Environmental Modelling & Software 47 (2013) 1–6

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

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

Short communication

Habitat prediction and knowledge extraction for spawning European grayling (*Thymallus thymallus* L.) using a broad range of species distribution models

Shinji Fukuda [a,*], Bernard De Baets [b], Willem Waegeman [b], Jan Verwaeren [b], Ans M. Mouton [c]

[a] *Faculty of Agriculture, Kyushu University, Hakozaki 6-10-1, Fukuoka 812-8581, Japan*
[b] *KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, 9000 Ghent, Belgium*
[c] *Research Institute for Nature and Forest (INBO), Kliniekstraat 25, 1070 Brussels, Belgium*

ARTICLE INFO

Article history:
Received 1 October 2012
Received in revised form
23 April 2013
Accepted 27 April 2013
Available online 5 June 2013

Keywords:
Habitat modelling
Data-driven model
Machine learning
Knowledge extraction
Accuracy
Presence/absence data

ABSTRACT

This study aims to apply seven data-driven methods (i.e. artificial neural networks [ANNs], classification and regression trees [CARTs], fuzzy habitat suitability models [FHSMs], generalized additive models [GAMs], generalized linear models [GLMs], random forests [RF] and support vector machines [SVMs]) to develop data-driven species distribution models (SDMs) for spawning European grayling (*Thymallus thymallus*), and to compare the predictive performance and the ecological relevance, quantified by the habitat information retrieved from these SDMs (i.e. variable importance and habitat suitability curves [HSCs]). The results suggest RF to yield the most accurate SDM, followed by SVM, CART, ANN, GAM, FHSM and GLM. However, inconsistencies between different performance measures were observed, indicating that different models may obtain a high score on a particular aspect and perform worse on other aspects. Despite their lower predictive ability, GAM, GLM and FHSM proved to be useful, since HSCs could be obtained and thus these techniques allow testing of ecological relevance and habitat suitability. Water depth and flow velocity appeared to be important variables for spawning grayling. The HSCs clearly indicate higher habitat suitability at a lower water depth, a low to medium flow velocity and a higher percentage of medium-sized gravel, whereas the models disagreed on the habitat suitability for the percentage of small-sized gravel. These findings demonstrate the applicability of data-driven SDMs for both habitat prediction and ecological knowledge extraction that are useful for management of a target species.

1. Introduction

Data-driven species distribution models (SDMs) are useful tools to analyse species–environment relationships (Mouton et al., 2008, 2009, 2011; Pino-Mejías et al., 2010; Brás et al., 2013; Marsili-Libelli et al., 2013). Despite the increasing number of species distribution modelling studies, these studies have focused mainly on accuracy, often disregarding ecological relevance such as habitat suitability (but, see Meynard and Quinn (2007) and Pino-Mejías et al., (2010) in which variable importance was considered in model comparison). The habitat information obtained from an SDM can illustrate the species' response to a given habitat condition and provide insight into how the SDM works under a given condition (Olden et al., 2004; Jowett and Davey, 2007; Fukuda, 2011; Fukuda et al., 2011; Mouton et al., 2011). It is therefore important to compare a broad range of SDMs regarding predictive performance as well as habitat information for a better understanding of target species and an improved design of conservation and restoration plans.

This study aims to demonstrate the applicability of seven data-driven methods, namely artificial neural networks (ANNs), classification and regression trees (CARTs), fuzzy habitat suitability models (FHSMs), generalized additive models (GAMs), generalized linear models (GLMs), random forests (RF) and support vector machines (SVMs), to model species–habitat relationships of spawning European grayling (*Thymallus thymallus*). These methods include a broad range of data-driven models from statistics (GAMs and GLMs), artificial intelligence (ANNs and FHSMs [genetic-algorithm-optimized fuzzy models]) and machine learning (ANNs,

[*] Corresponding author. Tel.: +81 92 642 2918; fax: +81 92 642 2917.
E-mail address: shinji-fkd@agr.kyushu-u.ac.jp (S. Fukuda).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.04.005

CARTs, RF and SVMs). We assess the predictive performance using multiple performance measures, and then compare the habitat information, namely variable importance and habitat suitability curves (HSCs), to ecological knowledge on the spawning European grayling in literature.

2. Methods

2.1. Spawning habitat data

The spawning habitat data were obtained from an intensive monitoring campaign conducted in a 1300-m stretch of the Aare river in the Bern department, Switzerland (Mouton et al., 2008). The monitoring results were combined with the results of the hydraulic simulations, from which a data set (18,409 points) consisting of five variables (flow velocity (hereafter velocity) [m s^{-1}], water depth (depth) [m], percentage of small-sized gravel (%SG) [2 mm–20 mm; %], percentage of medium-sized gravel (%MG) [20–50 mm; %] and presence/absence of spawning ground [prevalence of 0.22]) was obtained (Fig. 1). Of these variables, the first four variables were used as model inputs, whereas the last variable was used as model output.

2.2. Species distribution models

For the ANN, we used the ‘e1071’ package (Dimitriadou et al., 2011) of the R environment (R Development Core Team, 2011), in which all habitat variables were normalized to their maximum. While the default settings for ANNs (i.e. three-layered perceptron with logistic activation functions) were used, two hyper-parameters, namely the number of units in the hidden layer (size: 5 × 1, 5 × 2,..., 5 × 10) and the weight decay constant (decay: 10^{-5}, 10^{-4},..., 10^{0}), were tuned using 3-fold cross-validation. In addition, the connection weight method (Olden et al., 2004) was used (personal code) for evaluating variable importance in the ANN computation.

