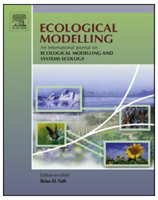
Ecological Modelling 363 (2017) 48–56

Contents lists available at ScienceDirect

Ecological Modelling

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



Minimizing effects of methodological decisions on interpretation and prediction in species distribution studies: An example with background selection

Catherine S. Jarnevich[a,*], Marian Talbert[b], Jeffery Morisette[b], Cameron Aldridge[c], Cynthia S. Brown[d], Sunil Kumar[e], Daniel Manier[a], Colin Talbert[a,b], Tracy Holcombe[a]

[a] U.S. Geological Survey, Fort Collins Science Center, 2150 Centre Ave Bldg C, Fort Collins, CO 80526, USA
[b] Department of Interior, North Central Climate Science Center, Colorado State University, Fort Collins, CO 80523, USA
[c] Natural Resource Ecology Laboratory, Colorado State University, in cooperation with the U.S. Geological Survey, Fort Collins Science Center, 2150 Centre Ave Bldg C, Fort Collins, CO 80526, USA
[d] Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO 80523-1177, USA
[e] Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523-1499, USA

ARTICLE INFO

Article history:
Received 23 December 2016
Received in revised form 16 June 2017
Accepted 8 August 2017
Available online 6 September 2017

Keywords:
Species distribution modeling
Habitat modeling
Niche modeling
Correlative models
Maxent
Boosted regression trees
Random forest
GLM
Background data

ABSTRACT

Evaluating the conditions where a species can persist is an important question in ecology both to understand tolerances of organisms and to predict distributions across landscapes. Presence data combined with background or pseudo-absence locations are commonly used with species distribution modeling to develop these relationships. However, there is not a standard method to generate background or pseudo-absence locations, and method choice affects model outcomes. We evaluated combinations of both model algorithms (simple and complex generalized linear models, multivariate adaptive regression splines, Maxent, boosted regression trees, and random forest) and background methods (random, minimum convex polygon, and continuous and binary kernel density estimator (KDE)) to assess the sensitivity of model outcomes to choices made. We evaluated six questions related to model results, including five beyond the common comparison of model accuracy assessment metrics (biological interpretability of response curves, cross-validation robustness, independent data accuracy and robustness, and prediction consistency). For our case study with cheatgrass in the western US, random forest was least sensitive to background choice and the binary KDE method was least sensitive to model algorithm choice. While this outcome may not hold for other locations or species, the methods we used can be implemented to help determine appropriate methodologies for particular research questions.

Published by Elsevier B.V.

1. Introduction

Understanding environmental conditions that allow a species to persist has been a fundamental question in ecology (Grinnell, 1917; Hutchinson, 1957; Soberón, 2007) and continues to be a pressing conservation priority. As originally described, the Grinnellian or fundamental niche considered a series of scenopoetic (i.e., abiotic) conditions that allowed for a species to persist (Grinnell, 1917; Hutchinson, 1957). However, Hutchinson (1957) and others recognized that biotic interactions, such as competitive exclusion, resulted in a species rarely utilizing its entire fundamental niche, referring to this smaller occupied space as the realized niche (Pulliam, 2000).

There has been a recent proliferation in the application of species distribution models (hereafter SDMs) in the ecological literature partly in response to the large availability of species occurrence data (Anderson, 2012) and spatial datasets (e.g., Porter et al., 2012), but also in part due to the increase in development and application of multiple SDMs (Zimmermann et al., 2010). SDMs attempt to understand the niche conditions (typically realized niche) that allow a species to persist, contrasting known presence locations with either known absence locations or some representative sample of potential available locations across space that characterize the range of environmental conditions available to the species, alternatively called background, available, or pseudo-absence locations. We will use the term 'background'. Presence data are often the only data collected and available (i.e., no absence data;

[*] Corresponding author.
E-mail address: jarnevichc@usgs.gov (C.S. Jarnevich).

http://dx.doi.org/10.1016/j.ecolmodel.2017.08.017
0304-3800/Published by Elsevier B.V.

Soberon and Peterson, 2005), especially over large spatial extents for which the time and cost to adequately sample is prohibitive, and for poorly sampled parts of the world. In these cases, SDMs are limited to those methods that only use presence information to define the niche (e.g., Ecological Niche Factor Analysis; Hirzel and Arlettaz, 2003) or background methods.

Several choices must be made during development of SDMs that can influence results, and there is not a quantitative methodology to direct decisions. Alternative choices add uncertainty to predictions, some of which can be quantified by comparing alternatives. In these studies that partitioned uncertainty among various choices, comparisons were made between modeling algorithm selection, location data choice and accuracy, predictor choice, climate change scenarios, method to control for collinearity in predictors, and variable selection method (e.g., Diniz-Filho et al., 2009; Dormann et al., 2008). Previous analyses of quantifiable uncertainty in model predictions highlight that modeling algorithm is often one of the greatest sources of uncertainty (e.g., Diniz-Filho et al., 2009; Dormann et al., 2008).

