Ecological Modelling 330 (2016) 50–59

Contents lists available at ScienceDirect

Ecological Modelling

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

On the dangers of model complexity without ecological justification in species distribution modeling

David M. Bell[a,*], Daniel R. Schlaepfer[b,c]

[a] Pacific Northwest Research Station, U.S. Forest Service, Corvallis, OR, USA
[b] Department of Botany, University of Wyoming, Laramie, WY, USA
[c] Department of Environmental Sciences, Section of Conservation Biology, University of Basel, Switzerland

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form 18 March 2016
Accepted 20 March 2016
Available online 8 April 2016

Keywords:
Prediction
Extrapolation
Model fitting
Species distribution modeling
Transferability

ABSTRACT

Although biogeographic patterns are the product of complex ecological processes, the increasing complexity of correlative species distribution models (SDMs) is not always motivated by ecological theory, but by model fit. The validity of model projections, such as shifts in a species' climatic niche, becomes questionable particularly during extrapolations, such as for future no-analog climate conditions. To examine the effects of model complexity on SDM predictive performance, we fit statistical models of varying complexity to simulated species occurrence data arising from data-generating processes that assume differing degrees of distributional symmetry in environmental space, interaction effects, and coverage in climate space. Mismatches between data-generating processes and statistical models (i.e., different functional forms) led to poor predictive performance when extrapolating to new climate-space and greater variation in extrapolated predictions for overly complex models. In contrast, performance issues were not apparent when using independent evaluation data from the training region. These results draw into question the use of highly flexible models for prediction without ecological justification.

Published by Elsevier B.V.

1. Introduction

Monumental increases in the availability of ecological data and computing resources allows increasingly complex ecological models to be leveraged for predicting changes in biogeography. Increasing complexity in ecological models developed to represent species distributions in both geographic and environmental space is supported by the fact that those same distributions depend on a suite of processes associated with physiology (Buckley et al., 2011), demography (Pagel and Schurr, 2011), dispersal (Elith and Leathwick, 2009; Iverson et al., 2004), and biotic interactions (Parmesan and Yohe, 2003; Vanderwel et al., 2013; Walther et al., 2002). However, model complexity is sometimes motivated by the maximization of predictive performance, not ecological theory, as has been noted for correlative species distribution modeling (Austin, 2002, 2007). Correlative species distribution models (SDMs) are commonly used to assess habitat suitability as it relates to key environmental gradients (Elith and Leathwick, 2009) and generate global change predictions at regional to continental scales, painting a portrait of extreme biogeographic change under most, if not all, future scenarios of climate change (Parmesan and Yohe, 2003; Walther et al., 2002). While the SDMs do not generally model ecological processes constraining species occurrences as mechanistic approaches might (Ibáñez et al., 2006), the advent of virtual species simulation as a method of testing key assumptions of these models offers opportunities to assess model robustness and appropriateness (Meynard et al., 2013; Zurell et al., 2010). Model assessments exploring the impacts of failing to meet assumptions on predictive performance are not only needed to guide ecologists in choosing SDMs, testing their reliability, and interpreting their results (Aguirre-Gutierrez et al., 2013; Austin, 2007; Elith and Graham, 2009; Jimenez-Valverde et al., 2008), but for any ecological models used for prediction.

In part, the diversity of SDMs available for modeling reflects differences in assumptions about how species respond to environmental gradients. The prevalence of unimodal patterns of species occurrences along environmental gradients is well-supported (Austin, 2005; Gauch and Whittaker, 1972) and is assumed to represent suitability declines as conditions depart from the optimum (Austin and Smith, 1989; Heikkinen and Mäkipää, 2010). Asymmetric and symmetric unimodal patterns are both common (Austin and Gaywood, 1994; Austin and Van Niel, 2011; Boucher-Lalonde et al., 2012; Ellenberg, 1953), suggesting that

* Corresponding author at: 3200 SW Jefferson Way, Corvallis, OR 97331, USA. Tel.: +1 5417507298.
E-mail address: dmbell@fs.fed.us (D.M. Bell).

http://dx.doi.org/10.1016/j.ecolmodel.2016.03.012
0304-3800/Published by Elsevier B.V.

both can be reasonable representations of reality. The degree of symmetry is often interpreted as evidence of certain mechanisms controlling species distributions in environmental space, such as physiological constraints producing asymmetric distributions (Austin and Gaywood, 1994; Austin and Smith, 1989).

An apparent asymmetry in a species distribution may arise from truncation in the climate-space (Normand et al., 2009). Given that no-analog climates are likely to be common in the future (Williams and Jackson, 2007), ecological models that perform well under contemporary conditions may be unable to predict future changes. For example, contemporary patterns of conifer budbreak dates in western North America are negatively correlated with temperatures (i.e., earlier budbreak in warmer regions), but budbreak under future conditions may be delayed as chilling requirements are no longer met (Harrington and Gould, 2015). Because there may be no contemporary analogs to some future climates, models sensitive to truncation in the climate space will struggle in extrapolating to future conditions.

