author: SPM

Department of Environmental Science, Policy, & Management

University of California, Berkeley

130 Mulford Hall #3114

Berkeley, California 94720-3114

smaher02@gmail.com

Summary

1. Identifying the boundary of a species' niche from observational and environmental data is a common problem in ecology and conservation biology and a variety of techniques have been developed or applied to model niches and predict distributions. Here, we examine the performance of some pattern recognition methods as ecological niche models (ENMs). Particularly, one-class pattern recognition is a flexible and seldom used methodology for modeling ecological niches and distributions from presence-only data. The development of one-class methods that perform comparably to two-class methods (for presence/absence data) would remove modeling decisions about sampling pseudo-absences or background data points when absence points are unavailable.
2. We studied 9 methods for one-class classification and 7 methods for two-class classification (5 common to both), all primarily used in pattern recognition and therefore not common in species distribution and ecological niche modeling, across a set of 106 mountain plant species for which presence-absence data was available. We assessed accuracy using standard metrics and compared trade-offs in omission and commission errors between classification groups as well as effects of prevalence and spatial autocorrelation on accuracy.

3. One-class models fit to presence-only data were comparable to two-class models fit to presence-absence data when performance was evaluated with a measure weighting omission and commission errors equally. One-class models were superior for reducing omission errors (i.e. yielding higher sensitivity) and two-classes models were superior for reducing commission errors (i.e. yielding higher specificity). For these methods, spatial autocorrelation was only influential when prevalence was low.
4. These results differ from previous efforts to evaluate alternative modelling approaches to build ENM and are particularly noteworthy because data are from exhaustively sampled populations minimizing false absence records. Accurate, transferable models of species' ecological niches and distributions are needed to advance ecological research and are crucial for effective environmental planning and conservation; the pattern-recognition approaches studied here show good potential for future modeling studies. This study also provides an introduction to promising methods for ecological modeling inherited from the pattern recognition discipline.

Keywords: machine learning, species distribution model, Swiss flora, realized distribution, potential distribution

Introduction

The relation between a species' distribution and its environment is determined by its fundamental niche, modified by dispersal limitations, perturbations and biotic interactions (Pulliam 2000, Williams et al. 2007, Thuiller et al. 2013). Essentially, the niche, as described by Grinnell (1917) and Hutchinson (1957), is the set of environmental conditions over which a species average fitness is positive, which translates into the habitable regions of a landscape (Chase and Leibold 2003, Soberón 2007). The niche concept, and its representation as a set of suitable environmental conditions, is now foundational to many core ecological and evolutionary theories (Holt 2009) and central to the debate over neutral theory (Adler et al. 2007, Hubbell 2001). With continued development of technologies associated with geographic information systems and acquisition of distributional and environmental data, the practice of constructing models of species' niches for research and conservation will undoubtedly continue to increase (Guisan and Thuiller 2005, Elith & Leathwick 2009, Peterson et al. 2011, Guisan and Rahbeck 2011, Guisan et al. 2013).

It has become customary to categorize these models, whether referring to them as ecological niche models (ENMs) or niche-based species distribution models (SDMs; see Guisan et al. 2013 S1), according to the type of data used in model fitting: “presence/absence” models are trained on both positive occurrences (sampled point locations where the species has been documented) and negative occurrences (sampled point locations where the species has not been documented), while “presence-only” models are trained using only positive occurrences (Hirzel et al. 2002, Brotons et al. 2004, Elith et al. 2006, Tsoar et al. 2007). A previous comparative study of logistic regression (a presence-absence model) and Ecological Niche Factor Analysis (a presence-only model; Hirzel et al. 2002) found the presence-absence model to be superior and preferable “in most situations" (Brotons et al. 2004), and Elith et al. (2006) found presence-absence algorithms to be superior in general. As there may be conceptual grounds for preferring presence-only methods or because only presence data are available (i.e. avoiding pseudo-absence data; Drake et al. 2006, Drake & Bossenbroek 2009, Hastie and Fithian 2013), a major goal of our study was to introduce techniques from the pattern recognition field, and determine if differences between presence-only and presence-absence implementations yield similar conclusions.

Regardless of the data used for fitting, the goal of an ENM is to delineate regions of environmental space that designate niche environments from non-niche environments (Figure 1). Thus, niche modeling poses a boundary estimation problem (Cuevas and Fraiman 2009) with proposed solutions ranging from simple bounding box representations (Nix 1986, Busby 1991) to complex statistical or computational models (e.g., Stockwell and Peters 1999, Hirzel et al. 2002, Drake et al. 2006, Elith et al. 2006; but see Godsoe 2010 for an alternative view). Similar problems have been studied in the machine learning and pattern classification literatures, in which "one-class classification" is a solution when presence-only data are available (Tax & Duin 1999, Tax 2001) and classical two-class classification is a solution for presence/absence datasets (van der Heijden et al. 2004). Unfortunately, some terminology that may be in common usage will be ambiguous with respect to our task, for instance whether "classifier" refers to a family of models (e.g. "neural network"), a specific untrained instance of that family (e.g. "single-layer radial basis network"), the algorithm used for training that model, or a trained model that can be used to classify new cases. For clarity, we use "mapping" to refer to a specific trained instance, as it maps a set of inputs *x* (environmental variables) onto a response *y* (niche or non-niche). We use "representation" to refer to the more general untrained class of models (e.g., "nearest neighbor model") which might be implemented or tuned in different ways (i.e., the algorithm by which the set of nearest neighbors to a point is identified). Some representations may be implemented in either a one-class or a two-class "form", resulting in either a set boundary or a decision boundary (Figure 1) and yielding different mappings for empirical comparison of one-class or two-class models. Just as several currently popular methods for presence-only and presence/absence modeling were adapted from solutions known to solve analogous problems in other disciplines (Elith et al. 2006, Phillips et al. 2006), we studied these pattern recognition approaches as ways to estimate the niche/non-niche boundary from occurrence data.

We tested one-class mappings and two-class mappings of representations of various levels of complexity, using records of a well-surveyed alpine Swiss flora (Drake et al. 2006, Randin et al. 2006). Often, it has been determined or assumed that mappings fit to two-classes of data would be generally superior to one-class mappings (e.g., Brotons et al. 2004, Peterson et al. 2011). In our view, this is an empirical, not conceptual, question. We observe that if absences are common in a dataset (either because of sampling error or because the species prevalence on the landscape is low), then one-class mappings may actually be expected to accurately predict new positive occurrence data better than two-class mappings, resulting in a reduced false negative rate in test examples (reduced omission error). Alternatively, we recognize that it is generally difficult to accurately estimate the boundary of a multidimensional distribution if the data are sparse or unevenly sampled–both of which are common problems with ENMs. In such cases, the absence (or pseudo-absence) data used by two-class mappings might stabilize the estimate of the boundary (i.e., low variance with respect to perturbations of the training data set) and result in models of more limited volume and with reduced false positive errors on new data (reduced commission error). However, absences can in turn increase the rate of false negatives when they are caused by human and natural (e.g., geomorphic) perturbations that prevent species from occurring at locations that are otherwise suitable (Randin et al. 2009a, b). Essentially, one-class mappings based on environmental data may better represent conditions associated with the Grinnellian niche (sensu Soberón 2007), whereas two-class mappings will tend to designate the realized niche and distribution.

