Environmental Modelling & Software 53 (2014) 112–120

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Optimizing biodiversity prediction from abiotic parameters

Androniki Tamvakis [a,*], Vasilis Trygonis [a], John Miritzis [a], George Tsirtsis [a], Sofie Spatharis [a,b]

[a] University of the Aegean, Department of Marine Sciences, University Hill, 81100 Mytilene, Greece
[b] University of Glasgow, Institute of Biodiversity, Animal Health and Comparative Medicine, Glasgow G12 8QQ, Scotland, UK

ARTICLE INFO

Article history:
Received 1 March 2013
Received in revised form
29 November 2013
Accepted 2 December 2013
Available online

Keywords:
Diversity indices
Phytoplankton
Simulated assemblages
Machine learning
Instance based learning

ABSTRACT

An integrated methodology is proposed for the effective prediction of biodiversity exclusively from abiotic parameters. Phytoplankton biodiversity was expressed as richness, evenness and dominance indices and abiotic parameters included temperature, salinity, dissolved inorganic nitrogen and phosphates. Prediction was based on three machine learning techniques: model trees, multilayer perceptron and instance based learning. To optimize diversity prediction, indices were calculated on a large number of phytoplankton field assemblages, but also on corresponding noise-free simulated assemblages. Biodiversity was most accurately predicted by the instance based learning algorithm and the efficiency was doubled with simulated assemblages. Based on the optimal algorithm, indices, and dataset, a software package was developed for phytoplankton diversity prediction for Eastern Mediterranean waters. The proposed methodology can be adapted to any group of organisms in marine and terrestrial ecosystems whereas important applications are the integration of community structure in ecological models and in assessments of global change scenarios.

Software availability

Name of software: PREPHYB
Developers: Androniki Tamvakis and Vasilis Trygonis
Contact address: Department of Marine Sciences, University of the Aegean, University Hill, 81100 Mytilene, Greece
Tel.: +30 22510 36811
Fax.: +30 22510 36809
E-mail: atamvaki@mar.aegean.gr
First available: 2013
Software required: (a) for MATLAB users: MS Windows or Mac or Linux; (b) for non-MATLAB users: MS Windows
Programming language: MATLAB R2010a
Program size: (a) for MATLAB users: zipped file of 0.5 MB; (b) for non-MATLAB users: zipped file of 160 MB
Availability and online documentation: http://www.mar.aegean.gr/biodiv/Prephyb
Cost: freely available

1. Introduction

Diversity prediction through a number of biotic and abiotic parameters is currently a challenging issue in ecology (Ingram and Steel, 2010; Gontier et al., 2006). Obtaining a measure of diversity from field data is not always feasible due to constraints related to the taxonomic analysis of samples (Maurer, 2000). However, estimates of diversity are essential when it comes to prioritizing sites for management purposes (Lockwood et al., 2012), for assessing the ecological status of ecosystems (WFD, 2000; Spatharis and Tsirtsis, 2010) or for predicting effects of global change on ecosystem diversity and function (Dawson et al., 2011). In this context, it is essential to develop methodologies that provide a realistic prediction of diversity based on a small number of abiotic parameters that are more straightforward to measure.

Recently, the emergence of powerful tools as the Machine Learning (ML) techniques and their application in ecology has significantly advanced the predictive power of models (Fielding, 1999; Kuo et al., 2007; Li et al., 2011). These techniques are effective for exploring complex ecological processes, and can handle nonlinearity without relying on implicit assumptions on the relationships between parameters (Dzeroski and Drumm, 2003; Jeong et al., 2008; Junker et al., 2012; Kanevski et al., 2004). However, few attempts have been made so far to apply ML techniques for biodiversity prediction. Most studies are still based on classical statistical

Abbreviations: CV, cross validation; IBk, instance based learning algorithm; LOOCV, leave one out cross validation; ML, machine learning; MAD, mean absolute deviation; MTs, model trees; MLP, multilayer perceptron; MLR, multiple linear regression; NN, neural network; PO_4, phosphates; RMSE, root mean square error; *S*, salinity; *T*, temperature.

