Ecological Informatics 35 (2016) 9–18

Contents lists available at ScienceDirect

Ecological Informatics

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

Assessing bluegill (*Lepomis macrochirus*) habitat suitability using partial dependence function combined with classification approaches

Risa Shiroyama *, Chihiro Yoshimura

Department of Civil and Environmental Engineering, Tokyo Institute of Technology, 2-12-1-M1-4, Ookayama, Meguro-ku, Tokyo 152-8552, Japan

ARTICLE INFO

Article history:
Received 27 April 2016
Received in revised form 24 June 2016
Accepted 27 June 2016
Available online 29 June 2016

Keywords:
Species distribution model
Partial dependence function
Bluegill
Habitat suitability
Classification approach

ABSTRACT

Classification approaches have been used to understand the habitat suitability of key species. Partial dependence function is an especially useful concept despite of a lack of studies that compare the results of the function against observations. Furthermore, there has been scant research investigating the relative performance of classification approaches for describing habitat suitability. Thus, we aim to assess the applicability of partial dependence function combined with classification approaches to describe habitat suitability of the bluegill *Lepomis macrochirus*, a riverine fish in the Kanto region of Japan. A total of 425 samples, along with eight environmental variables, were surveyed by the National Censuses on River Environments and were used throughout this research. Five classification approaches were combined with a partial dependence function to assess the habitat suitability of bluegill. The areas under the curves based on the training and test data were calculated 100 times using each of the five classification approaches. Additionally, partial dependence on individual and paired environmental variables was estimated using each combination of a five-classification approach and partial dependence function, and this dependence was plotted to determine the habitat suitability of bluegill. As a result, random forest approach demonstrated high predictive accuracy compared to other classification approaches. The combination of the partial dependence function and random forest described the peaks of habitat suitability observed in the field, both using individual and paired environmental variables. Moreover, habitat suitability based on pairs of environmental variables also indicated that bluegill changes their habitat throughout the year.

1. Introduction

Understanding habitat suitability specifically enhances the development of long-term conservation plans (Ju and Santos, 2012). Thus, there is increasing interest in understanding species distributions, habitat suitability, and habitat ecology in the context of environmental management across terrestrial, freshwater, and marine systems (Cheung et al., 2009; Hermoso et al., 2015; Rydgren et al., 2003). A relevant number of researches on species distribution and habitat suitability have been published in the last decade and classification approaches have consequently been applied to ecological data in order to explain their relation (Guisan et al., 2013; Zimmermann et al., 2010). However, the application of classification approaches to environmental management has not been sufficiently described (França and Cabral, 2015) and requires more practical assessments on the use of such approaches (Guisan et al., 2013).

Advances in computer technology have enabled advanced classification approaches such as classification and regression trees (CART; Breiman et al., 1984), random forests (RF; Breiman, 2001), and generalized boosted regression models (GBM; Friedman, 2001). These approaches do not require any assumptions regarding population distributions and estimate model parameters directly from empirical data. In recent years, ecologists have realized the usefulness of these approaches and have widely applied them to aquatic ecosystems. The capability of these approaches for modeling the presence and absence of freshwater fish has been supported by published studies (Conti et al., 2015; Hopkins and Roush, 2013). The advantage of these approaches is now widely known, yet they have been used less frequently compared to conventional classification approaches, such as generalized linear models (GLM; Nelder and R. Jacob, 1972) and generalized additive models (GAM; Hastie and Tibshirani, 1990). This tendency may be explained by the complexity of the advanced classification approaches. The relationship between the fitted response variables and environmental variable is relatively transparent in traditional classification approaches, whereas this relationship is placed into a “black box” in advanced classification approaches (Elith et al., 2008).

In addition, the visualization of species' responses to environmental variables is valuable for understanding habitat suitability and the

* Corresponding author at: Dept. of Civil Engineering, Tokyo Institute of Technology, 2-12-1-M1-4, Ookayama, Meguro-ku, Tokyo 152-8552, Japan.
E-mail addresses: shiroyama.r.aa@m.titech.ac.jp (R. Shiroyama), yoshimura.c.aa@m.titech.ac.jp (C. Yoshimura).

http://dx.doi.org/10.1016/j.ecoinf.2016.06.005
1574-9541/

behavior of organisms. These visualized plots constructed from a classification approach have been given various names such as "prediction curves," "response curves," "species response curves," and "habitat preference curves" (Austin, 2002; Fukuda, 2011; Rydgren et al., 2003), and for simplicity, the present study collectively refers to these terms as "habitat suitability plot." Some methods have been demonstrated to improve our understanding of habitat suitability. For instance, suitability indexes have been used as a means of assessing habitat suitability for managers and decision makers in natural resources management (e.g., Brooks, 1997; Vélez-espino, 2006). Suitability indexes depict the relationship between an environmental variable and scored habitat suitability of target species. In general, these plots are constructed using ecological traits of target species as determined from the published literature and field surveys (Vinagre et al., 2006). Other research has proposed a method of visualizing the fitted response of target species to sets of variables (Elith et al., 2005; Liang et al., 2013); this method keeps all variables constant except for the variable of interest, and fitted response variables can then be plotted over the range of the target variables (e.g., averaging method, Elith et al., 2005). This method can be applied to any classification approach, and it has demonstrated the importance of using bivariate plots to detect complex interactions (Elith et al., 2005; França and Cabral, 2015; Liang et al., 2013). Additionally, Fukuda (2011) examined the habitat suitability of *Oryzias latipes* by using fuzzy neural networks and visualizing the habitat suitability of the fish against each environmental variable. In these studies, and many others in the literature, the visualization of species' responses to environmental variables continues to be illustrated (e.g., Fukuda and De Baets, 2016; Olden et al., 2004; Zurell et al., 2012).