Because overall accuracy is a combination of both kinds of error and because the amount of data corruption (e.g., false absence records) and the dimensionality of the true niche can vary from case to case, we submit that if there is any general rule about the superiority of one-class or two-class forms, then it must be at best a generalization that could be discovered through comparative analysis (Guo et al. 2005). This study aims thus to: (i) introduce alternative techniques to model species' ecological niches and distributions, taken from the field of pattern recognition; and (ii) compare, within these, measures of accuracy of one-class and two-class forms. We found that one-class mappings were comparable to two-class mappings with respect to a common measure of discriminative performance, and had preferable (lower) rates of omission error; commission error was typically lower in two-class mappings. Finally, the number of positive occurrence records used for model fitting did not improve models on average, although the variability in accuracy declined with sample size.

Methods

A first set of species distribution data was collected by a team led by AG and CR from 2002 to 2004. The study region was located in the Western Alps in the state of Vaud, Switzerland, and comprised 550 vegetation survey plots (Figure 1). The set of 106 species included in this study are easy to detect by experts in the field so above-ground vegetation may be considered comprehensively sampled. Covariates include only those conjectured to have a physiological effect on species (Randin et al. 2006) and were sampled from interpolated topographic and climatic GIS layers at plot coordinates. Variables included measures of moisture and precipitation, solar radiation, snow cover and topography. Additional details concerning this subset of the data are reported in Drake et al. (2006). Spatial autocorrelation was evaluated by generating semi-variograms for each predictor variable using the gstat package (Pebesma 2004) and visual inspection to determine detectable ranges and sills. To robustly assess the predictive power of models, we used a dataset collected in 2009 consisting of 312 plots that did not spatially coincide with any sampling sites in the initial dataset (Pottier et al. 2013). Because there is no suitable way to split the data into geographically distinct subsets given the study area configuration and spatial correlation was avoided between the two sets (Pottier et al. 2013), these data represent a useful time-for-space substitution for independent evaluation. Hereafter we refer to this dataset as the transfer dataset.

To compare the accuracy of one-class and two-class mappings we selected 11 different model representations, five of which can be tuned as a one-class form (OCF) or two-class form (TCF), yielding 16 mappings (Table 1). Representations varied between density-based representations (derived from classical methods in probability and statistics; similar to BIOCLIM (Busby 1991) and Ecological Niche Factor Analysis (Hirzel et al. 2002)), distance-based representations (emphasizing similarity to known examples; similar to DOMAIN (Carpenter et al. 1993)), and concept learning representations (algorithmic representations from machine learning theory; some techniques included here have been used elsewhere (e.g. Guo et al. 2005, Drake et al. 2006)). Details for each of the pattern recognition representations are supplied in Supplementary Table 1. Most other common ENM approaches – including Maxent, generalized linear models (GLMs), generalized additive models (GAMs), generalized boosting machines (GBM), and random forests – do not correspond to these categories and do not admit one-class tuning. For these methods, we followed the recommendation of Barbet-Massin et al. (2012) to limit the amount of absence data to the number of observations for each species to train and test these approaches. Details for fitting and calibrating these methods are provided in Supplementary Table 2.

For evaluation, data were randomly split into independent training (80%; 440 plots) and testing sets (20%; 110 plots) following Drake et al. (2006). All pattern recognition models were fit using the Pattern Recognition Tools (www.prtools.org) and DD Tools (http://prlab.tudelft.nl/david-tax/dd_tools.html) MATLAB libraries in MATLAB R2009b (van der Heijden 2004, Tax 2012; Supplementary Table 1). PR Tools provides routines to fit two-class mappings for a variety of representations, while DD Tools provides functionality for tuning one-class mappings. Two-class mappings were optimized to maximize AUC in the training set. One-class mappings were optimized for a target rejection threshold of 5% in the training set. That is, applying the trained mapping to the training data would classify the 5% most extreme cases as non-suitable, where what counts as most extreme varied according to the representation.

To compare performance among trained mappings, we used withheld testing data to calculate three measures of accuracy commonly used in ENMs: one measure, the area under the curve (AUC) of a receiver-operator plot, comparing predicted probabilities to presence-absence observations, and two measures, sensitivity, and specificity, comparing binarized predictions to presences-absences (Swets 1988). We also calculated two performance measures for binary predictions more commonly used in engineering: precision (or positive predictive value) and overall accuracy (van der Heijden 2004). In each of these measures, performance increases as the summary statistic, ranging from zero to one, increases. Sensitivity and precision quantify the mapping's ability to identify positive instances (i.e., presences), with sensitivity reporting the proportion of true positives predicted correctly and precision reporting the proportion of true positive of all points predicted positive. Specificity quantifies the ability of a mapping to exclude negative cases. Overall accuracy is the proportion of cases correctly predicted. Of course, a good model should discriminate presence from absence, which is quantified in the form of the AUC (Fawcett 2006), which takes values between 0 to 1, where values approaching 1 reflect agreement with data, values approaching 0 reflect an inverse relationship, and a fair coin yields AUC=0.50. To be more conservative, we evaluated the number of models in which the AUC value exceeded 0.70 and determined if this proportion was attributable to random chance.

Performance measures for each form of each representation were statistically summarized to investigate effects of species identity, methodology (representation and one-class vs. two-class form) and sample size. With the exception of overall accuracy, we square-root-arcsine transformed performance measures and calculated means, standard deviations, and coefficients of variation over species, mapping, and for pools of species grouped by frequency of occurrence in the training data for use in parametric analysis, unless otherwise noted. To test for an effect of species identity and the one-class vs. two-class forms, we used an ANOVA with both species and form as factors and an interaction term. Further, we examined the effect of prevalence on AUC (e.g., Wisz et al. 2008) using AIC values of linear models to discriminate among candidate models and predictor variables (Akaike 1974, Burnham and Anderson 2002; e.g., Wisz et al. 2008). We explored whether high or low prevalence (based upon whether the sample size was in the upper or lower half of the training set) yielded different trends through a fixed effect, and if there were effects of one-class versus two-class form or specific representation (e.g., Parzen density estimator versus support-vector machine) using random effects to account for within-group variation in an efficient manner. Five mappings had corresponding one-class and two-class forms. For these we used a two-sided paired t-test to test for differences in transformed values, where differences were calculated between the pair of performance measurements for each species. We tested for an effect of spatial clustering in sampling using a linear model of mean transformed AUC value and mean nearest neighbor distance, as well as this distance adjusted for sample size. If spatial sampling biased our results, we would expect a relationship with either positive or negative slopes. Statistical analyses were performed in R (c. 3.0.2; R Core team 2013).

