Journal of Biogeography (*J. Biogeogr.*) (2009) **36**, 2290–2299

ORIGINAL ARTICLE

Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models

Samuel D. Veloz*

Department of Environmental Science and Policy, University of California, One Shields Avenue, Davis, CA, USA

ABSTRACT

Aim Environmental niche models that utilize presence-only data have been increasingly employed to model species distributions and test ecological and evolutionary predictions. The ideal method for evaluating the accuracy of a niche model is to train a model with one dataset and then test model predictions against an independent dataset. However, a truly independent dataset is often not available, and instead random subsets of the total data are used for 'training' and 'testing' purposes. The goal of this study was to determine how spatially autocorrelated sampling affects measures of niche model accuracy when using subsets of a larger dataset for accuracy evaluation.

Location The distribution of *Centaurea maculosa* (spotted knapweed; Asteraceae) was modelled in six states in the western United States: California, Oregon, Washington, Idaho, Wyoming and Montana.

Methods Two types of niche modelling algorithms – the genetic algorithm for rule-set prediction (GARP) and maximum entropy modelling (as implemented with Maxent) – were used to model the potential distribution of *C. maculosa* across the region. The effect of spatially autocorrelated sampling was examined by applying a spatial filter to the presence-only data (to reduce autocorrelation) and then comparing predictions made using the spatial filter with those using a random subset of the data, equal in sample size to the filtered data.

Results The accuracy of predictions from both algorithms was sensitive to the spatial autocorrelation of sampling effort in the occurrence data. Spatial filtering led to lower values of the area under the receiver operating characteristic curve plot but higher similarity statistic (*I*) values when compared with predictions from models built with random subsets of the total data, meaning that spatial autocorrelation of sampling effort between training and test data led to inflated measures of accuracy.

Main conclusions The findings indicate that care should be taken when interpreting the results from presence-only niche models when training and test data have been randomly partitioned but occurrence data were non-randomly sampled (in a spatially autocorrelated manner). The higher accuracies obtained without the spatial filter are a result of spatial autocorrelation of sampling effort between training and test data inflating measures of prediction accuracy. If independently surveyed data for testing predictions are unavailable, then it may be necessary to explicitly account for the spatial autocorrelation of sampling effort between randomly partitioned training and test subsets when evaluating niche model predictions.

Keywords

Accuracy assessment, *Centaurea maculosa*, environmental niche model, GARP, invasive species, Maxent, similarity statistic *I*, spatial autocorrelation, spatially autocorrelated sampling, western United States.

*Correspondence: Samuel D. Veloz, Department of Environmental Science and Policy, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA.
E-mail: sdveloz@ucdavis.edu

www.blackwellpublishing.com/jbi
doi:10.1111/j.1365-2699.2009.02174.x

Journal of Biogeography

INTRODUCTION

Ecological niche models are increasingly being employed to test predictions from ecological and evolutionary theory (Peterson *et al.*, 1999; Anderson *et al.*, 2002; Graham *et al.*, 2004b), to test the invasion potential of species in novel areas (Peterson & Vieglais, 2001; Estrada-Peña *et al.*, 2007) and to predict the effects of climate change on species distributions (Thuiller, 2004). The models work by predicting a species' distribution given a set of georeferenced occurrences and a corresponding set of environmental layers, typically including climate and land cover derived from geographical information system sources. The models are based on a concept of the niche which assumes that species have a limited range of environmental conditions under which they can persist without immigration (Hutchinson, 1957). To estimate the niche space occupied by a species, these models use correlations between species occurrences and the environmental layers to predict their potential distributions across a given study extent.

Two general categories of niche models have been used extensively in recent years: those using presence and absence species occurrence data, and those that use presence-only data. The popularity of presence-only niche models has arisen in part due to the increased availability of presence-only records from museum databases and other non-systematic surveys (Graham *et al.*, 2004a). Despite the increasing use of presence-only methods in niche modelling, there has been a lack of critical evaluation of the conditions under which these models perform well. Models developed from these data may suffer from several forms of bias (Stockwell & Peterson, 2002), yet little work has been attempted to show how the bias in occurrence data affects model predictions or the accuracy assessment of those predictions.

Occurrence records may exhibit spatial bias when a species distribution has only been sampled in a non-random subset of the total area where the species occurs, such as only in easily accessible areas. Spatially biased sampling may result in poor model calibration, and this issue has received some recent attention in the literature (Phillips *et al.*, 2009). On the other hand, spatially autocorrelated sampling effort, in which each site sampled is much closer together than would be expected if all sites were visited randomly, may also lead to inaccuracy in model prediction and evaluation, even if there is no other bias. Little consideration has been paid to how spatially autocorrelated sampling effort may influence the assessment of niche model predictions.

When an independent source of occurrence data is unavailable for validating model predictions, the original data must be used. Simply using the same data to build the model and evaluate predictions (resubstitution) leads to overly optimistic views of the model's accuracy (Elith & Burgman, 2002). In order to provide a better assessment of the accuracy of niche model predictions, the occurrence data are often partitioned into 'training' and 'test' subsets (Fielding & Bell, 1997). Niche models are developed using the training data and validated using the test data. The data are usually split into each group randomly. If the data were collected with a spatially autocorrelated sampling effort, the test data would not be independent of training occurrences. Thus, using a random partition of occurrence records collected with spatially autocorrelated sampling effort for the assessment of model performance may generate an overly optimistic assessment of the accuracy of the model. Frequently used test statistics assume independence of samples, and the assessment of niche model performance may therefore exhibit a similar inflation of model significance or performance if training and test data are not independent (Araújo & Guisan, 2006).

