Ecological Informatics 32 (2016) 69–78

Contents lists available at ScienceDirect

Ecological Informatics

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

Data prevalence matters when assessing species' responses using data-driven species distribution models

Shinji Fukuda [a],*, Bernard De Baets [b]

[a] Institute of Agriculture, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu, Tokyo 183-8509, Japan
[b] KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, 9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form 14 January 2016
Accepted 15 January 2016
Available online 25 January 2016

Keywords:
Data-driven model
Ecological trait
Prevalence
Species distribution model
Species–environment relationship
Virtual species

ABSTRACT

The study of species' response is a key to understand the ecology of a species (e.g. critical habitat requirement and biological invasion processes) and design better conservation and management plans (e.g. problem identification, priority assessment and risk analysis). Predictive machine learning methods can be used as a tool for modeling species distributions as well as for describing important variables and specific habitat conditions required for a target species. This study aims (1) to demonstrate how habitat information such as species response curves can be retrieved from a species distribution model (SDM), (2) to assess the effects of data prevalence on model accuracy and habitat information retrieved from SDMs, and (3) to illustrate the differences between three data-driven methods, namely a fuzzy habitat suitability model (FHSM), random forests (RF) and support vector machines (SVMs). Nineteen sets of virtual species data with different data prevalences were generated using field-observed habitat conditions and hypothetical habitat suitability curves under four interaction scenarios governing the species–environment relationship for a virtual species. The effects of data prevalence on species distribution modeling were evaluated based on model accuracy and habitat information such as species response curves. Data prevalence affected both model accuracy and the assessment of species' response, with a stronger influence on the latter. The effects of data prevalence on model accuracy were less pronounced in the case of RF and SVMs which showed a higher performance. While the response curves were similar among the three models, data prevalence markedly affected the shapes of the response curves. Specifically, response curves obtained from a data set with higher prevalence showed higher tolerance to unsuitable habitat conditions, emphasizing the importance of accounting for data prevalence in the assessment of species–environment relationships. In a practical implementation of an SDM, data prevalence should be taken into account when interpreting the model results.

1. Introduction

Species distribution models (SDMs) are an essential tool to analyze species–habitat relationships, offering valuable information for conservation decisions (see Guisan et al., 2013 for a comprehensive review). In conservation planning, habitat suitability maps have been used to identify potential sites that are important for a target species, whereas species response curves, describing a species' response to a given habitat condition, can illustrate specific habitat requirements for the species. Specifically, response curves can support the decision-making process for identifying and comparing designs and options for species conservation and management.

It is widely known that the model accuracy of an SDM is not independent of the quality and quantity of data such as size (i.e. the number of data points in a data set) and data prevalence (i.e. the proportion of presences in a data set). This is partly because a species absence can have multiple meanings: environmental absences (i.e. habitat conditions are not suitable for a species), contingent absences (i.e. habitat conditions are suitable but other factors such as biotic interactions, barriers to dispersion or local extinction are responsible for the absence of the species), and methodological absences (i.e. the species is present but not detected) (Lobo et al., 2010), whereas observed presence is factual. The effects of data prevalence on SDMs have been studied using real (Fukuda and De Baets, 2012; Václavík and Meentemeyer, 2012) and virtual species data (Austin et al., 2006; Barbet-Massin et al., 2012; Jiménez-Valverde et al., 2009; Lauzeral et al., 2012; Zurell et al., 2012). Virtual species data are commonly used in these studies because they provide perfect knowledge and allow for control over the uncertainties, whereas the mechanisms that drive real species distributions are unknown or only qualitatively understood by experts. Santika (2011) reviewed and summarized that the effects of data prevalence on model accuracy can be influenced by modeling algorithms, species–habitat dependency, performance measures and threshold selection methods that are used to convert the fitted probability of occurrence into presence–absence of a species (see also references

* Corresponding author. Tel./fax: +81 42 367 5604.
E-mail address: shinji-f@cc.tuat.ac.jp (S. Fukuda).

http://dx.doi.org/10.1016/j.ecoinf.2016.01.005
1574-9541/© 2016 Elsevier B.V. All rights reserved.

Nomenclature

AM	Arithmetic mean
AUC	Area under the receiver operating characteristic curve
CART	Classification and regression tree
FHSM	Fuzzy habitat suitability model
GA	Genetic algorithm
GAM	Generalized additive model
GLM	Generalized linear model
MX	Mixed interaction
MSE	Mean squared error
PR	Product
RF	Random forests
SDM	Species distribution model
SISO	Single-input single-output
SVM	Support vector machine

therein). Despite abundant literature, these studies focus mainly on model accuracy, often disregarding the ecological relevance of a species' response which can be retrieved quantitatively using SDMs. While accurate models are often regarded as complex and uninterpretable, there have been several approaches proposed from ecological modeling and informatics (Fukuda, 2011; Olden et al., 2004) in order to visualize so-called “black-box” models. As such, extracting ecological information from predictive SDMs is of great importance for conservation decisions. Such information can improve existing knowledge and methods for tackling important questions in ecology such as niche shift and biological invasions (Guisan et al., 2014).