Finally, we applied the mappings to the transfer set, and calculated AUC, sensitivity, specificity, and precision. We analyzed these measures to determine if differences in one-class and two-class forms were consistent with observations from the testing data. We used a subset of statistical analyses from above to examine whether these approaches correctly predicted data from a novel set. To measure the number of “good” mappings, we used a less conservative AUC value threshold of 0.6.

Results

Frequency of occurrence in the training data set ranged from 5 to 155 (median 28.5) and exhibited strong right skew (Figure 2). Frequency of occurrence in the test set was similarly skewed (minimum=1, maximum=38, median=7; Figure 2). Visual inspection of semivariograms did suggest spatial autocorrelation in the predictor variables, although semivariance values were quite low (< 0.08) across variables (Supplemental Figure 1), and mean distances between points generally were well beyond the range (Supplemental Figure 2).

Summaries of model performance by representation and one-class vs. two-class forms are shown in Figure 4. Overall, models predicted species presence better than random (Table 1; untransformed AUC values were above 0.7 in 1,164 of 1,696 or 68.63% of models; binomial test $p<0.001$). In general, one-class mappings (640 of 954, 67.1%) were no more likely than two-class mappings (524 of 742, 70.6%) to meet this minimal criterion ($\chi^2_1=3.688$, $p=0.055$). Of the poorly fitting mappings, the clear majority (86 of 98) were estimated from species with low prevalence (less than 29 positive observations; Figure 4); one species, *Cirsium oleraceum* (Cabbage thistle), performed poorly in 14 of the 16 mappings, probably due to only 2 positive observations in our testing set, although other species with similarly low testing prevalence did not perform as poorly. The outcome for this species is probably anomalous and due to the random construction of the test data set. There was no evidence for a difference in untransformed AUC between the one-class and two-class forms ($\bar{x}_{OCF}=0.75$ vs. $\bar{x}_{TCF}=0.75$; Wilcoxon rank test for two samples, $W=343687.5$, $p=0.306$). These values were comparable to values for two-class methods frequently used ($\bar{x}_{GAM}=0.786$, $\bar{x}_{GBM}=0.784$, $\bar{x}_{GLM}=0.742$, $\bar{x}_{Maxent}=0.794$, $\bar{x}_{RF}=0.820$). Overall accuracy was greater in two-class forms than one-class forms ($\bar{x}_{OCF}=0.61$, $\bar{x}_{TCF}=0.89$; $W= 142022$, $p<0.001$) and was correlated with the number of true negatives ($r=0.986$, $t_{1694}=246.823$, $p<0.001$). Mappings from one-class forms were more sensitive ($\bar{x}_{OCF}=0.66$ vs. $\bar{x}_{TCF}=0.24$; Wilcoxon rank test for two samples, $W=560943$, $p<0.001$), less specific ($\bar{x}_{OCF}=0.58$ vs. $\bar{x}_{TCF}=0.94$; Wilcoxon rank test for two samples, $W=140447.5$, $p<0.001$) and less precise ($\bar{x}_{OCF}=0.15$ vs. $\bar{x}_{TCF}=0.27$; Wilcoxon rank test for two samples, $W=140447.5$, $p<0.001$) than two-class forms. Results restricted to species with prevalence in the upper 0.5 quantile were similar (Supplementary Table 3). Within the one-class forms, the Parzen density mapping (see Supplement for definition) performed poorly in sensitivity and precision, especially within the upper 0.5 prevalence quantile.

Comparison of species with respect to form and representation

Analysis of effects on transformed AUC within one-class and two-class forms showed effects of species and representation, regardless of form (OCF: species—$F_{105,840}$=12.67, p<0.001 and mapping—$F_{8,840}$=30.73, p<0.001; TCF species—$F_{105,630}$=6.636, p<0.001, mapping—$F_{6,630}$=22.438, p<0.001). Species identity had an effect on transformed AUC value for the complete data set ($F_{105,1484}$=14.375, p<0.001), but form and the interaction between form and species did not ($F_{1,1484}$=0.210, p=0.647, interaction—$F_{105,1484}$=0.731, p=0.980). Restricting analysis to species in the upper 0.5 frequency of occurrence quantile revealed effects of both species ($F_{52,742}$=13.396, p<0.001) and form ($F_{1,742}$=26.396, p<0.001), but not an interaction ($F_{52,742}$=0.443, p=0.9998). Species explained a greater proportion of variation (0.484) than form (0.034).

Comparison of prevalence with respect to form and representation

The simple linear model of AUC from prevalence yielded a negative slope ($\beta_{prevalence}$= -0.004, t_{1694}= -2.878, p=0.004, AIC= -1213.339), but the model accounting for potential difference between high and low prevalence yielded greater support ($\beta_{prevalence}$=0.001, t_{1694}=0.960, p=0.337, $\beta_{quantile}$=0.082, t_1=3.998, p<0.001, interaction= -0.002, t_1= -2.038, p=0.042, AIC= -1228.864). The mixed effect models accounting for form as a random effect was uninformative ($\beta_{prevalence}$=0.001, t= 0.960, p=0.337, $\beta_{quantile}$=0.082, t=3.998, p<0.001, interaction= -0.002, t= -2.038, p=0.042, AIC= -1226.864). However, accounting for representation improved explanatory power ($\beta_{prevalence}$=0.001, t= 1.001, p=0.317, $\beta_{quantile}$=0.082, t=4.172, p< 0.001; interaction= -0.002, t= -2.137, p=0.033; AIC= -1342.689). Finally accounting for both form and representation yielded the most supported model ($\beta_{prevalence}$=0.001, t=1.001, p=0.315; $\beta_{quantile}$=0.082, t=4.192, p< 0.001, interaction= -0.002, t= -2.137, p=0.033, AIC -1346.458). Linear models of mean AUC as a function of clustering in training data showed a negative effect ($\beta_{clustering}$= -3.175e-05, t= -3.248, p=0.002, AIC = -155.204). Similar patterns emerged when adjusting clustering for prevalence

($\beta_{clustering\_adj}$= -1.307e-04, t_1= -2.13, p=0.035), but this model was less supported (AIC = -149.513). Allowing for differences between high and low abundant species in how clustering affects AUC, the model maintained this negative effect overall ($\beta_{clustering}$= -4.835e-05 t1 -4.035, p<0.001), the effect was stronger for abundant species ($\beta_{quantile}$= -0.063, t1= -2.305, p=0.023) and was more supported (AIC= -158.537).