This study tests the accuracy of two presence-only environmental niche model algorithms: the genetic algorithm for rule-set prediction (GARP; Stockwell & Peters, 1999) and maximum entropy modelling (Maxent; Phillips *et al.*, 2006), for predicting the distribution of an exotic plant in the western United States. In several recent studies, Maxent niche model predictions have been found to have higher predictive accuracy on average than GARP niche models (Elith *et al.*, 2006; Hernandez *et al.*, 2006; Phillips *et al.*, 2006), but GARP has a longer history of application in the field. Spatial bias in the occurrence data is predicted to affect the evaluation of niche model predictions if test data are not spatially independent of training data. The effects of spatial bias on niche model evaluation are examined by repeating each analysis with a spatial filter applied to the training data. In this paper I test the hypotheses that: (1) the spatial autocorrelation of sampling effort between training and test data will bias the assessment of niche model predictions using standard methods of accuracy assessment, and (2) the random partition of occurrence data, collected with spatially autocorrelated sampling effort, into training and test data will result in an overly optimistic assessment of model accuracy.

MATERIALS AND METHODS

Study species

Centaurea maculosa Lam. (spotted knapweed; Asteraceae) is a biennial or short-lived perennial plant native to Eurasia that has become a serious pest in rangelands throughout the western United States and Canada. *Centaurea maculosa* can form dense stands on disturbed soils, especially along roads and railways and on grazed rangelands (Watson & Renney, 1974). The species was initially chosen for distribution modelling in collaboration with the California Department of Food and Agriculture's Noxious Weed Program because the species was one of the plants on the state's most noxious weed list that had possibly not yet reached its maximum distribution (S. Schoenig, Noxious Weed Information Project, California Department of Food and Agriculture, personal communication). Environmental niche models assume that the distribution of a species is at equilibrium with the environment in which it occurs and thus the inclusion of areas where the weed

has not fully invaded suitable environmental conditions within the study extent may have clouded the results of this study. However, because *C. maculosa* has successfully invaded a large portion of the western United States and seems to have reached the limits of its distribution within certain subregions, the species represents a suitable candidate for testing the ability of environmental niche models to model the distribution of a species at various stages during the invasion process.

Models were only trained with occurrences where the species is invasive. This may allow for a further exploration of the potential climate space which is suitable for the persistence of *C. maculosa* populations, as the realized distribution of the species in its native range in Eurasia may be more severely limited by competitors than its realized niche within its invaded range in the United States (Broennimann *et al.*, 2007).

Niche modelling

Two environmental niche models, GARP, as implemented in openModeller version 0.6.1 (available for download from http://openmodeller.sourceforge.net/; accessed 18 May 2009) and Maxent version 3.2.2 (available for download from http://www.cs.princeton.edu/~schapire/maxent/; accessed 18 May 2009), were used to predict the spatial distribution of spotted knapweed across six states in the western United States. I used the new openModeller implementation of GARP because it outperformed the older desktop version in a recent comparative analysis (Elith *et al.*, 2006) and it was easier to implement batch runs using the command line interface.

For this study, I used 100 runs for each GARP experiment using a 0.01 convergence limit with all rule types. I used the best subset feature (Anderson *et al.*, 2003) to create an optimal set of runs for each experiment. The 40 runs with the lowest omission rate were initially selected. From the 40 models with the lowest omission rate, the area predicted to be occupied in each model was compared with the median area predicted to be occupied across all 40 models. The 20 models for which the area predicted to be occupied was closest to the median area predicted to be occupied were selected as the final best subset of models. The final output from the analysis was a map with cell values equal to the number of models predicting the cell as occupied divided by 20 (which equals the total number of models in the best subset). For each Maxent experiment, the default values for the algorithm were chosen. I used the logistic threshold output format which results in each grid cell in the map having values ranging continuously from 0 to 1 and can be interpreted as the probability of presence of suitable environmental conditions for the target species.

Study region

Niche models were created for *C. maculosa* within California, Oregon, Washington, Idaho, Montana and Wyoming. Analyses were run over all possible combinations of the groups of states (62 different study extents) to examine the effects of including a wider range of environmental conditions than those present in the sampled area. The six states were chosen based on the availability of data, the variability in prevalence of *C. maculosa* within each state and the environmental variability between states within the region. The variability in prevalence of *C. maculosa* allows for comparisons of the ability of each algorithm to predict the potential distribution of *C. maculosa* over areas where the weed has already become well established and also areas where it has yet to, or cannot, invade.