[*] Corresponding author.
E-mail address: atamvaki@mar.aegean.gr (A. Tamvakis).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.12.001

approaches such as regression analysis (Arias-Gonzalez et al., 2012; Brakstad et al., 1994; Denisenko, 2010; Thrush et al., 2001) which are constrained by assumptions on data such as normality, homoscedasticity or colinearity. In this context, application of ML approaches seems most prominent in marine ecosystems that are affected by multidimensional, complex and stochastic phenomena often characterized by non-linearity (Olden et al., 2008).

Among the most frequently applied ML algorithms are Model Trees (MTs), Neural Networks (NNs) and Instance Based learning (IBk). These algorithms represent the three main ML categories (trees, neural networks and lazy algorithms) that use completely different predictive approaches (Solomatine et al., 2008). These span many applications in ecology (Dzeroski, 2001; Lek and Guegan, 1999; Recknagel, 2001) whereas in the marine environment they have been used in hydrodynamics, wave forecasting, habitat modelling, biomass prediction, and pollution assessment (e.g. Dakou et al., 2007; Etemad-Shahidi and Mahjoobi, 2009; Millie et al., 2012; Solomatine et al., 2006; Tamvakis et al., 2012; Tian et al., 2011). Concerning biodiversity prediction in particular, application of ML techniques in both marine and terrestrial ecosystems has been based on habitat features, biotic characteristics or a combination of both with some abiotic parameters (Cheng et al., 2012; Debeljak et al., 2007; Demsar et al., 2006; Dominguez-Granda et al., 2011; Dzeroski and Drumm, 2003; Jurc et al., 2006; Knudby et al., 2010; Kocev et al., 2009; Pittman et al., 2007). These studies have focused on one biodiversity component (e.g. species richness or Shannon diversity) whereas so far there has been no attempt to predict different diversity components (richness, evenness, and dominance) exclusively from abiotic parameters related to the physical and chemical environment.

Diversity can be expressed through a number of indices which quantify community structure and the changes it undergoes due to natural or anthropogenic stress (Magurran, 2004). However, field communities are also driven by multiple stochastic factors such as seasonality and spatial heterogeneity which impose a degree of uncertainty and distortion on data (Van Straten, 1992). This 'environmental noise' inherent in field communities is also reflected on the subsequent calculation of indices (Vounatsou and Karydis, 1991). This problem can be overcome with the use of simulated communities *via* a species abundance distribution (e.g. the log-series, lognormal) however retaining the structure of field ones (Blackwood et al., 2007; Lyashevska and Farnsworth, 2012; Schloss and Handelsman, 2006; Spatharis and Tsirtsis, 2010). Calculations on noise-free simulated communities seem appropriate when trying to establish cause-and-effect relationships, e.g. between diversity and abiotic parameters, due to the removal of noise or distortion that more easily supports the revealing of possible signals.

In this paper we propose an integrated methodology for the optimization of diversity prediction exclusively from abiotic parameters (Fig. 1). The diversity is expressed by diversity, evenness, and dominance indices calculated on both field and simulated phytoplankton assemblages covering a wide productivity range typical of Eastern Mediterranean waters. Predictions were carried out based on three ML algorithms. The objectives of the study were thus: (a) to distinguish the ML technique offering the most accurate prediction, (b) to select the indices representative of all three diversity components (richness, evenness, and dominance) (c) to optimize prediction by calibrating the methodology with indices calculated on simulated assemblages, and (d) to develop a software tool for biodiversity prediction based on the proposed methodology.

