Environmental Modelling & Software 66 (2015) 17–26

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Predicting fish species richness in estuaries: Which modelling technique to use?

Susana França [a, *], Henrique N. Cabral [a, b]

[a] MARE – Marine and Environmental Sciences Centre, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[b] Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 14 July 2014
Received in revised form
1 December 2014
Accepted 12 December 2014
Available online

Keywords:
Predictive modelling
Species richness
Estuaries
GLM
GAM
CART
BRT

ABSTRACT

Four different modelling techniques were compared and evaluated: generalized linear models (GLM), generalized additive models (GAM), classification and regression trees (CART) and boosted regression trees (BRT). Each method was used to model fish species richness variation throughout several Portuguese estuarine systems. Model comparisons were based on goodness-of-fit and predictive performance via cross-validation. The relative influence of the most important predictors according to each of the four models was also examined. Fitted BRT, CART, GAM and GLM models accounted for 70.6%, 57.0%, 34.6% and 23.7% of total model deviance, respectively. No single variable was consistently responsible for the larger amount of percentage of relative deviance explained by the models, but several variables were selected by the four models. Nevertheless, their relative importance was highly variable, according to each modelling technique. The tree-based models (CART and BRT) presented lower prediction errors after cross-validation. The limitations and usefulness of each technique are discussed.

1. Introduction

Assessing how geographical and spatial variability acts on biological patterns and in the processes at their origin has always been a major goal in ecology (Leathwick et al., 2005). Therefore, an increased demand to explore and predict the relationships between environment and biota distribution is now occurring (e.g. Olden, 2003; Bahn and McGill, 2012; Fukuda et al., 2013). Predictive modelling techniques have been increasingly used, and are currently helping researchers to determine major habitat requirements for species distribution. The use of such predictive models encompass areas such as biogeography, spatial ecology, conservation biology, climate change and environmental management and allow for considerable progress to be made in crucial ecological topics such as habitat loss and fragmentation and climate change impacts (Araújo and Guisan, 2006; Maggini et al., 2006; Meynard and Quinn, 2007). Consensus on the appropriate data to be used when building these models, methodology or interpretation, as well as the conceptual framework on which species predictive models are built, have not been reached (Austin, 2007). Linear regression is known to be one of the most widely used statistical techniques to investigate how environmental variables influence species' distribution patterns and occurrence, mainly due to its easy use and straightforward interpretability (Aertsen et al., 2010). Nevertheless, ecological data are often complex, unbalanced, present a non-constant variance distribution and contain missing values (Potts and Elith, 2006). Strong nonlinear relationships are often expected between ecological variables and species occurrence patterns. Conventional modelling strategies, such as linear regression will therefore result in non-significant predictions, leaving high unexplained variation (Austin, 1999; De'ath and Fabricious, 2000; Aertsen et al., 2010).

Important developments have benefited predictive distribution modelling (Guisan and Zimmerman, 2000; Segurado and Araujo, 2004) and new and sophisticated techniques have been developed in the types of statistical models applied to ecology (Leclere et al., 2011). Advanced statistical and machine-learning techniques allowed to improve the capacity to predict complex systems, namely through their capacity to detect and represent nonlinear and highly interactive relationships (Austin, 2007). These techniques combined with the growing accessibility of geodatasets at high spatial resolution are increasingly being used with greater accuracy in studies of the relationships between plant or animal

* Corresponding author. Tel.: +351 217500826; fax: +351 217500207.
E-mail address: sofranca@fc.ul.pt (S. França).

http://dx.doi.org/10.1016/j.envsoft.2014.12.010
1364-8152/

communities and their environment (Guisan and Thuiller, 2005; Aertsen et al., 2010; Knudby et al., 2010).

Several methods are now available for modelling either individual species distribution and occurrence, as well species richness variation, each varying in the way they model the response distribution, choose relevant predictors, define fitted functions, weight variable contributions, allow for interactions and predict patterns of occurrence (Elith et al., 2006). Several studies have compared the performances of these statistical techniques and concluded that there is no general best modelling technique that should be applied to all situations. The type of model used must be chosen according with the type of the relationships between the environment and the occurrence and distribution of the species, as well as the most adequate for a particular situation and goal (Segurado and Araujo, 2004; Elith et al., 2006; Aertsen et al., 2010; Meynard and Kaplan, 2012).

Estimating species richness (i.e. the number of species present in a determined area) has become a common feature in ecological modelling as this parameter constitute a common and basic step of most field studies carried out in community ecology and is now subject of increased interest with recognition that provides a reasonably and useful way to measure biodiversity (Leathwick et al., 2006). Species richness may be predicted using different approaches: species distribution models may be applied to individual species in an assemblage, and then distributions are overlayed and predictions are summed for a determined area to derive species richness; or predictions may be performed using species richness directly, as the response variable for a given model (Gotelli et al., 2009). The present study is focused on the latter, mainly because comparison of different modelling techniques in predicting species richness variation are scarce and use mostly the first approach.