Presence data were obtained from the California Department of Food and Agriculture Noxious Weed Information Project (http://www.cdfa.ca.gov/phpps/ipc/noxweedinfo/noxweedinfo_hp.htm; accessed 28 May 2009) and the Invaders database (for states other than California: http://invader.dbs.umt.edu/; accessed 18 May 2009). All presence locations were projected into geographical coordinates (latitude, longitude) to correspond to the coordinate system of environmental layers. Points were initially spatially filtered such that only one point occurred within each 0.0083 decimal degree^2 (*c.* 1 km^2) grid cell of the environmental layers.

Presence locations within each state combination were divided into training and test data for model development and evaluation. The test data were created by randomly partitioning the points within each study combination by using 80% for training and 20% for testing (see Appendix S1 in Supporting Information). The test sets generated in this step were used for evaluation in all subsequent analyses.

Environmental layers

The environmental layers used in constructing models of the environmental niche of *C. maculosa* were obtained from WorldClim (http://www.worldclim.org/; accessed 28 May 2009) (Hijmans *et al.*, 2005). Niche models were created within each state combination with a set of four environmental layers: mean diurnal temperature range (mean of monthly maximum temperature minus minimum temperature), mean temperature of the wettest quarter, mean temperature of the coldest quarter and precipitation of the driest quarter. Potential predictor variables available from the WorldClim site were screened for collinearity by examining pairwise correlations between variables. The included variables all had Pearson product–moment correlation coefficients of ≤ 0.5.

Spatial filters for training data

The results from the initial model runs were used to assess the level of spatial structure within the occurrence data. At each occurrence site and within each study extent, the niche model prediction from each algorithm was compared with the average niche model prediction from all models for each algorithm. The difference between the average prediction and the individual prediction was treated as the model residual and

was used to model the spatial dependence in the prediction errors. Occurrence records were projected into USA Contiguous Albers Equal Area Conic USGS projection using ArcGIS 9.3 so that distances between points would be recorded in metres. I used the sgeostat package in R statistical software version 8.0 (R Development Core Team, 2008) to calculate semivariograms to assess the distance at which points are spatially independent based on niche model errors. Semivariograms model the variance between pairs of points as a function of distance. The distance at which the semivariance reaches a sill is called the 'range', and is the distance beyond which the semivariance between pairs of points is spatially independent.

An *a priori* spatial scale was selected to model the spatial structure of the results from the first set of niche models. Although spatial patterns may often be detected at multiple spatial scales, the intent of this analysis was to determine if spatial structure at relatively small spatial scales influences the evaluation of niche model predictions. For this reason, semivariograms were modelled with a maximum distance of 30–50 km. In all cases an isotropic spherical model was used for semivariogram modelling. A 1-km lag distance was chosen as this is the minimum sampling unit based upon the grid size of the environmental data.

The mean range for the residuals for all niche models was used to spatially filter the occurrence data. Use of the mean range was justified as a simple way to filter all the datasets in an equal manner because the actual independence of pairs of occurrence points was not necessary. Rather, the occurrence datasets were filtered so that the relative influence of the spatial structure within all datasets could be detected. The mean range was used to spatially filter points within each training data set so that training points are separated by a minimum distance but training and test points are not necessarily separated.

Because the spatial filtering of the data drastically reduced the sample size of the occurrence data (see Appendix S1), non-spatially filtered subsets of the training presence points were randomly selected so that the number of points in each study extent equalled the spatially filtered data sets. Spatially filtered and non-spatially filtered random subsets were created for each state and group of states. This was done to correct for any effect of reduction in sample size.

I used a bivariate spatial point pattern analysis, Ripley's *K*, to quantify the spatial relationships between training and test points (Andersen, 1992). This statistic estimates the expected number of points of one class centred around a point in another class at a given distance. The distance used was the mean range of the semivariogram analysis. Larger values of *K* indicate higher spatial clustering of the points from the two classes. Here I compare the spatial patterns of training points in relation to test points using the spatstat package (Baddeley & Turner, 2005) in R 2.8.0 (R Development Core Team, 2008). I used a translation edge corrected estimate of *K*, which yields a symmetric result when either the training or test points are used as the focal class.

Evaluating model accuracy

The area under the curve of a receiver operating characteristic (ROC) plot (AUC) was used as a threshold independent measure of accuracy (Swets, 1988). Although traditionally used with presence/absence data, the AUC has increasingly been advocated for use in the evaluation of environmental niche models (Fielding, 2002). AUC values range between 0 and 1, with maximum accuracy achieved with values of 1, accuracy no better than random with values of 0.5 and values < 0.5 indicating performance worse than random. When used with presence-only data, the maximum values should theoretically be < 1 (Phillips *et al.*, 2006). Because the calculation of the AUC requires absence locations, 10,000 points from the background were generated for each analysis extent. The results of each niche model were recorded at each presence/background location and used to calculate the AUC. ROC curves and AUC values were created using the ROCR package (Sing *et al.*, 2005) in R 2.8.0 (R Development Core Team, 2008).

It could be argued that spatially filtering the occurrence data could lead to the species being predicted as more of a generalist by the niche models and thus result in lower AUC values because generalist species are not as well discriminated from the background as specialists (whose occurrences would probably be more spatially and environmentally clustered). To compensate for this potentially confounding factor, I compared the overlap in predictions from each spatially filtered and random subset niche model with a model created with all occurrence records and across all the sites. Recent work has shown that using all the data to create a model will tend to make a more accurate prediction, especially in cases with larger environmental heterogeneity across the study extent (Broennimann & Guisan, 2008).