Comparing representation between forms

The one-class form of a representation yielded a significantly higher transformed AUC than the two-class form in four out of five representations available in both one-class and two-class forms (binomial test p≈1). These were the K-nearest neighbor, Parzen density estimator, Gaussian density, and support vector machine (Supplementary Table 4). The pattern changed slightly when analysis was restricted to species in the upper half of prevalence, where the K-nearest neighbor two-class form was more accurate (Supplementary Table 5). In general, the estimated difference between the forms was low, such that the improvement in non-transformed AUC value for the better scheme would not be expected to be greater than 0.015.

Comparisons of precision, sensitivity and specificity

Transformed sensitivity measures (Supplementary Table 6) were significantly affected by species and form, but there was no evidence for an interaction (species - $F_{105,1484}$=4.077, p<0.001; form – $F_{1,1484}$=681.183, p<0.001; species × form – $F_{105,1484}$=1.188, p=0.100). The ANOVA of transformed specificity values detected effects of species, form, and the interaction (species - $F_{105,1484}$=3.673 p<0.001; form – $F_{1,1484}$=943.370, p<0.001; species × form – $F_{105,1484}$=1.529, p<0.001). Likewise, the ANOVA of transformed precision values (Supplementary Table 6) showed significant effects of species, form, and the interaction (species - $F_{105,1484}$=11.041, p<0.001; form – $F_{1,1484}$=60.511, p<0.001; species × form –

$F_{105,1484}$=2.222, p<0.001). For species in the upper half of prevalence (Supplementary Table 7), the same predictors emerged for sensitivity and precision (sensitivity: species - $F_{52,742}$=1.314, p=0.072; form – $F_{1,742}$=536.825, p<0.001; species × form – $F_{52,742}$=0.948, p=0.580; precision: species – $F_{52,742}$=5.360, p<0.001; form – $F_{1,742}$=164.705, p<0.001; species × form – $F_{52,742}$=1.594, p=0.006), but the interaction term was not significant in the model of specificity (species – $F_{52,742}$=1.365, p=0.049; form – $F_{1,742}$=803.284, p<0.001; species × form – $F_{52,742}$=0.646, p=0.975). Interaction plots (Supplementary Figure 3) show the heterogeneity of responses between species where interactions were significant.

Significant differences in precision, specificity, and sensitivity of representations implemented as one-class and two-class forms were found in 11 of 15 cases (Supplementary Tables 8-10). For sensitivity, significant differences were observed in each case; the Parzen density estimator was significantly greater in the two-class form, while one-class forms were superior in the remaining four representations. The difference in precision between support vector mappings was marginally non-significant in favor of the two-class form (*p*=0.055); the mappings for which there were significant tests favored two-class mappings (nearest neighbor and Gaussian). Specificity was higher in the two-class forms in three of the four significant tests (K-nearest neighbor, nearest neighbor, and Gaussian), with the Parzen density estimator having larger values in the one-class form and support vector machines showing no difference between forms. Restricting results to the 50% most frequently occurring species yielded similar results (Supplementary Tables 11-13).

Comparisons of transferability

Accuracy values from the transfer dataset were consistent with those from the testing set, with respect to trends and which methods and forms performed best (Table 1, Supplementary Table 6, Supplementary Figure 4). Five species (*Artemisia umbelliformis*, *Campanula cenisia*, *Rhododendron hirsutum*, *Rosa pendulina*, and *Saxifraga stellaris*) were absent from the transfer dataset, so AUC and sensitivity comparisons were among the remaining 101 species. AUC values were greater than 0.6 in 1,014 of 1,696 or 59.79% of models; binomial test p<0.001). Differing from the pattern in the testing set, one-class mappings (605 of 954, 63.4%) were more likely than two-class mappings (409 of 742, 55.1%) to meet this minimal criterion (χ^2_1=19.31, p<0.001). ANOVAs of accuracy measures identified the effect of form and species for AUC (species - $F_{100,1414}$=17.892, p<0.001; form – $F_{1,1414}$=16.219, p<0.001; species × form – $F_{100,1414}$=0.904, p=0.737), sensitivity (species - $F_{100,1414}$=3.117, p<0.001; form – $F_{1,1414}$=504.709, p<0.001; species × form – $F_{100,1414}$=1.216, p=0.0782) and specificity (species - $F_{105,1484}$=2.999, p<0.001; form – $F_{1,1484}$=589.981, p<0.001; species × form – $F_{105,1484}$=0.889, p=0.779) without a significant interaction; the ANOVA for precision did show a significant interaction (species - $F_{105,1318}$=22.840, p<0.001; form – $F_{1,1318}$=7.525, p=0.006; species × form – $F_{105,1318}$=1.414, p=0.005). Pairwise comparisons for the matching methods were consistent in that AUC and sensitivity were higher in the one-class forms (Supplementary Tables 14-15) and specificity generally was higher in the two-class forms (Supplementary Table 16). However, precision was higher in the one-class form (Supplementary Table 17).

Discussion

We studied the performance of a variety of pattern recognition techniques for modeling species' ecological niches and predict distributions, and assessed the ability of one-class and two-class mappings to distinguish patterns of positive occurrence records. We found one-class mappings to perform satisfactorily, similar to two-class mappings in term of overall accuracy (AUC), but with higher sensitivity. In turn, two-class mappings yielded higher specificity. This pattern was consistent when we applied mappings to a temporally independent and spatially novel dataset. Presumably, this is because one-class mappings are able to identify a boundary of niche space with less data (e.g. only positive occurrences), and such boundaries are more likely to be inclusive of a broader set of environmental combinations than two-class mappings. The latter are thus better at predicting absences, i.e., where a species does not occur. For instance, within the testing set, a two-class form had the overall highest mean AUC values (0.804, K-nearest neighbor), yet the one-class form yielded a much higher mean value of sensitivity with a similar mean AUC value (0.794). However, the highest AUC value in the transfer set was a one-class form (0.672 support vector machine), and five one-class representations had higher mean AUC values than the highest two-class representation (Table 1).