The practice of generating background locations is a form of *a priori* definition of the area accessible of a species, akin to prior selection of (independent) predictor variables, making careful and informed consideration of the background sample region essential to interpretable and useful results when models are used to extrapolate beyond sample units. The selection of background locations in presence-background SDMs is a subject of ongoing debate because this decision can affect model estimates (e.g., Phillips et al., 2009) and inflate model evaluation statistics (e.g., Rodda et al., 2011), but has not been included in the analyses of partitioning uncertainty described above. Regardless of the approach, selection should be related to the biological question of interest when defining the niche conditions for a given species. Several background point selection approaches have been explored, but so far no consistent, optimally performing method has emerged. To generate background points both the extent within which points will be generated and how points are placed within the extent should be addressed. Three main considerations apply to these decisions, including the biology of the species, the questions being asked and the potential sampling bias that often exists in presence-only datasets. Many earlier SDM studies selected background points randomly from the entire extent of the study area (e.g., Elith et al., 2006; Phillips et al., 2006). For applications using herbaria and museum data, research suggests targeted background or inventory pseudo-absence approach (e.g., Elith and Leathwick, 2007; Phillips et al., 2009), thereby comparing observations (collections) to the “full range” of environmental conditions in the target region. If doing so encompasses a range of unsuitable conditions for the species of interest, model prediction success would be high, but biological understanding of the niche requirements for the species would not be enhanced (e.g., temperate regions predicted unsuitable for tropical species). Thus, linking background sampling to the question of interest and understanding implications of the background method is imperative. Additionally, for datasets aggregated from disparate sources, such as multiple, independent survey or mapping efforts, target background locations may not be available and the aggregated location data may be clustered geographically (e.g., spatially disparate clusters representing disparate mapping efforts). This resulting sample selection bias can reduce the accuracy of SDMs (see Fourcade et al., 2014). Thus it is important to explore the impact of background selection on model results because these locations will influence model results.

Building upon the efforts of Barbet-Massin et al. (2012) who examined background selection uncertainty, our goal was to investigate the effects of a broader spectrum of background methods to evaluate the spatial extent and the spatial placement of available locations within that extent on predictions of SDMs using a ‘real’ dataset rather than a virtual species. Previous work highlights the importance of testing background selection methods for each dataset rather than a best method for all species-geographic extent combinations, and we outline a process to evaluate the effects of methods to select background points in conjunction with different SDMs. We evaluated six different SDMs of varying complexity using four different background-selection methods. Although there are many more algorithms commonly used and other methods to select background locations, we felt that this set of 24 pairs was enough to demonstrate the methodology. Our purpose was not to say what the ‘best’ pairing was but rather to evaluate a methodology to choose a pairing that minimized the effects of subjective decisions on model results. We explored random placement within the study area, random placement within a minimum convex polygon defined by presence data, random placement within a region defined by a kernel density estimator (KDE), and placement weighted by density of presence locations through a KDE (Fig. 1).

To conduct this assessment, we required a readily available dataset depicting the presence of a species across a large spatial extent, where existing covariates were available spatially. Modeling the distribution of cheatgrass (*Bromus tectorum*), an exotic invasive grass, is a good test candidate for this exercise because previous conservation and management efforts have resulted in an abundance of location data across the USA. Management concerns and challenges associated with the species indicate a clear need to better understand the current and potential influence of this species on wildlands and wildlife habitats at local and continental scales (Miller et al., 2011). Given a better understanding of environmental factors that affect cheatgrass distribution and abundance, land managers may focus limited resources on areas with the greatest threat, susceptibility, or both. Regional models also provide a link to scenario modeling efforts (i.e., application of climate and land-use scenarios) to support planning for potential future conditions. Importantly, to be useful for management and planning, models must represent “reality” observed by field managers and biologists. Analytically elegant, but inaccurate models may have little practical value. Therefore, this project presents results from our efforts to refine and improve regional SDM for conceptual and practical applications.

2. Materials and methods

2.1. Location data

We processed data within the VisTrails (Freire et al., 2006) Software for Assisted Habitat Modeling package (SAHM v 1.1; Morisette et al., 2013). We compiled cheatgrass point location data for the western USA from a variety of sources, resulting in 36,971 locations (Supplementary Table 1), reduced to 16,651 unique locations within 230 m resolution pixels. This cell size (230 m) was selected to facilitate integration of datasets derived from Moderate Resolution Imaging Spectroradiometer (MODIS). For model evaluation we obtained two independent datasets of cheatgrass presence and absence for a region of southwestern Wyoming (907 presence; 4882 absence) and a region in north central Nevada (360 presence; 204 absence; Supplementary Table 1).

We generated background locations equal in number to our presence locations to use in model development, following advice of Barbet-Massin et al. (2012) to have a large number of background points with equal weight to the presence locations and results from preliminary tests we conducted. In these preliminary analyses we examined histograms for each potential predictor vari-

Download English Version:

https://daneshyari.com/en/article/5742012

Download Persian Version:

https://daneshyari.com/article/5742012

Daneshyari.com