As an alternative to the AUC, I also used the similarity statistic (*I*), an index of niche overlap, to assess prediction accuracy. *I* was estimated for each niche model by using ENMtools (Warren *et al.*, 2008, available for download from http://enmtools.blogspot.com/; accessed 18 May 2009). As used here, niche overlap can be thought of as the level of agreement between the predictions from two niche models. *I* is calculated by comparing the difference in predicted values at each cell across each study extent. *I* ranges from 0 to 1, with 0 indicating predictions with zero overlap and 1 indicating predictions with perfect overlap. Because all cells are used in the calculation of *I*, there is no influence of the spatial configuration between training and test points. If spatially clustered occurrences truly reflect the environmental conditions which limit the occurrence of the species, then *I* values for the spatially filtered treatments should be lower than *I* values associated with the random subset treatment. Alternatively, if spatial filtering results in a less biased sample of the environmental conditions that limit the species occurrence due to the reduction of the effect of spatially autocorrelated sampling, then the spatially filtered treatments should have equal or higher *I* values than the randomly chosen subset treatment.

Here I make the assumption that the model created with all the occurrence points (across all states) is the most accurate model. Under this assumption, model predictions generated from the reduced occurrence data (spatially filtered or random subset) and study extents (different state combinations), which overlap more with predictions from the full occurrence model, are more accurate based on the similarity statistic (*I*). For example, if a model from California built using a spatially filtered training dataset has higher overlap with the full model than with a model built with a random subset of training data from California, the model built with spatially filtered data would be interpreted as having higher accuracy. If the spatial clusters of occurrence data did, in fact, reflect the true distribution of the species, then the predictions from models built with spatially filtered data should have lower overlap with predictions from the full model (i.e. lower accuracy) because the spatially filtered training data would be a poor sample of the species' true distribution. However, if the sampling clusters were artefacts of spatially autocorrelated surveys, then the spatial stratification of the filtered training data used to build the models would result in greater prediction overlap with the full model (i.e. higher accuracy) than would models from the random subset treatment.

I used analysis of variance (ANOVA) to compare differences in AUC values from the different algorithm and sample size treatments. Factors that were included were: location (62 study extents), algorithm type (two levels: Maxent and GARP) and sample size treatment (two levels: random subset and spatial filter). Locations (state or state combination) were treated as blocks in the models to account for the variation introduced by the different study extents and sampling bias within each state. Interactions for algorithm type × sample size treatment, block × algorithm, and block × spatial treatment were also included in the model. All statistical analyses were conducted using R software, version 2.8.0 (R Development Core Team, 2008).

RESULTS

Spatial analysis

The mean range for all niche models was 38.6 km. This distance was used as the minimum distance allowed between training points for the spatially filtered occurrence dataset for spotted knapweed. Spatial filtering of the training data using this minimum distance resulted in extreme reductions in training sample size (Appendix S1).

Training points were less spatially clustered around test points in the spatially filtered data set compared with the randomly reduced subset treatment based on a Wilcoxon signed rank test of *K* values ($V = 9$, $P < 0.001$, $n = 61$). A linear regression of AUC by *K* revealed a significantly positive correlation for both GARP (Fig. 1a, slope ± SE = 0.01 ± 0.002, $P < 0.001$) and Maxent (Fig. 1b, slope ± SE = 0.02 ± 0.003, $P < 0.001$) models, meaning that models with training data that were more spatially clustered around test points also had higher AUC values.