This research extends previous efforts that used methods from classification and machine learning to fit niche models (Guo et al. 2005, Drake et al. 2006, Stockwell and Peters 2001, Dudik and Phillips 2004, Elith et al. 2006). Contrary to our findings, previous efforts to compare models using presence-only (one-class) and presence-absence (two-class) approaches demonstrated superiority of the latter, based upon higher AUC and COR values (Elith et al. 2006). Pattern recognition methods presented here performed similarly to frequently used presence-absence approaches. While we failed to detect consistent differences in testing AUC between one-class and two-class forms, idiosyncratic differences were detected that were complex and dependent on the species under consideration. For instance, representation interacted with prevalence (see *Comparison of prevalence with form and representation*), so that some methods were more accurate at high prevalence while others were preferable for species with low prevalence. Further, mappings for species with low prevalence were quite variable with respect to testing AUC (Figure 5), suggesting that some methods may not be suitable for species with few occurrence records, particularly support vector machines (in both one and two-class forms), Radial basis neural networks, and the Gaussian classifier in one-class form. The removal of species infrequently sampled reduced the coefficient of variation in testing AUC values for each of these implementations. Clustering of presence data had significant effects on performance, albeit with a small negative slope, and depended upon how we defined clustering. When we compared complementary mappings between forms using paired t-tests for the testing set, 4 of 5 (80%) tests showed higher values for one-class forms. AUC values from the transfer dataset were higher in one-class forms, although there was also a strong effect of species, and pairwise comparisons favored one-class forms for 4 of 5 (80%) representations.

Significant interaction terms in ANOVAs for the testing set make it difficult to separate the influence of species versus form on values of precision and specificity. The interaction was not significant for comparisons of sensitivity and specificity using the transfer dataset, although it was significant for comparison of precision. We suggest that one-class mappings should be preferred as ecological niche models used to answer questions about geographic expansion and potential distribution because they more successfully identified habitable environmental combinations in both the testing and transfer datasets, particularly when the species has reached all boundaries in environmental space. Two-class mappings could be preferable when the focus of investigation requires estimates of potential absence, such as in ENM-based climate-change projections (Guisan et al. 2013). We note the assignment of absence to an observation can be difficult, unreliable, or imprecise and use of absence as a data type can be problematic; in our case, absence data have been selected for maximal reliability. Such well-sampled absence data are rarely available due to site accessibility by the study species (Barve et al. 2011), historical factors (Dullinger et al. 2012), failure to sample, and sampling uncertainty (MacKenzie et al. 2006).

Some investigators have shown there to be strong relationships between sample size and model performance (Peterson and Cohoon 1999, Wisz et al. 2008). An intriguing result of our analysis was that increased frequency of occurrence did not universally improve performance of mappings. Indeed, our mixed effects model detected a very slight overall negative effect for more prevalent species. One possible explanation for this result could be that species with high prevalence are less specialized and thus more difficult to model, whereas species with lower prevalence are environmental specialists with clearer environmental requirements (Guisan & Hofer 2003). Other alternative explanations could be that widespread and abundant species (i.e. those with large source populations) could temporarily inflate their niche by "waves/events of strong dispersal" in a source-sink system when conditions are extremely favorable on the short term (see Fig.1c in Pulliam 2000). In such systems, populations could be decoupled from the long-term topo-climate and more dependent on the spatial history and configurations of the populations, or widespread/abundant species could have been favored under some specific human land-use treatments or natural disturbances not particularly correlated to climate and topography (e.g. N-fertilization; Randin et al. 2009a, b). In addition, this result could be a consequence of sampling variation: the skewed distribution in prevalence has few samples of common species to anchor our estimate of this relationship. Plots of the coefficient of variation for each species show that species with high coefficient of variation are those that showed the greatest prevalence (Figure 4). The consistency in testing AUC of different mappings within a species was much higher when sample sizes were greater than 29, which is consistent with the finding Wisz et al. (2008; 30 observations).

We found performance to vary considerably among species. Any of several explanations might account for this observation. Two prominent (and not mutually exclusives) hypotheses are that (1) spatial autocorrelation in the data yielded spatially biased model fits (which would be more severe in infrequently sampled species) and (2) that sample size was more important, but, due to our fixed prevalence for each species, this effect was aliased by the effect of species. With respect to hypothesis 1, we found little evidence for strong autocorrelation in environmental variables and analysis of the effect of clustering yielded small, negative slopes. We interpret this to suggest clustered datasets may yield a small improvement in model performance, and the effect is stronger in less abundant species. The effect of sample size could possibly be studied further by re-sampling subsets of the original data to separate species identity and the number of records, but how to study this problem in a principled way is not obvious and we did not pursue this idea further here.

How easy to use are these methods? The total amount of new code developed for this project amounted to only a few hundred lines. The classification methods we presented here are implemented in a widely available high-level scientific programming language (MATLAB) and use open source libraries (PR Tools and DD Tools). The more classical methods are all available in the freely available R software (e.g. Elith et al 2006, Thuiller et al 2009). The MATLAB toolboxes themselves are free to download, with source code available to adapt to other, free operating software packages, such as R. Thus, extension of these methods to other applications and mapping in other software should be straightforward.

The pattern-recognition techniques assessed in this study yielded similar AUC values to common techniques, and, thus, expand modeling options for investigators. We found that the one-class mappings may be preferable to the two-class mappings in some instances, mainly when absences are not reliable and evaluation is focused on optimizing sensitivity. The one-class mappings presented here require fewer data, show higher sensitivity than two-class mappings, and tended to estimate broader sets of suitable environments. It has been argued previously that one-class ENMs may be preferred on conceptual grounds for some applications, as one-class models more elegantly represent the niche idea than do two-class models (Drake et al. 2006, Drake & Bossenbroek 2009). Such applications include species introduced to novel geography and range shift forced by climate change. In turn, two-class mappings might be preferred when reliable absence data are available, especially if the focus is on optimizing specificity (i.e. predicting where a species is not). Both mapping forms presented here may be used when presence-absence are available and the focus is on overall evaluation (i.e., balancing sensitivity and specificity). Trade-offs between sensitivity and specificity are important for conservation decisions (Guisan et al. 2013), and the flexibility of these approaches for such assessments are evident.

Acknowledgements

This research was supported by a grant from the U.S. National Oceanic and Atmospheric Administration. The Swiss Office for the Environment (Climate Change Section), the Swiss National Science Foundation (SNF) and the Conservation Centre of the Canton de Vaud (SFFN) funded the initial survey on which this work is based. We thank Pedro Peres-Neto and two anonymous reviewers for comments that improved the manuscript.

Data Accessibility

Data used in this manuscript are available at Dryad Data Repository: doi:XXXXXX

The data already have been archived in the Swiss floristic data base at: XXXXXX

References

Adler, P.B., HilleRisLambers, J. & Levin, J.M.. (2007) A niche for neutrality. *Ecology Letters,* **10**, 95-104.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723

Barbet-Massin, M., F. Jiguet, C. H. Albert, & W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution,* **3**, 327-338.