Regarding marine and coastal environments, there has been an increased interest for modelling fish–habitat relationships, fish distribution (Eastwood et al., 2003; Le Pape et al., 2007; Vasconcelos et al., 2013) and fish community features, such as the response of fish species richness to environmental variables over large and local scales (Nicolas et al., 2010; França et al., 2012). Comparative studies of different modelling techniques in predicting fish species richness variation in these ecosystems are scarce and only few studies evaluated which modelling method produces the most accurate spatial predictions (Knudby et al., 2010; Leclere et al., 2011).

In the present study we compared the application of a set of different modelling techniques, namely: Generalized Linear Models (GLM) and Generalized Additive Models (GAM), which can be considered as conventional non-parametric statistical models; and two "Machine-Learning" methods: Classification and Regression Trees (CART) and Boosted Regression Trees (BRT). To do this we modelled fish species richness in several estuarine systems of the Portuguese coast, according to environmental variables and habitat characteristics. For each modelling approach, we determined which variables had highest influence in species richness variation, compared the influence of each selected variable and evaluated the predictive performance.

2. Material and methods

2.1. Study area

Five estuarine systems along the Portuguese coast were considered in the present study: Ria Aveiro, Tejo, Sado, Mira and Guadiana (Fig. 1).

These systems differ substantially in terms of their geomorphologic and hydrologic characteristics (Table 1): Tejo and Sado are large systems with areas of 320 km^2 and 180 km^2, respectively, while Mira is the smallest with 5 km^2. Mean river flow values are considerably higher in the Tejo estuary (300 m^3 s^{-1}) and this system also presents the largest estuary mouth width (5.3 km). In addition, this estuary also has the highest value of the anthropogenic pressure index (0.76), according to Vasconcelos et al. (2007), while Mira presents the lowest one (0.14). Moreover, habitat complexity (score based on the structure and patchiness of the habitats present in the estuary, with higher scores attributed to estuaries with more complex habitat structures and larger areas of the different habitats) was high for Ria de Aveiro (score 3), medium in Tejo and Sado (2) and low in Mira and Guadiana (score 1). Shallow areas are a common feature in all the estuarine systems, with mean depths varying between 1 and 6 m (Table 1).

Two areas presenting similar environmental variables were selected in each estuarine system, and were considered sampling replicates. In both areas, three

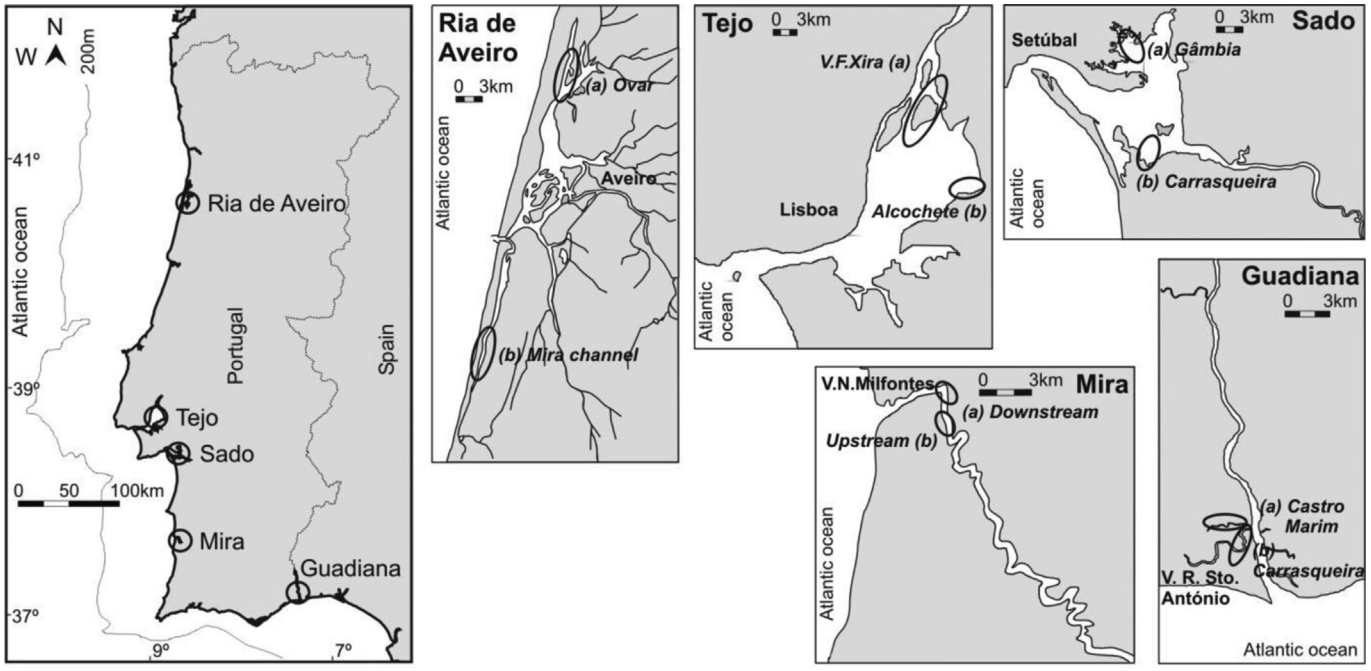

Fig. 1. Estuarine systems sampled in the Portuguese coast. Also shown is the location of sites within each estuary where the three habitats (saltmarsh, mudflat and subtidal) were sampled.