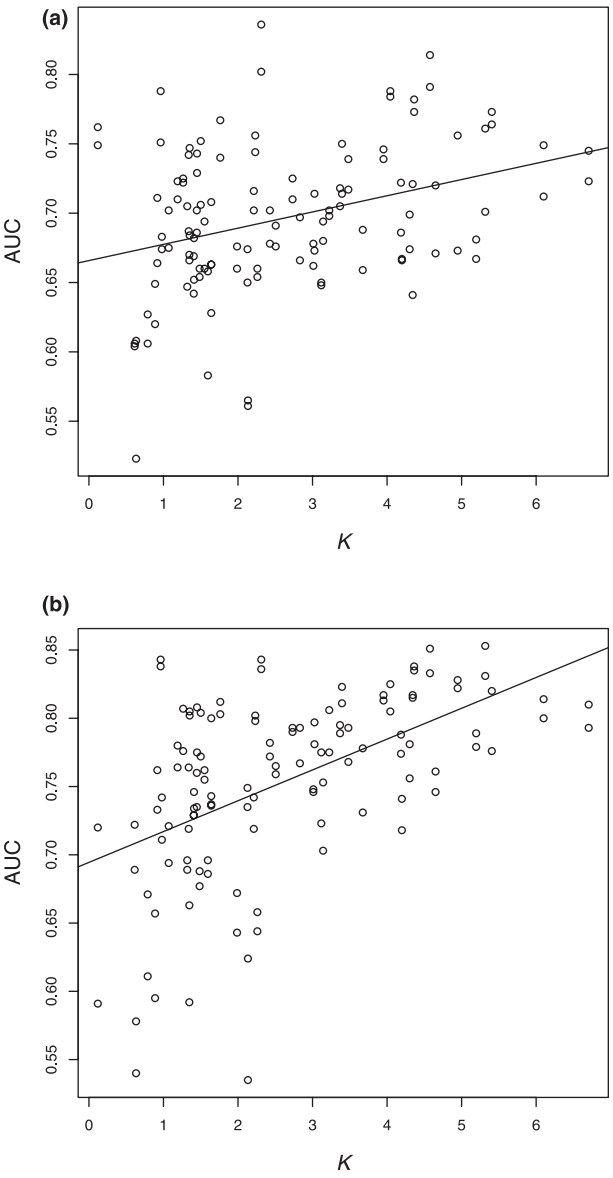


Figure 1 Linear regression of the area under the receiver operating characteristic curve plot (AUC) versus the spatial clustering of training points around test points (Ripley's *K*) for (a) GARP models and (b) Maxent models. High values for AUC indicate better discrimination of test occurrences from the background while high values of *K* indicate more spatial clustering of test and training points. The slope for both algorithms is significantly different from 0: (a) slope ± SE = 0.01 ± 0.002, $P < 0.001$, $n = 122$; (b) slope ± SE = 0.02 ± 0.003, $P < 0.001$, $n = 122$. *K* explains less of the variation in AUC for GARP models ($R^2 = 0.11$) than Maxent models ($R^2 = 0.27$).

Niche model comparisons

The predictions from models from Oregon by itself were omitted due to extremely small sample sizes. Including these locations in statistical models resulted in outlier points which led to violation of statistical model assumptions. In addition, the models from Idaho by itself were omitted because the

Table 1 (a) ANOVA comparison between the area under the receiver operating characteristic curve plot (AUC) values of niche model predictions. (b) ANOVA comparison between similarity statistic (*I*) values for predicted niche overlap for each treatment and a model built with all available data. Block refers to the spatial extent of sampling. Algorithm type refers to GARP or Maxent. Spatial treatments included both the reduced sample size occurrence data set and spatially filtered occurrence data set.

Effect	d.f.	*F*-value	*P*
(a)*			
Block	59	32.2553	< 0.001
Algorithm	1	574.667	< 0.001
Spatial treatment	1	19.4592	< 0.001
Block × algorithm	59	3.2907	< 0.001
Block × spatial treatment	59	1.4972	0.062
Algorithm × spatial treatment	1	0.8239	0.368
Residuals	59		
(b)†			
Block	59	6.4818	< 0.001
Algorithm	1	705.221	< 0.001
Spatial treatment	1	22.5381	< 0.001
Block × algorithm	59	3.6944	< 0.001
Block × spatial treatment	59	1.615	0.034
Algorithm × spatial treatment	1	1.8654	0.177
Residuals	59		

*Multiple R^2: 0.9792, adjusted R^2: 0.9158.
†Multiple R^2: 0.9603, adjusted R^2: 0.839.

random subset treatment had greater separation between training points than the spatially filtered treatment and inclusion of this block led to outliers in the statistical tests.

Significant differences were detected in AUC scores for the different algorithms and the spatial treatment (Table 1a). AUC scores for Maxent predictions were significantly higher than those for GARP predictions, although this result varied by block. Visual comparison of the marginal means from the block × algorithm interaction indicate that Maxent does tend to outperform GARP or is no different from GARP across the blocks in this study (see Appendix S2). The AUC for the random subset treatment was significantly higher than for the spatially filtered treatment. There was no significant interaction detected between the block and spatial treatment, indicating that the spatial treatment had a consistent effect across blocks. No significant interaction was detected between algorithm type and spatial treatment, indicating that the spatial treatments affected the two algorithms in the same way (Fig. 2a).

In contrast to the AUC analysis, models built with the spatially filtered occurrences had significantly higher niche overlap with the model parameterized with all occurrences than did those built with the random subset treatment (Table 1b). No significant interaction was detected between algorithm and spatial treatment, signifying that the spatial filter treatment affected results from the two algorithms similarly (Fig. 2b). There was a significant block × spatial treatment