2. Methodology

2.1. Datasets

The first dataset employed in the study includes 658 field samples and was compiled using existing data from coastal areas of the Aegean Sea, E. Mediterranean representing a wide range of productivity. At each station of a coastal area, repetitive sampling was carried out covering at least a full annual cycle on a monthly basis. Detailed information about the sampling sites and data collection are provided in Spatharis et al. (2008). Inner Saronikos Gulf, near Athens, and Kalloni Gulf in Lesvos Island are characteristic of eutrophic conditions (Simboura et al., 2005). Outer Saronikos Gulf and Gera Gulf in Lesvos Island are more typical of mesotrophic conditions (Arhonditsis et al., 2000; Ignatiades et al., 1992), while offshore stations in Rhodes Island have been characterized as oligotrophic (Kitsiou et al., 2002). A detailed account on the eutrophication level and ecological status of these areas is provided in Spatharis and Tsirtsis (2010). Among the various abiotic parameters available in the dataset, a subset was selected for the aims of the present study, including: (a) concentrations of limiting nutrients, Dissolved Inorganic Nitrogen (DIN) and Phosphates (PO_4), that directly influence the growth and composition of phytoplankton in the areas under consideration (Spatharis et al., 2008) and (b) Salinity (S) and Temperature (T), which may also indirectly affect phytoplankton synthesis through stratification in coastal waters (Spyropoulou et al., 2013). Nutrient concentrations were measured spectrophotometrically according to Parsons et al. (1984), whereas physical variables were recorded *in situ*. Moreover, phytoplankton species-abundance data were used in the current study analysed following the same protocol according to the inverted microscope method of Utermöhl (1958). Dataset information and summary statistics of the above parameters in each of the four areas are provided in Table 1. The dataset covers a wide range of phytoplankton abundance (10^3–9×10^6 cells/L) and species richness (4–39 species). There were no missing values in the dataset and no special treatment was performed for outlying values. It was considered that the latter often correspond to extreme events such as algal blooms due to episodic terrestrial inputs (Spatharis et al., 2007) or to the photoperiod increase during spring, that have to be included in the models to be developed. The variables' positive skewness (Table 1), that is almost always observed for environmental data, was taken into account in the application of the ML algorithms. According to the requirements of each algorithm standardization or normalization procedures were applied, described in detail below.

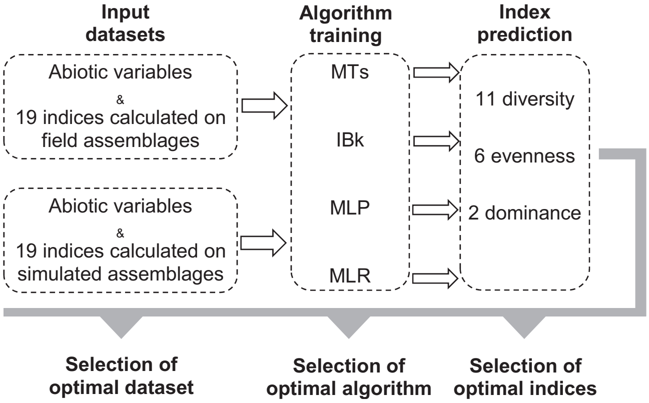
Input datasets
Algorithm training
Index prediction
Abiotic variables
&
19 indices calculated on field assemblages
Abiotic variables
&
19 indices calculated on simulated assemblages
MTs
IBk
MLP
MLR
11 diversity
6 evenness
2 dominance
Selection of optimal dataset
Selection of optimal algorithm
Selection of optimal indices

Fig. 1. Conceptual diagram of the methodological procedure followed in order to optimize diversity prediction from abiotic parameters.

The second dataset includes 658 simulated phytoplankton assemblages with abundances corresponding exactly to the abundances of the 658 field samples. The simulation was based on the log-series statistical distribution which assumes that most species in an assemblage are rare (Fisher et al., 1943). The log-series distribution is shaped by parameters x and a, that can be calculated knowing the ratio of species richness to total abundance (S/N) in an assemblage. The S/N ratio was estimated *via* a simple linear regression equation between S and N using the 658 field samples as described in Spatharis and Tsirtsis (2010). Regression analysis was also used to identify the relation of the abundance of the most dominant species N1 with the total phytoplankton abundance N in the 658 field samples. When parameters x and a were estimated, the expected number of species S was allocated for each abundance (total cells N). By feeding the previous two relationships which characterize field phytoplankton assemblages onto the log-series distribution, simulated assemblages are generated that retain the structure of the initial field ones (Fig. 2). This approach has been described in detail in previous studies (Spatharis and Tsirtsis, 2010; Tsirtsis et al., 2008) resulting in a wide range of assemblage diversity closely matching reality (Spatharis et al., 2011).

2.2. Indices expressing diversity components

Indices can express different aspects of biological diversity such as richness, evenness, and dominance. Thus, diversity indices weigh more on the richness