Ecological Modelling 251 (2013) 44–53

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

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

Application of genetic algorithm and greedy stepwise to select input variables in classification tree models for the prediction of habitat requirements of *Azolla filiculoides* (Lam.) in Anzali wetland, Iran

Roghayeh Sadeghi [a,*], Rahmat Zarkami [b], Karim Sabetraftar [b], Patrick Van Damme [a]

[a] Department of Plant Production, Faculty of Bio-Science Engineering, Ghent University, Coupure links, 653, 9000 Ghent, Belgium
[b] Department of Environmental Science, Faculty of Natural Resources, University of Guilan, P.O. Box 1144, Sowmehsara, Guilan, Iran

ARTICLE INFO

Article history:
Received 11 September 2012
Received in revised form
19 November 2012
Accepted 6 December 2012
Available online 9 January 2013

Keywords:
Azolla
Classification tree
Genetic algorithm
Greedy stepwise
Predictive models

ABSTRACT

The aim of the present study was to predict to what extent wetland characteristics can affect the habitat requirements of an exotic species, *Azolla filiculoides* (Lam.) in wetland. Biotic and abiotic variables were collected at the Selkeh wildlife refuge (a protected area in Anzali wetland, northern Iran) over the study period 2007–2008. Classification tree (CT) was used to find the relationship between the wetland characteristics and the cover percentage of *A. filiculoides*. Genetic algorithm (GA) and greedy stepwise (GS) were combined with CT in order to select the most important variables to explain the coverage of *A. filiculoides*. The applied method was assessed based on the percentage of correctly classified instances (CCI) and Cohen's kappa statistics (k). Different pruning confidence factors (PCFs) were tested in order to improve the predictive results regarding the complexity and accuracy of the prediction. The results showed that the prediction was reliable in terms of both performance criteria. Yet, after variable selection, the predictive performances of the CT improved. Due to potential collinear variables in the model, the GS method was less efficient than GA. The optimization of GA and GS resulted in an easy interpretation of the selected variables. The methods showed that both structural habitat (e.g. air temperature, humidity and depth) and physico-chemical variables (e.g. nutrients) can affect the habitat requirements of *A. filiculoides* in the wetland but the dependence of this aquatic fern on structural habitat was well confirmed by the CT before and after variable selection. Application of the given algorithms in combination with CT thus proved to have a better capability in selecting the most important variables explaining the cover of *A. filiculoides* and can be used by wetland managers in their decision-making for wetland conservation and management programs.

1. Introduction

The aim of habitat models is to link the presence/absence or abundance of a species at a site to abiotic factors that explain their general habitat (Guisan and Zimmerman, 2000; Zarkami, 2011; Zarkami et al., 2012). The predictive models can be important to ecological studies in order to assess, monitor and manage natural resources (D'heygere et al., 2006; Dakou et al., 2007; Zarkami et al., 2010; Sadeghi et al., 2012a). Predictive modeling is one of the most essential steps in the development of a standard habitat assessment protocol (Parsons et al., 2004).

Classification trees (CTs) (Quinlan, 1993) are an example of a predictive model that can be used even with small datasets (Goethals et al., 2007; Everaert et al., 2011; Zarkami et al., 2012) and they are being widely used due to their simplicity and flexibility (Ambelu et al., 2010; Zarkami et al., 2010, 2012). Based on their transparency, they can be useful for decision-makers to improve the efficiency of monitoring and assessment (Guisan and Zimmerman, 2000). Further, they are fairly simple to construct and their transparency allows for easy integration into an environmental decision support system. Džeroski et al. (1997) were among the first to describe applications of CT in river community analysis.

The availability of reliable datasets, choosing suitable analysis techniques, and the identification of environmental variables that are able to explain the system are very important stages when predicting the presence/absence, distribution, abundances, and habitat needs of organisms (Faraway and Chatfield, 1998; Sadeghi et al., 2012b). Too many variables involved in the dataset will be costly and unmanageable (Dom et al., 1989), and they can also lower the predictive performances and reliability of models (Hall and Holmes,

* Corresponding author. Tel.: +32 92646093; fax: +32 92646241.
E-mail addresses: roghayeh.sadeghipasvisheh@ugent.be, jila.sadeghi2003@yahoo.com (R. Sadeghi), rzarkami2002@yahoo.co.uk (R. Zarkami), karim.sabetrafter@gmail.com (K. Sabetraftar), Patrick.VanDamme@UGent.be (P. Van Damme).

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2012.12.010

2003). Therefore, selection of the suitable variables is important since it might improve the predictive performances (Goethals et al., 2007; Sadeghi et al., 2012b).

Genetic algorithm (GA) has been shown to be an important method to cope with variable selection (D'heygere et al., 2006; Zarkami et al., 2010). In fact, GA mimics aspects of biological information processing for data modeling and could be helpful in ecological science (Recknagel, 2001). GA can improve the accuracy and predictive power of a model so that the numbers of irrelevant variables are reduced (Harrell et al., 1996). It is a widespread method for the optimization of predictive models, in particular in the field of aquatic ecology (D'heygere et al., 2006; Hoang et al., 2010). A popular approach is the use of GA as an optimization tool for resetting the parameters in other classifiers such as CT, especially in the field of aquatic ecology (D'heygere et al., 2006). Greedy stepwise (GS) is also a variable selection method (forward selection and backward elimination search) that can be used to select the most important input variables (Gevrey et al., 2003; Butterworth et al., 2004; Ambelu et al., 2010).