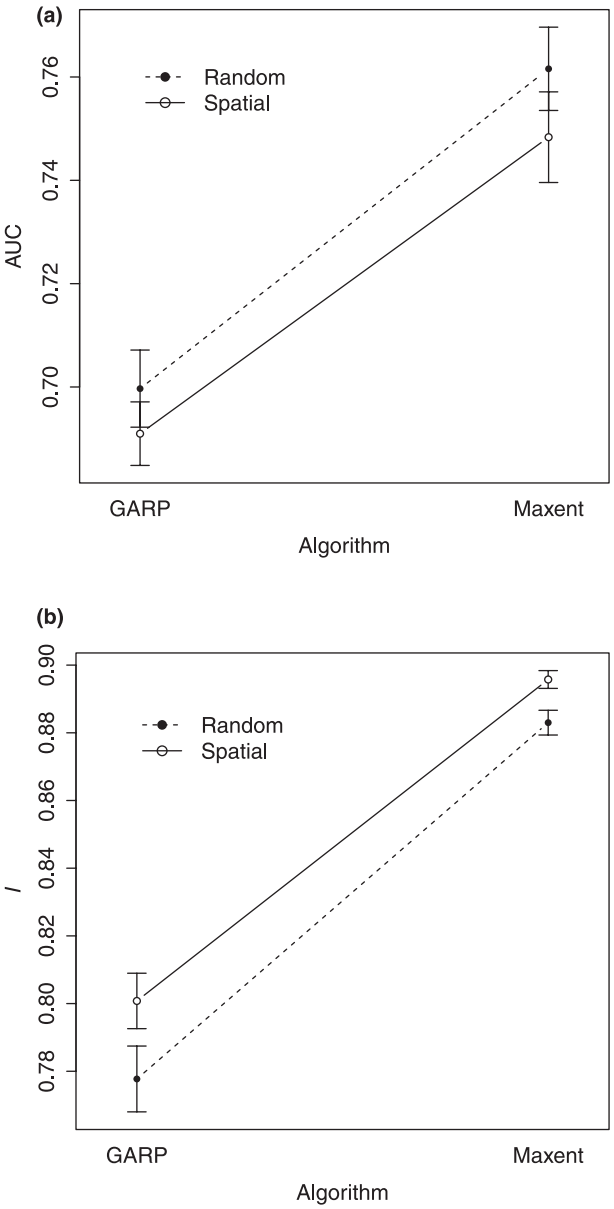
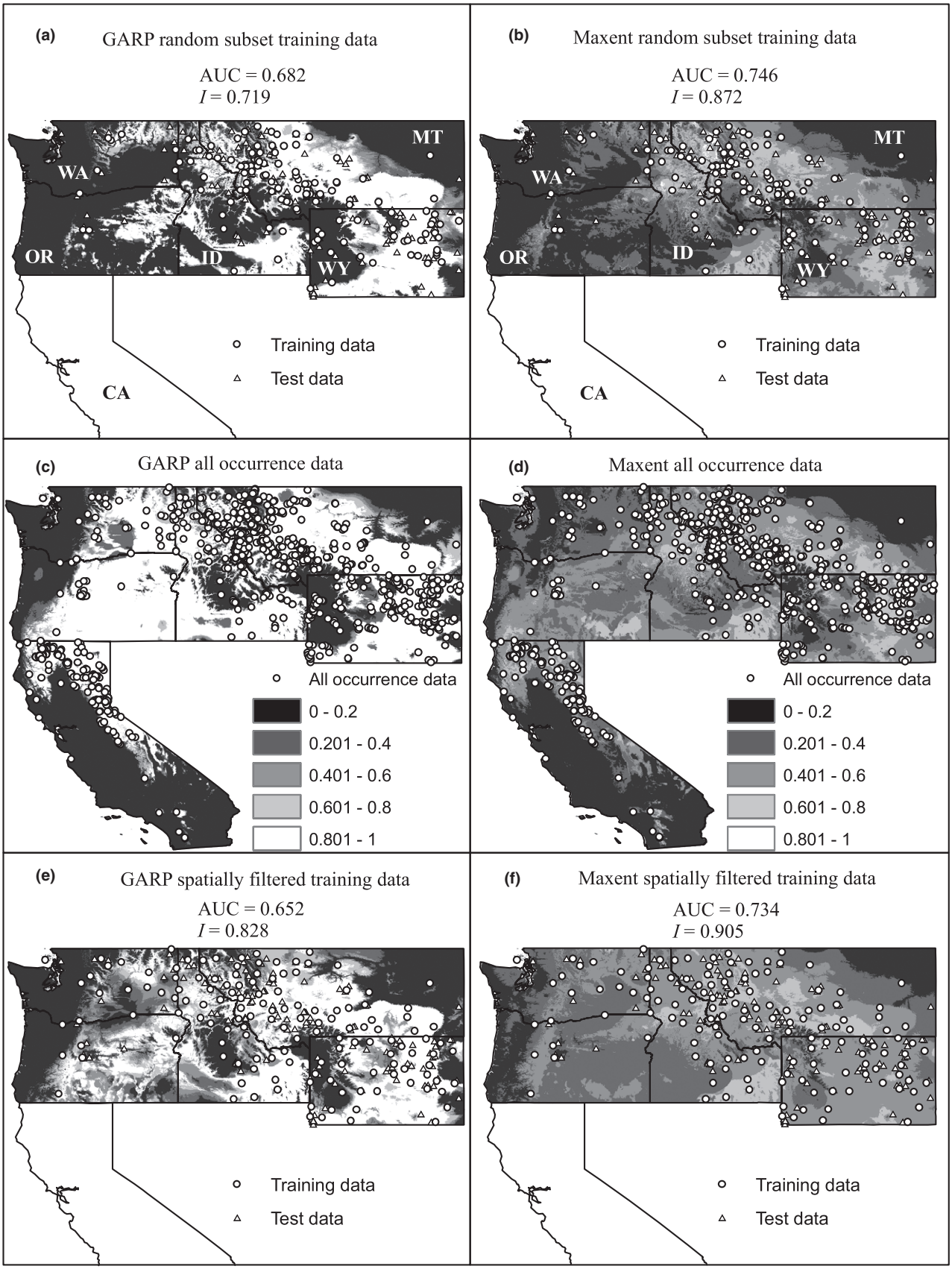