For the CART, we used the ‘e1071’ package of the R environment. While the default settings for CARTs were used, the complexity parameter (cp: 10^{-5}, 5×10^{-5}, 10^{-4}, 5×10^{-4}, 10^{-3}, 5×10^{-3}, 10^{-2}, 5×10^{-2}, 10^{-1}, 5×10^{-1}, 10^{0}) was tuned using 3-fold cross-validation.

In the FHSM (Fukuda et al., 2011), fuzzy if-then rules, consisting of two parts, namely an antecedent part stating conditions on an input variable (i.e. each of the four habitat variables) and a consequent part describing the corresponding values of the output variable (i.e. the habitat suitability for the habitat variable), were used to relate the habitat conditions with the habitat suitability for spawning European grayling. A set of HSCs can be obtained by providing consecutive values (in steps of 0.1) to the FHSM and plotting the output values against the corresponding habitat variable. We implemented the FHSM using our personal code (Fortran 90; freely available upon request), in which the membership functions (Fig. 2) in the antecedent part were defined based on Mouton et al. (2008) and the singleton values in the consequent part were determined by a binary-coded genetic algorithm using 3-fold repeated holdout validation with 20 different sets of initial conditions.

For the GAM, we used the ‘mgcv’ package (Wood, 2011) of the R environment, in which the default settings (the use of the logit link function and the assumption of binomial error distribution) were used except for the basis dimension for %MG ($k = 7$, corresponding to the number of its unique values: Fig. 1). To extract habitat information, the estimated degrees of freedom in a fitted GAM were used as a variable importance measure. In addition, the ‘plot.gam’ function was used to plot the fitting result for each term as a measure of habitat suitability (Jowett and Davey, 2007).

For the GLM, we used the ‘glm’ function of the R environment as a logistic regression model, in which the default settings (the use of the logit link function and the assumption of binomial error distribution) were used. To extract habitat information, Z scores (i.e. estimated coefficients of a fitted GLM divided by their standard errors), indicating significance of a habitat variable, were used as a variable

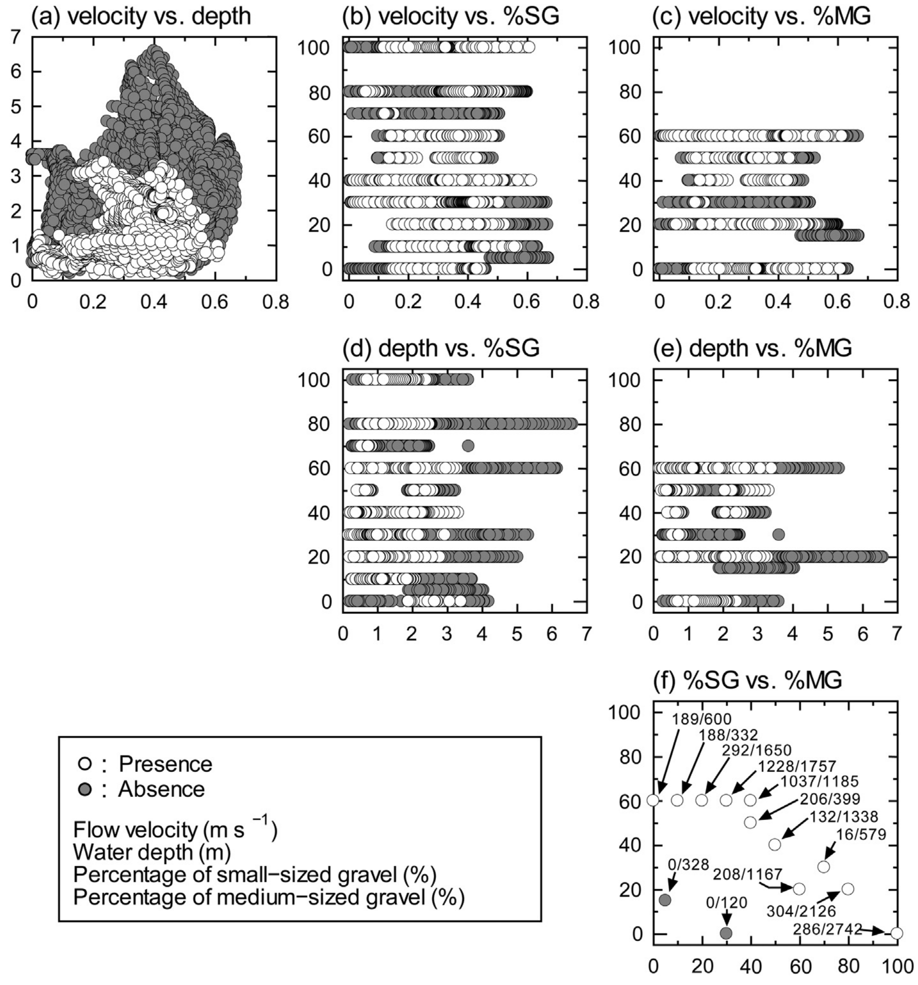

Fig. 1. Scatter diagrams of each pair of the four habitat variables: flow velocity (velocity), water depth (depth), percentage of small-sized gravel (%SG) and percentage of medium-sized gravel (%MG), in which the number of presence/absence points is presented for the plot of %SG versus %MG (f).