Ecological Modelling 341 (2016) 5–13

Contents lists available at ScienceDirect

Ecological Modelling

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

Novel methods to select environmental variables in MaxEnt: A case study using invasive crayfish

Zeng Yiwen, Low Bi Wei, Darren C.J. Yeo*

Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, Republic of Singapore

ARTICLE INFO

Article history:
Received 1 April 2016
Received in revised form
15 September 2016
Accepted 21 September 2016
Available online 28 September 2016

Keywords:
Tuning
Stepwise removal
A priori
Distribution
Ecological niche
Species distribution model

ABSTRACT

The popularity of MaxEnt in species distribution modeling has been driven by several factors including its high degree of accuracy, and flexibility to tailor efforts to species-specific situations. Although many recent studies have identified the importance of adjusting mathematical transformation (feature class) and regularization of coefficient values, collectively known as tuning, few studies have addressed the need to customize the variables used in species distribution modeling, and use unselected variable sets. This study presents two novel methods to select for environmental variables in MaxEnt. The first involves selecting from a priori determined environmental variable sets (pre-selected based on ecological or biological knowledge), and the second utilizes a reiterative process of model formation and stepwise removal of least contributing variables. Both methods were tested on eight known species of invasive crayfish, with results reinforcing the need for species-specific environmental variable sets. While the reiterative process generally performs better than the a priori selected variables, selection of method can be based on information availability. These techniques appear to outperform the current practice of utilizing unselected variable sets and is especially important considering the increasing application of species distribution modeling (across spatial and temporal barriers) in conservation and management efforts whereby inaccurate predictions might have adverse effects.

1. Introduction

Grounded in the machine-learning maximum entropy framework, the MaxEnt software package (Phillips et al., 2006) allows users to predict a species potential distribution by utilizing presence-only species distribution data and a set of environmental variables (e.g., elevation and temperature). It is currently one of the most popular tools used for species distribution and environmental niche modeling, and has been used in numerous fields of biology. These range from biogeography and phylogeny (e.g., Schmidt-Lebuhn et al., 2015) to conservation biology (e.g., Warren et al., 2014), epidemiology (e.g., Cardoso-Leite et al., 2014), and invasion biology (e.g., Iñiguez and Morejón, 2012; Palaoro et al., 2013). Its wide usage is owed to its robustness to low sample sizes and relatively high predictive accuracy (Pearson et al., 2006; Wisz et al., 2008), coupled with the ease of use and flexibility in construction of models (Elith et al., 2006; Peterson et al., 2007; Merow et al., 2013).

In particular, the flexibility to adjust the mathematical transformations (or features) applied to the environmental variables allows users to cater models for individual target species and their specific purposes (see Muscarella et al., 2014). Users of the program also have the freedom to select a range of regularization coefficient values in order to maximize predictive accuracy for species-specific studies (Warren and Seifert, 2011; Muscarella et al., 2014). Although the selection of species-appropriate regularization values might reduce the need to adjust feature selection (Merow et al., 2013), accounting for both a variety of feature selection and a range of regularization coefficient values (collectively known as tuning) have been noted to produce more accurate models (e.g., Shcheglovitova and Anderson, 2013; Radosavljevic and Anderson, 2014). Besides adjusting features and regularization coefficients, numerous studies have also identified other types of adjustments to develop more accurate models within MaxEnt. These adjustments include techniques such as spatial filtering (Boria et al., 2014), inclusion and adjustments of spatial bias files (Kramer-Schadt et al., 2013; Warren et al., 2014), and selecting appropriate resolution (Jiménez-Alfaro et al., 2012) and background extent (VanDerWal et al., 2009).

Among the approaches, one that has garnered relatively lesser attention is the selection of environmental variables (or predictors).

* Corresponding author.
E-mail address: dbsyeod@nus.edu.sg (D.C.J. Yeo).

http://dx.doi.org/10.1016/j.ecolmodel.2016.09.019
0304-3800/

Thus far, studies investigating the effects of adjusting environmental variables have shown that using different environmental datasets (e.g., from WorldClim or IPCC) (Peterson and Nakazawa, 2008) and a reduced number of variables (Warren et al., 2014) have a notable impact on models formed. This is perhaps unsurprising considering variable selection is a vital portion of any species distribution modeling effort (Araújo and Guisan, 2006). In spite of this, few applications of MaxEnt have accounted for the effect variable selection has on overall model performance. Instead, some studies construct models with a pre-selected set of variables, including variables chosen in previous studies of other species or chosen based on a biological understanding of target species (e.g., Rödder et al., 2009; Palaoro et al., 2013). This corresponds with the ideology that all models should be formed based on a deep understanding of the species biogeography, ecology, population dynamics and human disturbance (Araújo and Guisan, 2006). However, in most cases (that lack detailed large-scale studies on ecological factors that influence a species' distribution), determining which set of biologically meaningful variables should be used is not as straightforward. This difficulty in selecting the best set of environmental variables, coupled with the known importance and influence that environmental variables have on predictive models necessitates the development of a technique for variable selection used in species distribution models.