Partial dependence function (Friedman, 2001) may present a possibly superior method for illustrating habitat suitability in combination with classification approaches. This function represents the effect of selected explanatory variables, incorporates the effects of the other variables, and can be used to interpret "black box" models (Cutler et al., 2007; Elith et al., 2008). Some studies have applied such functions to revealing habitat suitability of their target species (Altermatt et al., 2013; Conti et al., 2015; Maloney et al., 2013; Martinez-capel et al., 2015), yet most studies did not validate these analyses against observations or other published research. Moreover, there is little research on the relative performance of classification approaches for describing habitat suitability combined with partial dependence function.

The present study aims to assess the performance of a combination of partial dependence function and classification approaches to describe habitat suitability of bluegill, one of the notable exotic riverine fish species in Japan. To achieve this objective, we focused on the following goals: (1) investigating which classification approaches are able to accurately predict the presence/absence of bluegill, (2) validating its habitat suitability using combinations of each of the classification approaches and a partial dependence function based on both field observations and published data, and (3) depicting suitable habitat conditions of bluegill in a quantitative manner using the developed distribution model.

2. Study region and dataset

2.1. Study region and monitored data

This study analyzed data from seven rivers in the Kanto region of Japan, which were selected for model construction and habitat suitability assessment of the bluegill *Lepomis macrochirus* (Table A.1). This region has a large, central plain with mountains in its northwestern areas, and it is open to the Pacific Ocean on its eastern shore. All the data used in this study were collected from the National Censuses on River Environments (NCRE). This survey has been conducted by the Ministry of Land, Infrastructure and Transport (MLIT) of Japan since 1990 (MLIT, 2006a) and aims to monitor riverine species distributions for use in managing river resources and promoting ecological research on riverine ecosystems (MLIT, 2006a). NCRE collects data on the presence and absence of riverine species (e.g., fish, invertebrates, birds, and plants) as well as relevant environmental variables from 109 Class-A rivers once every 5 years (for fish and invertebrates) or every 10 years (for other riverine species). This survey is conducted from spring to autumn at least two times in each river, and the timing and frequency of the survey are determined from the ecology of target communities and climate conditions (MLIT, 2006b). In each river, areas were established for detailed investigation, covering longitudinal sections measuring up to 1 km (MLIT, 2006a). Within each study area, one or more sampling locations were established, and environmental variables and presence or absence of the riverine species were monitored at each sampling location.

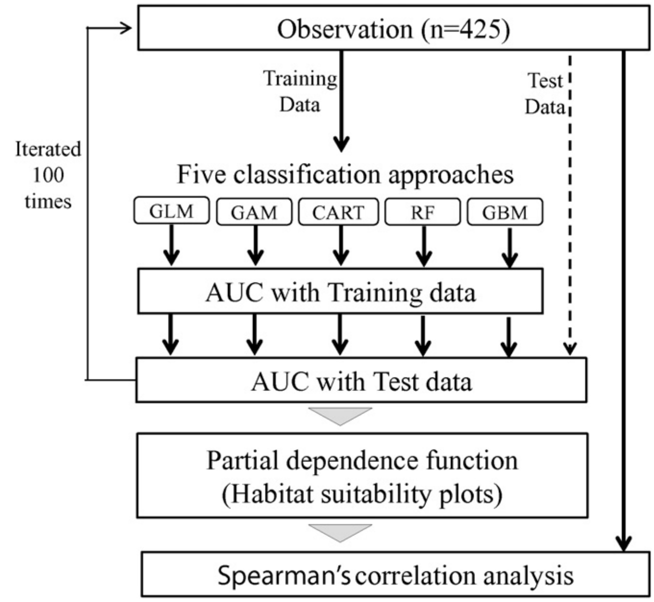

Fig. 1. Flowchart describing the process of estimating habitat suitability using classification approaches combined with a partial dependent function: AUC, the area under the curve; GLM, generalized linear model; GAM, generalized additive model; CART, classification and regression trees; RF, random forests; and GBM, generalized boosted regression models.

In the present study, the sample data collected in the period from 2006 to 2010 were used for constructing classification models and validating habitat suitability. These samples were collected mainly in June and October, thus we divide them into the two periods to investigate the seasonal habitat variability of bluegill: June for the first period and October for the second period. If the survey was not conducted in June and October, the closest survey to these months was used instead (Table A.1). Sampling locations that have survey data from both the first and the second periods were considered in this study. The numbers of all study areas and sampling locations in the Kanto region for the target period were 132 and 527, respectively. After removing data from those sampling locations that were not sampled twice in a year, the dataset was comprised of data from 131 study areas with 425 sampling locations, which were subsequently analyzed in this study.

2.2. Environmental variables

For the habitat assessment, we included eight environmental variables throughout this literature: distance from the river mouth (in case of tributaries distance from confluence), longitudinal slope in each area, as well as depth, velocity, and water temperature in the