Busby, J. R. (1991) bioclim – a bioclimatic analysis and prediction system. *Nature Conservation: Cost Effective Biology Survey and Data Analysis* (eds C.R.Margules & M.P.Austin), pp. 64-68. CSIRO, Australia.

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

Burnham, K.P. & Anderson, D.R. (2002) Model Selection and Inference: A Practical Infromation Theoretical Approach. 2nd Ed. New York Springer-Verlag.

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

Chase, J.M. & Leibold, M.A. (2003) *Ecological Niches: Linking classical and contemporary approaches*. University of Chicago Press, Chicago.

Cuevas, A., & Fraiman, R. (2009) *Set estimation*. New Perspectives in Stochastic Geometry, pp 608. Oxford University Press, Oxford.

Dullinger, S., Willner, W., Plutzar, C., Englisch, T., Schratt-Ehrendorfer, L., Moser, D., Ertl, S., Essl, F., & Niklfeld, H. (2012) Post-glacial migration lag restricts range filling of plants in the European Alps. *Global Ecology and Biogeography*, **21**, 829-840.

Drake, J.M., Guisan, A. & Randin, C. (2006) Modelling ecological niches with support vector machines. *Journal of Applied Ecology,* **43**, 424-432.

Drake, J.M. & Bossenbroek, J.M. (2009) Profiling vulnerability to invasion by zebra mussels with support vector machines. *Theoretical Ecology*, **4**, 189-198.

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

Elith, J. & Leathwick, J. (2009) Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, **40**, 677-697

Elith, J., Phillips, S., Hastie, T., Dudik, M., Chee, Y., & Yates, C. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43-57.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27, 861-874.

Godsoe, W. (2010) I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche. *Oikos*, **119**, 53-60.

Grinnell, J. (1917) The niche-relationships of the California Thrasher. *The Auk*, **34**, 427-433.

Guisan, A. & Hofer, U. (2003). Predicting reptile distributions at the mesoscale: relation to climate and topography. *Journal of Biogeography*, **30**, 1233-1243.

Guisan, A. & Rahbeck, C., (2011) SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography,* **38**, 1433-1444.

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

Guisan A., Tingley R., Baumgartner J.B., Naujokaitis-Lewis I., Sutcliffe P.R., Tulloch A.I.T., Regan T.J., Brotons L., McDonald-Madden E., Mantyka-Pringle C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, Y.M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, **16**, 1424-1435.

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.

Hastie, T., & Fithian, W. 2013. Inference from presence-only data; the ongoing controversy. *Ecography*, **36**, 864-867.

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

Holt, R.D. (2009) Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences USA*, **106**, 19659-19665.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, Princeton.

Hutchinson, G.E. (1957) Concluding remarks: Population studies: Animal ecology and demography. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415-427.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, Burlington, Massachusetts, United States.

Nix, H.A. 1986. A biogeographic analysis of Australian Elapid Snakes. *Atlas of Elapid Snakes of Australia*. pp 4-15. Australian Government Publishing Service: Canberra.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**, 683-691.

Peterson, A.T. & Cohoon, K.P. (1999) Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling*, **117**, 159-164.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderseon, R.P., Martinez-Meyer, E., Nakamura, M. & Araújo, M.B. (2011) *Ecological niches and geographic distribution*. Princeton University Press, Princeton.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231-259.

Pottier, J, Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C.F., Vittoz, P., & Guisan, A. (2013) The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Global Ecology and Biogeography*, **22**, 52-63.

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349-361.

R Core Team. (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. URL http://R-project.org.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography,* **33**, 1689-1703.

Randin, C.F., Jaccard, H., Vittoz, P., Yoccoz, N.G. & Guisan, A. (2009a) Land use improves spatial predictions of mountain plant abundance but not presence-absence. *Journal of Vegetation Science*, **20**, 996-1008.

Randin, C.F., Vuissoz, G., Liston, G.E., Vittoz, P. & Guisan, A. (2009b) Introduction of snow and geomorphic disturbance variables into predictive models of alpine plant distribution in the Western Swiss Alps. *Arctic Antarctic and Alpine Research*, **41**, 347-361.

Soberón, J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, **10**, 1115-1123.

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

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

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

Tax, D.M.J. & Duin, R.P.W. (1999) Support vector domain description. *Pattern Recognition Letters* **20**, 1191-1199

Tax, D.M.J. (2012) Ddtools: The data description toolbox for Matlab. v. 1.9.1.

Thuiller, W., Lafourcade, B., Engler, R. & Araujo, M.B. (2009) BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369-373.

Thuiller W., Münkemüller T., Lavergne S., Mouillot D., Mouquet D., Schiffers K., & Gravel D. (2013) A road map for integrating eco-evolutionary processes into biodiversity models *Ecology Letters*, **16** (s1), 94-105.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. & Kadmon R. (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* **13**, 397-405.

van der Heijden, F., Duin, R.P.W., de Ridder, D. & Tax, D.M.J.. (2004) *Classification, Parameter Estimation, and State Estimation: An engineering approach using Matlab*. Wiley, West Sussex, England.

Williams, J.W., Jackson, S.T. & Kutzbach, J.E. (2007) Projected distributions of novel and disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences*, **104**, 5738-5742.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. & NCEAS Predicting Species Distributions Working Group. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763-773.

Table 1

	Training form		Training form	
	(Testing Set)		(Transfer Set)	
	One-class	Two-class	One-class	Two-class
Representation	(presence-only)	(presence/absence)	(presence-only)	(presence/absence)
Density-based representations				
Gaussian density	0.792(sd: 0.118)	0.765 (sd: 0.134)	0.629 (sd: 0.167)	0.605 sd(: 0.164)
Parzen density estimator	0.798 (sd: 0.108)	0.786 (sd: 0.115)	0.653 (sd: 0.175)	0.627 sd(: 0.160)
Principal components classifier	0.737 (sd: 0.131)		0.553 (sd: 0.180)	
Distance-based representations				
K-means clustering	0.789 (sd: 0.111)		0.658 (sd: 0.181)	
K-nearest neighbor	0.794 (sd: 0.107)	0.804 (sd: 0.112)	0.653 (sd: 0.174)	0.613 sd(: 0.178)

Nearest Neighbor	0.703 (sd: 0.127)	0.756 (sd: 0.137)	0.597 (sd: 0.163)	0.634 sd(: 0.195)
Concept learning representations				
Self-organizing map	0.709 (sd: 0.136)		0.653 (sd: 0.163)	
Auto-encoder	0.673 (sd: 0.148)		0.597 (sd: 0.156)	
Radial basis neural network		0.702 (sd: 0.171)		0.553 sd(: 0.172)
Naive Bayes classifier		0.763 (sd: 0.119)		0.652 sd(: 0.179)
Support vector machine	0.719 (sd: 0.161)	0.665 (sd: 0.162)	0.672 (sd: 0.180)	0.538 sd(: 0.162)

Table 1. Mean AUC on withheld data of 9 one-class and 7 two-class models. The different representations of the classification problem may be classified as density-based representations (derived from classical methods in probability and statistics), distance-based representations (emphasizing similarity to known examples), and concept learning representations (algorithmic representations from machine learning theory).