This consideration is further compounded by the issue of "transferability", where species distribution models calibrated using knowledge of where a species occurs naturally (its realized niche) do not account for disparate environmental conditions where the species might otherwise occur if freed of dispersal and biotic constraints (its fundamental niche) (Soberón, 2007; Larson et al., 2010; Rodda et al., 2011). Given that many recent studies on species distribution models are concerned with taxa undergoing significant range shifts into novel environmental space, such as in the case of climate change (e.g., Warren et al., 2014) or human-mediated dispersal (e.g., Palaoro et al., 2013), the uncertain "transferability" of models over time and geographic space poses a serious obstacle to improving current and future species prediction techniques. This is because many range-shifting species violate assumptions of equilibrium upon translocation to a new habitat, and require extrapolation when predicting range extent in novel environmental space that have not been adequately sampled (Elith et al., 2010). However, invasive species with long histories of human-mediated translocation and establishment outside their native distributions could present excellent case studies for the species distribution modeler, as these species are likely to be 1) at equilibrium with the new environment after a long period of establishment, and 2) already occupying their full fundamental environmental (Grinnellian) niche across native and invaded ranges (Araújo and Pearson, 2005; Václavík and Meentemever, 2012), thereby improving "transferability" (Capinha et al., 2011; Larson et al., 2014).

Therefore, utilizing invasive crayfish as a case study (as species at equilibrium), and driven by the growing use of AICc (Akaike information criterion corrected for small sample sizes) as an evaluation tool (Warren and Seifert, 2011; Muscarella et al., 2014; Ficetola et al., 2014; Moreno-amat et al., 2015), we investigated two novel methods that have the potential to select environmental variables for species distribution models. The first technique draws from the ideas of information theory (IT) (see Whittingham et al., 2006; Hegyi and Garams, 2011) and data modeling (DM) approaches (see Breiman, 2001; Warren and Seifert, 2011) whereby several sets of models are constructed using a pre-selected list of environmental variables. Based on several biologically driven hypotheses, the best a priori-determined variable set is then chosen based on the best fit to the data. The second technique on the other hand is based the ideas of stepwise regression (SR) (Whittingham et al., 2006; Hegyi and Garams, 2011) and algorithmic modeling (AM) approaches (Breiman, 2001; Warren and Seifert, 2011), despite its use of AICc as an evaluation technique. This latter technique assumes that the best performing combination of environmental variables is an unknown, and aims to approximate this optimum set though a stepwise removal of the least contributing variables. To this end, we applied these two techniques to eight invasive crayfish with well-documented invasion histories and known negative impacts. We suggest that our methods can inform management of other invasive species and prove useful in prioritizing conservation efforts of threatened ecosystems, as well as contribute greatly to the toolbox of species distribution modelers working on various fields of biology.

2. Materials and methods

2.1. Study species and occurrence records

We selected eight crayfish—*Cherax destructor*, *Cherax quadricarinatus*, *Orconectes immunis*, *Orconectes limosus*, *Orconectes rusticus*, *Orconectes virilis*, *Pacifastacus leniusculus* and *Procambarus clarkii*—as target organisms for this study for several reasons. Firstly, these crayfish include closely related species (two species from the genus *Cherax* and four species from the genus *Orconectes*), and representatives from all three crayfish families (Cambaridae, Astacidae and Parastacidae) (Lodge et al., 2012). These crayfish also have a range of distribution sizes—from globally distributed invasive species such as *Pr. clarkii* and *Pa. leniusculus* to invasive species with ranges not exceeding a single continent (e.g., *O. rusticus* and *C. destructor*). These species have a long history of invasion, with non-native populations are likely to have reached a state of equilibrium (Václavík and Meentameyer, 2012). Also known for their detrimental effects to local ecosystems and economies (e.g., Gherardi, 2007; Lodge et al., 2012) these crayfish represent well-documented examples of globally relevant invasive species.

To form species distribution models, we collated a total of 219 occurrence records for *Cherax destructor*, 160 records for *Cherax quadricarinatus*, 309 records for *Orconectes immunis*, 2112 records for *Orconectes limosus*, 511 records for *Orconectes rusticus*, 930 records for *Orconectes virilis*, 1680 records for *Pacifastacus leniusculus,* and 1141 records for *Procambarus clarkii*. These records were gathered from a variety of sources including published literature (e.g., Beatty et al., 2005; Snovsky and Galil, 2011; Chucholl, 2012; Torres and Álvarez, 2012), online databases (e.g., Global Biodiversity Information Facility (GBIF) (http://www.gbif.org); Invasive Species Compendium (ISC) (http://www.cabi.org/isc/)), and museum collection records (e.g., Smithsonian Institute National History, Australian Museum) (Appendix A. Supplementary data). These species occurrence records represent both the native and alien (established) ranges of each species, with records less than 10 years, and eradicated populations, filtered out of alien distribution records.

2.2. Spatial extent and pseudo-absences

Considering the importance of including both alien and native range occurrence data to determine the invasive potential for species (see Mandle et al., 2010), coupled with the need to limit background extent (see Barve et al., 2011), this study utilizes a 250-arc-minutes buffer radius (roughly 500 km) around each occurrence record to limit spatial extent. This buffer was included for both native and alien (established) ranges, with the choice of buffer radius motivated by known ranging capabilities of invasive crayfishes (e.g., Gherardi et al., 2000; Bubb et al., 2004; Anastaício et al., 2015). Following recommendations by Phillips (2008) and