Ecological Informatics 36 (2016) 126–134

Contents lists available at ScienceDirect

Ecological Informatics

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

Field validation of an invasive species Maxent model

Amanda M. West [a,*], Sunil Kumar [a], Cynthia S Brown [b], Thomas J. Stohlgren [a], Jim Bromberg [c]

[a] Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA
[b] Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO 80523, USA
[c] Rocky Mountain National Park, U.S. National Park Service, Estes Park, CO 80517, USA

ARTICLE INFO

Article history:
Received 16 April 2016
Received in revised form 27 October 2016
Accepted 4 November 2016
Available online 5 November 2016

Keywords:
Biological invasions
Field validation
GLM
Maxent
Habitat suitability
Regression analysis
Model comparison
Bromus tectorum

ABSTRACT

Accurate and reliable predictions of invasive species distributions are urgently needed by land managers for developing management plans and monitoring new potential areas of establishment. Presence-only species distribution models are commonly used in these evaluations, however they are rarely tested with independent data over time or compared with presence-absence models fit with the same presence data. Using Maxent, we developed a presence-only model of invasive cheatgrass (*Bromus tectorum* L.) distribution in Rocky Mountain National Park, Colorado, USA in 2007 fit with limited data, and then tested the model with independent presence and absence data collected between 2008 and 2013. This model was verified using threshold dependent and threshold independent evaluation metrics. Next, we developed a Maxent model with cheatgrass presence data from 2007 through 2013 (i.e. Maxent 2013), and compared this model to a presence-absence method (i.e., generalized linear model; GLM 2013) using the same data. Threshold dependent and threshold independent evaluation metrics suggested Maxent 2013 outperformed GLM 2013, and a two-tailed Wilcoxon signed rank test indicated relative probability outputs were not significantly different between the models in geographic space. Based on known presences and absences of cheatgrass collected in the field, the Maxent 2013 and GLM 2013 relative probability outputs were highly correlated at absence locations but less correlated at presence locations. A Kappa comparison of Maxent 2007 and Maxent 2013 binary output provides evidence that Maxent is robust when fit with limited data. Our results indicate Maxent is an appropriate model for use when land management objectives are supported by limited resources and thus require a conservative, but highly accurate estimate of habitat suitability for invasive species on the landscape.

1. Introduction

The importance of predicting species distributions is increasing rapidly with global changes and their influences on native ecosystems. Scientists or land managers may need to locate and protect populations of a rare species or identify habitat that may be threatened by an invasive species, to name two of many reasons for the need of accurate predictive tools. Distributions of species vary according to an array of biological and physical conditions underlying the fundamental niche (Hutchinson, 1957), and correlative species distribution models (SDMs) provide a tool that enhances our understanding of this niche in geographic space. The maximum entropy model (Maxent; Phillips et al., 2006) is one of the most widely used presence-only SDMs; as of 04/15/2016, searching for '"maxent" and "species distribution"' in Web of Science yields 1292 results. This approach has demonstrated comparable ability to predict a species' range to models that use both locations where the species is known to occur and known not to occur (i.e., presence-absence models; Elith et al., 2006). Presence-only models use background points rather than true absences, and do not assume that absence precludes the possibility of occurrence (Evangelista et al., 2008; Kumar et al., 2009). Much uncertainty exists with absences, since they may indicate either unsuitable habitat or suitable habitat into which the species has not yet dispersed (Jarnevich et al., 2015).

While many of these models have been determined to effectively predict where species are likely to occur, they may not be rigorously validated. Many species habitat models use a subset of the original data to validate the model (Elith et al., 2006; Fielding and Bell, 1997). In such cases, the data are partitioned into training data to generate model predictions and testing data that are used to assess the accuracy of the model predictions. If the testing data are sufficiently predicted correctly by the model, then the model is considered to accurately predict the species' range. Since the testing data are a random sub-sample of the original dataset, information cannot be obtained on the accuracy of the model when applied to a larger region than that from which the original data came. Improved model evaluation can be obtained by incorporating independent field based presence and absence data, but this method is rarely used, particularly for invasive plant species (see Costa et al., 2010 and Rebelo and Jones, 2010 for examples using reptiles and bats, respectively).

* Corresponding author.
E-mail address: amanda.west@colostate.edu (A.M. West).

http://dx.doi.org/10.1016/j.ecoinf.2016.11.001
1574-9541/

Model comparisons can be used to evaluate multiple SDMs using both threshold dependent and threshold independent evaluation metrics. The area under the receiver operating characteristic (ROC) curve (AUC) is a commonly used threshold independent metric for evaluation of SDMs fit to true presence and absence data (Elith et al., 2006; Evangelista et al., 2008; Hosmer and Lemeshow, 2000; Swets, 1988). Test AUC (AUC_{TEST}) measures the ability of model predictions to discriminate between observed presence and absence for a test dataset (e.g., data held aside in a 10-fold cross validation split or independent test data), regardless of the absolute value of the predictions (Fielding and Bell, 1997; Elith and Graham, 2009). However, the use of AUC has its drawbacks. A low AUC value may indicate low discrimination between presences and absences even with a model that fits the data accurately (Lobo et al., 2008). AUC values also provide no information on the spatial distribution of incorrectly predicted presences and absences of a species (Lobo et al., 2008). Thus, AUC is useful in measuring how well presence locations can be discriminated from absences based on predictor variables, while providing little information about how well the model predictions fit the species distribution.