Figure 1.

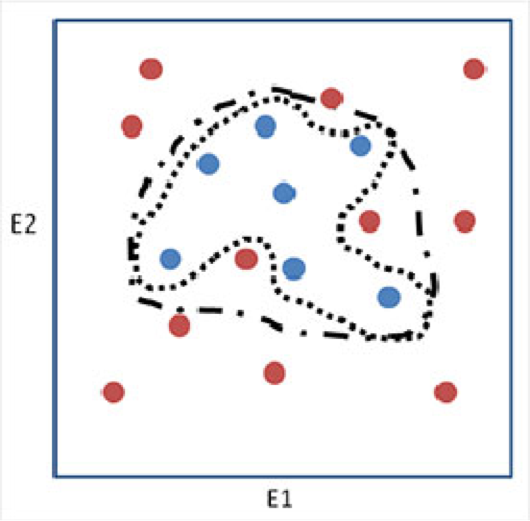

Figure 1. Conceptual models illustrating the difference between one-class and two-class methods for ecological niche modeling. Presence data are shown as blue dots in a coordinate space defined by two environmental variables, E_1 and E_2. Absence data are shown as red points. The dotted line separates the blue points from the red points in this space. Such a *decision boundary* is the typical fitted outcome of a presence-absence model. The dashed line, by contrast is the estimated *set boundary* containing all the blue points, without consideration for whether or not any red points are contained. This figure also shows why presence-absence models may sometimes yield predictions exclusive of suitable sites: all datasets will contain some points that may truly belong to the niche, but are scored as absences due to detectability, dispersal limitations, or other constraints.

Figure 2

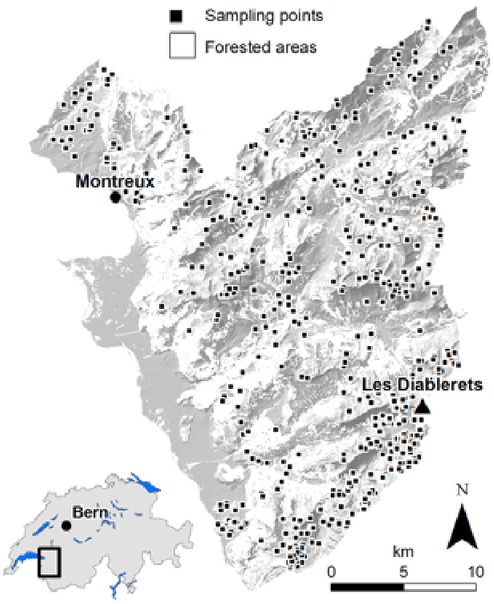

Figure 2. Locations of *n*=550 vegetation survey plots in the Swiss pre-alps.

Figure 3

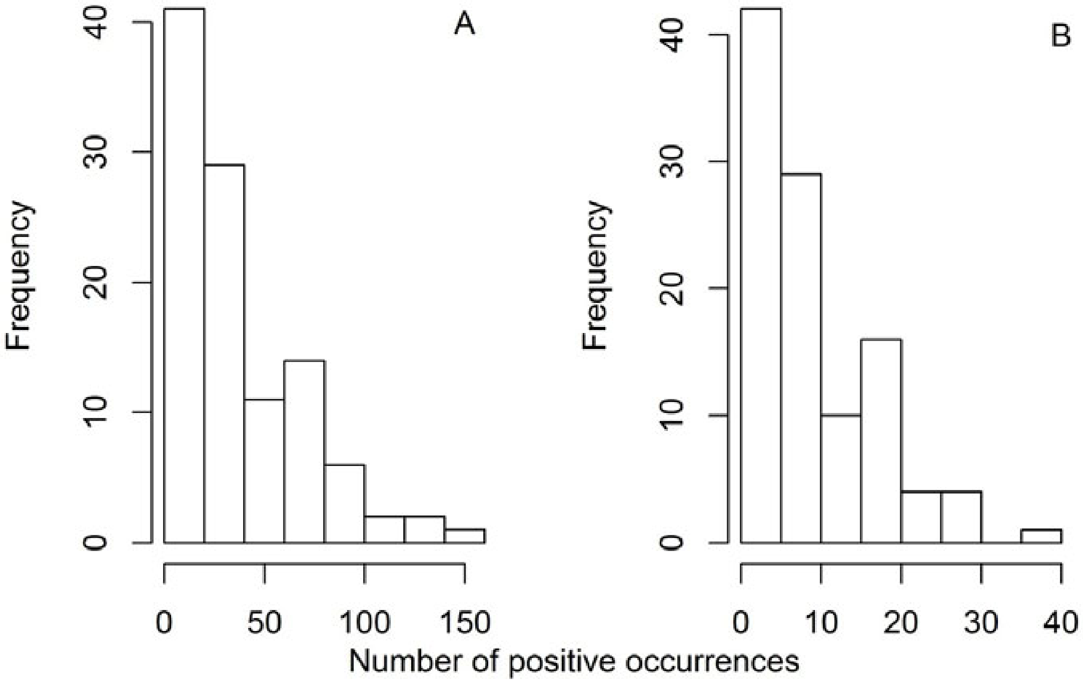

Figure 3. Histograms of the number of observations for species (out of *n*=440 plots in the training set, A, and *n*=110 in the test set, B). Both plots show that positive occurrence information was strongly skewed in each data set.