Figure 2 Plots for ANOVA results comparing (a) area under the receiver operating characteristic curve plot (AUC) values and (b) similarity statistic (*I*) values to the algorithm (GARP or Maxent) by spatial treatment (random subset versus spatially filtered training data) interaction. In both cases, no significant interaction was detected. (a) Random subset models have significantly higher AUC scores than spatially filtered models. (b) Spatially filtered models have significantly higher *I* scores than randomly chosen subset models. Error bars indicate ± 1 SE.

interaction, indicating that effect of the spatial treatment varied by block. The *I* scores were higher for Maxent predictions than GARP predictions, but there was a significant block × algorithm interaction. In a majority of the cases, models trained with the spatially filtered data had higher *I* scores across the blocks than the random subset training data (Appendix S2). Regions with a low density of occurrence points are expected to be less sensitive to the spatial filter



treatment. The largest departure from the average response in the block × spatial treatment interaction occurred in the block containing Oregon, Washington and Idaho, the three states with the lowest density of occurrence points (Appendix S2). Across all blocks, visual inspection of the marginal means from the block × algorithm interaction shows that Maxent models always had higher *I* scores than GARP models (Appendix S2).

Maps of the predictions from this analysis reveal clear differences between both the spatial filter treatment and the two algorithms (Fig. 3). GARP models primarily predict areas with either very high or very low values while Maxent models predict much more of a gradient in habitat suitability. In Fig. 3, for both GARP and Maxent, the random subset treatments fail to predict areas in Oregon, Washington and south-west Idaho (Fig. 3a,b) that are predicted to be moderately to highly suitable by both the full models (Fig. 3c,d) and the spatially filtered models (Fig. 3e,f), resulting in lower *I* values compared with the spatially filtered treatments. However, because the training occurrences in the random subset treatment are much closer to the test points, AUC values are higher for the random subset models.

DISCUSSION

The practice of creating subsets of data for model training and testing purposes is common in ecological modelling when independent data for model evaluation are unavailable. It is assumed that a random subdivision of the sample data constitutes a reasonable substitute for an independent dataset with which to evaluate the model. This study illustrates how occurrence data collected with a spatially autocorrelated sampling effort can lead to lack of independence between the training and test data sets. This lack of independence then leads to a bias in model predictions and frequently an inflated assessment of the model's accuracy using AUC tests.

The results from this study show that the commonly used AUC statistic is very sensitive to the spatial autocorrelation between training and test points. Considerable variation in the AUC can be explained purely by the spatial clustering of training points around test points (Fig. 1). Although not tested here, other commonly used metrics to evaluate the accuracy of niche models, such as omission error and the binomial χ^2 (Peterson *et al.*, 1999; Anderson, 2003) test, are also likely to have a similar sensitivity to the spatial autocorrelation of sampling effort between training and test data. Specifically, the binomial χ^2 assumes that samples are random and independent of one another, which I have shown to be violated in the presence of spatially autocorrelated sampling effort.

The similarity statistic (*I*) employed here benefits from having each cell across the study extent contribute to its calculation such that no test points are needed and the spatial configuration of predictions across the study extent are explicitly incorporated. However, the use of *I* as an accuracy measure may be limited due to the need for a probability surface that approximates the truth to which predictions from the niche model are compared. The statistic is thus only able to estimate the relative accuracy of a given model with respect to some other estimation of the true distribution. When occurrence data are limited it may not be feasible to compare models built with a subset of the data with models built with a full dataset as was done here. However, randomization tests could be employed to test how well predictions from a niche model built from a subset of an occurrence dataset overlap with predictions from a niche model built with the full dataset, compared with predictions from niche models built from random points drawn from across the study extent, to see if predictions are significantly better than random (Raes & ter Steege, 2007; Warren *et al.*, 2008).

The results for both AUC and *I* are consistent across the two algorithms tested here and are likely to be robust to other commonly used algorithms. Consistent with other comparison studies, Maxent predictions were more accurate than those of GARP models, although there was some variation by block. The spatial filter treatment affected both algorithms in the same manner and thus I expect that other algorithms would respond in a similar manner, although this would need to be confirmed by further tests. Explicit methods for incorporating the spatial structure of occurrences have been developed for algorithms designed for presence/absence data (Dormann *et al.*, 2007) but little work has focused on incorporating these methods directly into presence-only algorithms. Recent work has shown that the influence of spatially biased occurrences can be reduced by comparing the occurrences with background points which also reflect the same spatial bias (Phillips *et al.*, 2009). Although this methodology will reduce the bias in the spatial predictions, it will not resolve the lack of independence between training and test data. To ensure the least biased prediction when using spatially biased occurrence records, background samples should also be chosen to reflect the spatial bias, while to ensure accurate prediction assessment some effort must be made to create truly independent sets of training and testing samples.