Invasive species have always been a major concern for biologists and ecologists dealing with conservation and management of aquatic ecosystems. Invaders may threaten the biological diversity of native species, resulting in permanent changes in community structure and ecosystem (Vitousek et al., 1996; Ricciardi and MacIsaac, 2000; Sax et al., 2005; Van der Zanden and Olden, 2008). Therefore, knowing the habitat requirements of invasive species is important for predicting how the species will use a new environment.

Azolla is a particular example of such an invasive species in Iran. It is a genus of aquatic ferns and small-leafed floating plants, native to the tropics, subtropics, and warm temperate regions of Africa, Asia, and the Americas (Sweet and Hills, 1971). This genus is considered to constitute as the world's fastest growing aquatic macrophytes. Based on many reports (e.g. Lumpkin and Plucknett, 1982; Zimmerman, 1985; Taghi-Ganji et al., 2005), the doubling time of *Azolla* is between 2 and 5 days. Although, *Azolla* species have many advantages (e.g. N_2-fixation and removal of heavy metals), they are considered as aggressive weeds. Among *Azolla* species, *Azolla filiculoides* is one of the most common global invaders (Barreto et al., 2000), including Iran (Delnavaz and Ataei, 2009). In the last few years, many aquatic ecosystems in Iran (especially in the northern part) have been invaded by this fern.

In the literature, many studies can be found about the *Azolla–Anabaena* association and N_2-fixation in *Azolla* (e.g. Van Hove and Lejeune, 2002; Fernández-Zamudio et al., 2010) while almost no studies have been conducted in order to predict the habitat requirements of *Azolla* as an exotic species invading a new ecosystem.

This present study is the first practical work that aims to use the CT method combined with the two optimzers to predict the habitat needs of *A. filiculoides* based on a set of wetland characteristics at Selkeh wildlife refuge which is a protected area in the Anzali wetland, northern Iran (the Selkeh is currently under protection by various agencies such as fishery research organization and Guilan Environmental Protection Bureau). Furthermore, another aim of the present study was to determine if wetland variable selection by the GA and GS methods could improve predictive performances of the CT, and hence lead to a reliable prediction for *A. filiculoides* in the study area. The given techniques were chosen because they can be used when only a small datasets is available. This study presents a method for selecting the most important explanatory variables for wetland conservation, restoration and management programs. Therefore, this type of study can be used to identify variables for monitoring to prevent future spreading or options for treating *A. filiculoides*.

2. Methods

2.1. Description of study area

Anzali wetland is an international aquatic ecosystem based on Ramsar convention (1971) which is located in northern Iran and south west of the Caspian Sea (Fig. 1). This wetland consists of shallow impoundments, large, shallow, eutrophic freshwater

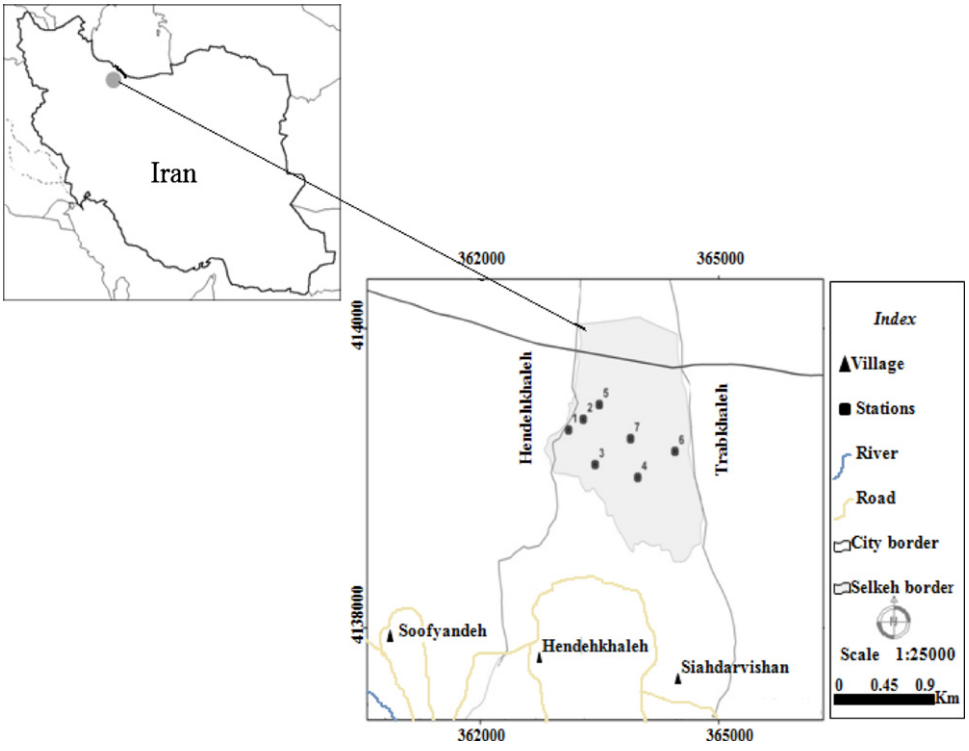

Fig. 1. Location of the sampling sites at Selkeh wildlife refuge, south of Anzali wetland, northern Iran.