Figure 4

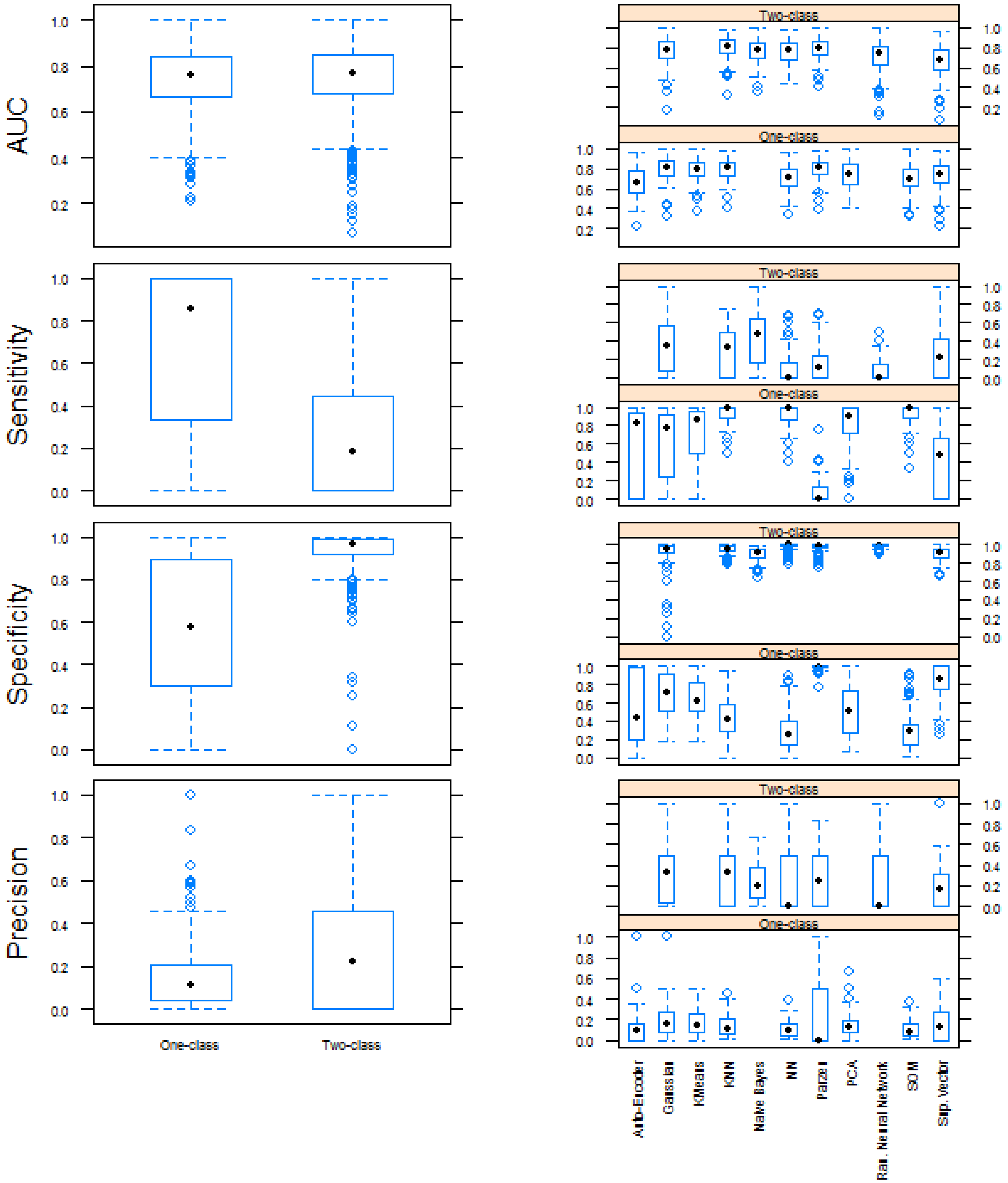
AUC
Sensitivity
Specificity
Precision
0.0
0.2
0.4
0.6
0.8
1.0
One-class
Two-class
Auto-Encoder
Gaussian
KMeans
KNN
Naive Bayes
NN
Parzen
PCA
Rad. Neural Network
SOM
Sup. Vector

Figure 4. Model performance measures varied considerably within and between representations. Box plots of untransformed performance are shown by representation (left column) and for one-class vs. two-class forms (right column). Preferred values of the statistic are those that approach one, whereas poor performance in AUC is designated by values of 0.5 and below and by values that approach zero in the other measures.

Figure

5

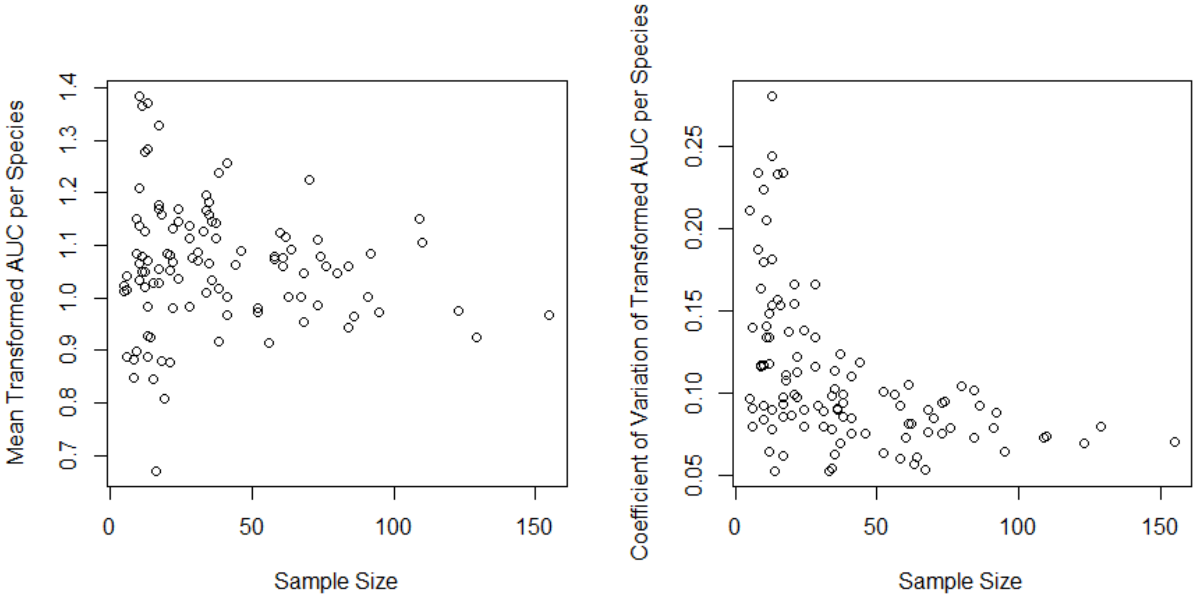

Figure 5. Model performance (transformed AUC; left panel) and consistency (coefficient of variation of transformed AUC; right panel) as a function of sample size. Frequently, models generated from samples of less than 29 occurrence records had AUC values below the random threshold (AUC=0.50, transformed AUC=0.79). Variation at larger sample size was not as extensive, but there were fewer species with such a larger pool of occurrences.