While AUC provides the ability of a model to discriminate between presences and absences, additional metrics can be used to evaluate SDMs developed using threshold selection methods based on study objectives. In the case of invasive plant species, management objectives may be tied to gaining the best possible understanding of where a given species exists on the landscape currently, which would encourage a model threshold based on maximizing sensitivity. Sensitivity measures the percentage of correctly classified presences, while specificity measures the percentage of correctly classified absences. Percent correctly classified (PCC) index considers both sensitivity and specificity. The true skill statistic (TSS = sensitivity + specificity − 1) places equal weight on model sensitivity and specificity, with values ranging between −1 and 1 (Allouche et al., 2006). Values above zero indicate better model performance than chance alone. Often, studies using SDMs rely on these threshold dependent metrics to evaluate and compare model performance and do not consider alternative indicators of model robustness, including comparisons in geographic space. Examples of model comparisons in geographic space include relative probability raster output comparisons using the two-tailed Wilcoxon signed-rank test and regression analysis, and binary raster output comparisons using the Kappa statistic.

The focus of this study was to examine a Maxent model developed with limited presence- only data for an invasive species, and evaluate its usefulness in a management context using threshold independent, threshold dependent, and geographic similarity comparison metrics. We modeled the distribution of cheatgrass because of the concern that land managers have about spread of this non-native species throughout high elevation plant communities (Bromberg et al., 2011; West et al., 2015). While modeling potential ranges of other species may be of interest as well, cheatgrass was of high priority to land managers in our study area, Rocky Mountain National Park. Tied to management objectives, the primary motivation of this study was to determine whether the predicted Maxent relative probabilities were strong indicators of where cheatgrass would be present. We used an independent presence and absence dataset collected during new field campaigns to validate initial Maxent model predictions, highlighting statistical robustness that cannot be obtained from partitioning the original data into training and testing subsets. Finally we combined the newly collected field data with the existing cheatgrass presence (and absence) data and compared Maxent to a commonly used presence-absence model, generalized linear model (GLM).

Our objectives were to: (1) generate an initial potential habitat suitability model for cheatgrass using Maxent fit with presence-only data, and use field sampling to test the predictions; (2) compare Maxent and GLM model predictions fit with the split-sample approach using threshold-dependent and threshold-independent metrics and comparisons in geographic space, and (3) identify the best fit model for management purposes.

2. Materials and methods

2.1. Study area

The study was conducted in Rocky Mountain National Park (referred as the Park hereafter), near the Colorado Front Range in the southern region of the Rocky Mountains. The elevation of the Park ranges from approximately 2300 m (7500 ft) in Estes Park to over 4300 m (14,100 ft) on Longs Peak. The Park is situated at latitudes of approximately 40°10′N to 40°32′N and longitude of 105°31′W to 105°41′ W (Peet, 1981). One main road traverses the Park running generally east to west, while additional roads run along the eastern border of the Park. The backcountry is accessible through 578 km (359 miles) of trails as they meander throughout the Park. Grasslands, shrub lands, and forests as well as rocky, non-vegetated areas were included in the study region. All of the sampling sites occurred within the Park and ranged in elevation from 2490 m to 3540 m.

The Park experiences an arid climate east of the continental divide with average annual precipitation of approximately 400 mm in Estes Park at the east side of the Park (WRCC, 2009). Approximately 480 mm of precipitation fall annually in Grand Lake at the west side of the Park (WRCC, 2009). Most of the total precipitation comes in the form of summer rain although the west side of the Park receives much more winter snowfall (WRCC, 2009). The growing season is short with snow often occurring into early June and returning in September and the potential for snow any month of the year. Average high temperatures in July are 25.7 °C with lows around 7.8 °C (WRCC, 2009). Average temperatures for the month of January range from a high of 3.5 °C to a low around −8.7 °C (WRCC, 2009). Extremely rapid changes in weather are a common occurrence in the Park.

2.2. Field methods

Cheatgrass presence data ($n = 21$) were collected in the Park using a modified Whittaker plot design between 1993 and 2007 (Stohlgren et al., 1995). A presence-only model for cheatgrass was developed in Maxent using these data (see Maxent 2007 in Modeling procedure). Relative probability output from this model was used to stratify field samples taken in 2008 through 2013; these field samples would later be used to validate the model. To stratify the field samples, random site coordinates in Universal Transverse Mercator (UTM) projection were generated in ArcGIS 9.2 (ESRI Inc., Redlands, CA, USA) and stratified among five relative probability classes (>0.1, 0.1–0.3, 0.3–0.5,0.5–0.8, and 0.8–1.0) of cheatgrass habitat suitability from Maxent 2007 (Bromberg et al., 2011). The coordinates were also stratified among vegetation communities and elevation to capture the available environment for cheatgrass; these covariates were two of the most influential environmental predictors from Maxent 2007. Distance to the nearest road or trail was also one of the top three environmental predictors, but was not used for stratifying sample locations. An array of distances from roads and trails would automatically be captured in the randomness of the stratified sampling. Elevation was grouped into six classes (<2500 m, 2500–2700 m, 2700–2900 m, 2900–3100 m, 3100–3300 m, >3300 m) for the purpose of stratifying site locations. Elevation of randomly generated sites ranged from 2396 m to 4023 m. Sites actually visited ranged from 2490 m to 3540 m in elevation. Missing presences of cheatgrass at higher elevations was not a concern, since the highest recorded specimen in Colorado was collected in 2004 at approximately 3050 m (Rocky Mountain Herbarium). That is substantially lower in elevation than many of the highest sites visited in this study. Distance to the nearest road or trail of randomly generated sites ranged from 30 m to 12,046 m with the farthest site visited at 8574 m from a road or trail. The sites were stratified among six vegetation communities, which comprised non-vegetated, shrubland, grassland, deciduous forest, coniferous forest, and tundra.