In this study it was evident that the spatial structure in the training data biased predictions because *I* values were lower for

Figure 3 Maps of predicted distributions of spotted knapweed (*Centaurea maculosa*) across Oregon (OR), Washington (WA), Idaho (ID), Montana (MT) and Wyoming (WY) in the western United States, using the GARP (a,c and e) and Maxent (b,d and f) algorithms. All maps have the same colour scale and show increasing climatic suitability from 0 to 1 indicated by the legends in (c) and (d). The area under the receiver operating characteristic curve plot (AUC) compares how well the models discriminate test points from randomly generated background points. The similarity statistic (*I*) shows how well predictions from the random subset (a,b) or spatially filtered treatments (e,f) overlap with models created using all the occurrence data (c,d). Data from California (CA) were used to train the full models (c,d) but not for the calculation of *I*.

randomly chosen subset models that retained the spatial clusters of occurrences. The clusters of occurrence records in this experiment did not truly reflect areas of higher suitability of *C. maculosa*, as clusters do not consistently sample from a similar range of environmental conditions across the total study extent. Because clusters sample from different subsets of environmental conditions, the effect of these clusters is moderated in the full model calibrated with the entire occurrence dataset. However, models calibrated with data retaining spatial clusters from the subregions examined here are over-fitted to the environmental conditions sampled by the local clusters and have lower prediction overlap with the full model.

Whenever possible, an independent evaluation dataset should be used to evaluate presence-only ecological niche models (Elith *et al.*, 2006). Independence between training and test datasets can best be achieved by ensuring that sampling effort used to collect test data is independent of the effort that was used to collect the training samples. When the sampling effort used to collect the training data is unknown and spatial autocorrelation of sampling effort is suspected, the results from this study suggest that ensuring that clusters of training data are not excessively clustered around test data will provide a better assessment of prediction accuracy.

ACKNOWLEDGEMENTS

The UC Davis Biological Invasions IGERT, NSF-DGE no. 0114432 provided initial funding for this project. I would like to thank Dan Warren, Robert Anderson, Deborah Elliott-Fisk, Jim Quinn, Susan Ustin, Sanne Sokolow, Elizabeth Chamberlin and two anonymous referees for insightful comments on early drafts.

REFERENCES

Andersen, M. (1992) Spatial analysis of two-species interactions. *Oecologia*, **91,** 134–140.

Anderson, R.P. (2003) Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography*, **30,** 591–605.

Anderson, R.P., Peterson, A.T. & Gómez-Laverde, M. (2002) Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos*, **98,** 3–16.

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.

Araújo, M.B. & Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33,** 1677–1688.

Baddeley, A. & Turner, R. (2005) spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **6,** 1–42.

Broennimann, O. & Guisan, A. (2008) Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters*, **4,** 585–589.

Broennimann, O., Treier, U.A., Müller-Schärer, H., Thuiller, W., Peterson, A.T. & Guisan, A. (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10,** 701–709.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M. & Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30,** 609–628.

Elith, J. & Burgman, M. (2002) Predictions and their validation: rare plants in the central highlands, Victoria, Australia. *Predicting species occurrences: issues of accuracy and scale* (ed. by J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall and F.B. Samson), pp. 303–313. Island Press, Washington, DC.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Lucia, J.L., Lohmann, G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.McC., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29,** 129–155.

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

Fielding, A.H. (2002) What are the appropriate characteristics of an accuracy measure? *Predicting species occurrences: issues of accuracy and scale* (ed. by J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall and F.B. Samson), pp. 271–280. Island Press, Washington, DC.

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., Ferrier, S., Huettmann, F., Moritz, C. & Peterson, A.T. (2004a) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19,** 498–503.

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

Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29,** 773–785.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25,** 1965–1978.

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

Peterson, A.T. & Vieglais, D.A. (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *BioScience*, **51,** 363–371.

Peterson, A.T., Soberón, J. & Sánchez-Cordero, V. (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285,** 1265–1267.

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

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19,** 181–197.

R Development Core Team (2008) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/.

Raes, N. & ter Steege, H. (2007) A null-model for significance testing of presence-only species distribution models. *Ecography*, **30,** 231–259.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21,** 3940–3941.

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

Stockwell, D.R.B. & Peterson, A.T. (2002) Controlling bias in biodiversity data. *Predicting species occurrences: issues of accuracy and scale* (ed. by J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall and F.B. Samson), pp. 537–546. Island Press, Washington, DC.

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

Thuiller, W. (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10,** 2020–2027.

Warren, D.L., Glor, R.E. & Turelli, M. (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution*, **62,** 2868–2883.

Watson, A.K. & Renney, A.J. (1974) The biology of Canadian weeds: 6. *Centaurea diffusa* and *C. maculosa*. *Canadian Journal of Plant Science*, **54,** 687–701.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Sample size for each block in the analysis from models trained with 80% of the occurrence data and for models with reduced sample size due to the spatial filter treatment.

Appendix S2 Analysis of significant block × algorithm and block × spatial treatment interactions for analysis of variance results comparing the area under the receiver operating characteristic curve plot (AUC) and similarity statistic (*I*) values.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

BIOSKETCH

Samuel D. Veloz is currently a post-doctoral researcher at the University of California, Davis, in the Department of Environmental Science and Policy. He is interested in exploring the various ways in which ecological niche modelling can be applied to test hypotheses in evolution and ecology as well as in predicting the impacts of invasive species and the effects of climate change. Specifically he is interested in exploring how the methodologies commonly applied in ecological niche modelling can be made more rigorous.

Editor: Richard Pearson