Although not generally discussed in relation to species distribution modeling, interactions might project asymmetries from one environmental gradient to another, as noted for 23% of European tree species (Boucher-Lalonde et al., 2012). As a result, an asymmetric species distribution might arise because an asymmetry or truncation along one environmental gradient influences suitability along the other gradient. Therefore, the source of observed asymmetries in species distributions is not trivial and is not necessarily easily accounted for in SDMs.

The source of complexity in species distributions, both environmentally and geographically, is at the heart of the debate concerning the complexity of SDMs. In recent years, SDMs increasingly utilize highly flexible correlative statistical models capable of accommodating a diverse suite of species distributional responses to environmental gradients (Elith and Leathwick, 2009). Increasing flexibility seems to improve model performance based on traditional cross-validation techniques (Santika and Hutchinson, 2009), but compared to simple models such as the generalized linear model (GLM), more flexible models such as the generalized additive model (GAM), random forest (RF) models, maximum entropy (MaxEnt) models, or boosted regression trees (BRT) may not perform well in terms of predicting into other regions or the future (Araujo et al., 2005; Randin et al., 2006; Schibalski et al., 2014; Merow et al., 2014). This dichotomy indicates that complex models may be fitting spurious patterns that are difficult to identify if evaluating model performance using cross-validation within the same region used to train the models; for instance, spatial autocorrelation is a concern for cross-validation when predicting within the same region (Le Rest et al., 2014) and also when predicting into novel climate space (Crase et al., 2014). While species responses to environment may be highly conditional on local landscapes and communities, increased model complexity may be difficult to interpret or may explain random variation not related to any ecological processes (i.e., overfitting). Thus, our confidence in model predictions to novel climate space is based on the ability of a model to reproduce the underlying processes contributing to species distributions (Evans et al., 2013).

In this study, we examine the influence of process and model complexity on ecological inference and prediction. Our objectives were to (1) determine how model complexity impacted performance when predicting species occurrence within a training region as well as extrapolating to other regions and (2) to explore the factors contributing the variation in performance, such as the underlying process generating the data (including random and spatial error), the model employed, and the sampling of data. We used a virtual species approach to simulate presence and absence data (e.g.; Meynard and Quinn, 2007) based on four different underlying climatic suitability processes, we fit SDMs of varying complexity, and we evaluated model performance in terms of observed species occurrence and underlying suitability processes within a single region and across regions (i.e., independent validation and transferability, respectively).

2. Materials and methods

2.1. Study area and climate data

In this study, we focus on the dry domain of the United States because it encompasses a large, climatically complex region in which simulation of species distributions could produce complex patterns (Fig. 1). The dry domain the of U.S. encompasses ecosystems ranging from the eastern slope of the Sierra Nevada and Cascade Mountains to the western Great Plains, from lowland deserts to montane forests to alpine meadows (Bailey, 1995). We divided the region into sub-regions based on state boundaries so that we could fit models to data from a single region (Northern Rocky Mountains [NR]) and test transferability of these models to other regions representing different climate-spaces (Southern Rocky Mountains [SR], Southwest [SW], and the Great Plains [GP]). All regions overlap climatically, but none share similar climatic extents (Fig. 1b–e), ensuring that species distribution models developed in NR would need to extrapolate to predict species geographic and environmental distributions in SR, SW, and GP. Therefore, these regions provide an appropriate case for testing model transferability.

Climate data were extracted for a grid of sample locations, with points located uniformly at 1/32° intervals, resulting in 12,583 total sample points. To ensure climatic realism, we extracted 30-year climate normal (1981–2010) from the 30 arc-second (approximately 800 m) PRISM data set (PRISM Climate Group, 2012) and calculated log winter (November to March) precipitation (dm) and minimum annual temperature (°C). We chose these variables because (1) snowpack and extreme winter temperatures are often incorporated into plant species distribution models in the region (e.g., Rehfeldt et al., 2006), (2) the correlation between these variables was intermediate (Pearson correlations $r = 0.22$, -0.22, -0.58, and 0.63 for NR, SR, SW, and GP, respectively), and (3) these variables resulted in somewhat divergent climate-space among the four sub-regions (Fig. 1b–e). The coverage of the climate space for different regions was assessed by examining the similarity of univariate and multivariate climate within the training region (NR) to the projection regions (SR, SW, and GP) as measured with the NT1 and NT2 indices defined by Mesgaran et al. (2014). Both in terms of univariate and multivariate climate space, large portions of the projection regions were climatically similar to the training region, but that greatest dissimilarity between training and projection regions occurred in the southern portions of the study region (Fig. 2).

2.2. Simulation experiment design

To examine the influences of the species occurrence data generating process on model prediction, we employed a virtual ecologist approach (Meynard et al., 2013; Meynard and Quinn, 2007; Zurell et al., 2010) wherein we designed a simulation experiment that allowed us to vary the data generating process, the models used to describe species environmental distributions, and the climate-space provided to the models for statistical inference. While there are many SDM approaches reported on in the literature, in this research, we focus on GLM, GAM, RF, MaxEnt, and BRT models with and without interactions to represent a gradient in model complexity because (1) many ecologists and species distribution modelers are familiar with them, (2) computation with model fitting is relatively simple and fast, and (3) the main objective was to test the effect of model complexity and not to evaluate specific models. As a result, the current study uses a relatively small series of models