Various techniques have been applied for building SDMs based on presence–absence data, including statistical methods such as generalized linear models (GLMs) and generalized additive models (GAMs), and computational intelligence techniques such as artificial neural networks, classification and regression trees (CARTs), fuzzy systems, genetic algorithms (GAs), random forests (RF) and support vector machines (SVMs) (see Guisan and Zimmermann, 2000, Ahmadi-Nedushan et al., 2006 and Elith and Leathwick, 2009 for reviews). Hybridizations of these techniques have also been used, but such application studies are limited (see Van Broekhoven et al., 2007 and Fukuda et al., 2011 for genetic fuzzy systems and Fukuda et al., 2006 and Fukuda, 2011 for fuzzy neural networks). Alternatively, ensemble modeling is becoming popular in SDM studies (Lauzeral et al., 2012; Oppel et al., 2012). Specifically, Lauzeral et al. (2012) demonstrated an iterative ensemble modeling approach using GLMs, GAMs, CARTs and RF in order to reduce noisy absences for better and reliable species distribution modeling. These methods were not designed for deriving response curves and no study has assessed the effects of data prevalence on species' responses. Extracting habitat information from such advanced modeling methods can contribute to a better understanding of ecological traits of a target species and better applications of predictive but complex and less interpretable modeling methods in ecology.

This study assesses how data prevalence affects model accuracy and habitat information retrieved from SDMs based on virtual species data with different prevalences. The virtual species data were generated based on a set of hypothetical univariate habitat suitability curves aggregated with different forms of interactions between habitat variables when calculating composite habitat suitability for the species. Three correlative SDMs were developed using fuzzy habitat suitability models (FHSMs), RF and SVMs in order to fit the virtual species data. Based on the results from the three SDMs, this study demonstrates how a univariate response curve, describing a partial response for a habitat variable, can be derived from a multivariate SDM and illustrates the differences in how these SDMs respond to data prevalence in species distribution modeling and the assessment of species–environment relationships.

2. Methods

2.1. Virtual species data

In this study, virtual species data were generated in order to avoid the uncertainties in the species distributions in response to habitat conditions. Three physical habitat variables, namely water depth (cm, henceforth referred to as depth), flow velocity (cm s^{-1}, velocity) and percent vegetation coverage (%, vegetation; defined as the proportion of area covered with aquatic plants in a surveyed point), were obtained from a series of field surveys conducted in an agricultural canal in Kurume City, Fukuoka, Japan (33°20′N, 130°42′E). Four independent data sets were derived from the surveys conducted on October 14, and November 5 and 9, 2004 and April 25, 2005. Each data set was comprised 139 data points in the first survey, 130 data points in the second survey, 86 data points in the third survey, and 88 data points in the fourth survey (i.e. 443 data points in total).

For generating virtual species data, hypothetical habitat suitability was defined and composite habitat suitability was calculated using observed habitat conditions. The hypothetical habitat suitability curves (Fig. 1) are based on the ecology of Japanese medaka (*Oryzias latipes*), a small freshwater fish found in and around paddy fields in Japan (see Fukuda and Hiramatsu, 2008 for habitat suitability), and were expressed by a sigmoid function for depth (Eq. (1)), a Gaussian function for velocity (Eq. (2)), and a piecewise linear (trapezoidal) function for vegetation (Eq. (3)):

$$S_{\mathrm{d},i} = \frac{1}{1+\exp(-0.5\times(d_i-10))} \tag{1}$$

$$S_{\mathrm{v},i} = \exp\left(-\frac{(4-v_i)^2}{2\sigma^2}\right) \tag{2}$$

$$S_{\mathrm{veg},i} = \begin{cases} \dfrac{veg_i}{30} & , \quad \text{if } 0 \le x_i < 30 \\ 1 & , \quad \text{if } 30 \le x_i < 40 \\ \dfrac{100-veg_i}{60} & , \quad \text{if } 40 \le x_i \le 100 \end{cases} \tag{3}$$

where $S_{\mathrm{d},i}$, $S_{\mathrm{v},i}$ and $S_{\mathrm{veg},i}$ are the habitat suitability for the individual habitat variables of depth, velocity and vegetation, respectively, d_i, v_i, and veg_i are the habitat variables in the ith survey point of the corresponding

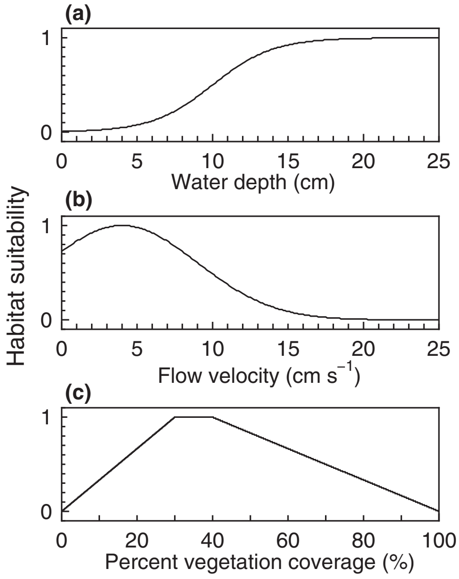

Fig. 1. Hypothetical habitat suitability curves: (a) water depth, (b) flow velocity and (c) percent vegetation coverage.