Ecological Informatics 25 (2015) 35–42

Contents lists available at ScienceDirect

Ecological Informatics

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

Evaluating machine-learning techniques for recruitment forecasting of seven North East Atlantic fish species

Jose A. Fernandes [a,b,c,*], Xabier Irigoien [b,d], Jose A. Lozano [c], Iñaki Inza [c], Nerea Goikoetxea [b], Aritz Pérez [c]

[a] Plymouth Marine Laboratory, PL1 3DH Plymouth, UK
[b] AZTI—Tecnalia, Marine Research Division, Herrera Kaia z/g, E-20110 Pasaia, Spain
[c] University of the Basque Country, Department of Computer Science and AI, Intelligent Systems Group (ISG), Paseo Manuel de Lardizabal, 1, E-20018 Donostia-San Sebastián, Spain
[d] King Abdullah University of Science and Technology (KAUST), Red Sea Research Center, Thuwal 23955-6900, Saudi Arabia

ARTICLE INFO

Article history:
Received 25 September 2014
Received in revised form 16 November 2014
Accepted 18 November 2014
Available online 24 November 2014

Keywords:
Pelagic fish
Fisheries management
Recruitment forecasting
Bayesian networks
Supervised classification
Kernel density estimation

ABSTRACT

The effect of different factors (spawning biomass, environmental conditions) on recruitment is a subject of great importance in the management of fisheries, recovery plans and scenario exploration. In this study, recently proposed supervised classification techniques, tested by the machine-learning community, are applied to forecast the recruitment of seven fish species of North East Atlantic (anchovy, sardine, mackerel, horse mackerel, hake, blue whiting and albacore), using spawning, environmental and climatic data. In addition, the use of the probabilistic flexible naive Bayes classifier (FNBC) is proposed as modelling approach in order to reduce uncertainty for fisheries management purposes. Those improvements aim is to improve probability estimations of each possible outcome (low, medium and high recruitment) based in kernel density estimation, which is crucial for informed management decision making with high uncertainty. Finally, a comparison between goodness-of-fit and generalization power is provided, in order to assess the reliability of the final forecasting models. It is found that in most cases the proposed methodology provides useful information for management whereas the case of horse mackerel is an example of the limitations of the approach. The proposed improvements allow for a better probabilistic estimation of the different scenarios, i.e. to reduce the uncertainty in the provided forecasts.

1. Introduction

Early on in fisheries research, recruitment was identified as a key element in management. As a result, recruitment and the factors determining it have been the subject of intense research (e.g. Cushing, 1971; Myers et al., 1995; Ricker, 1954; Rothschild, 2000). Such research has evolved from considering only the biomass of spawners, to including also environmental factors that can modulate recruitment (e.g. Planque and Buffaz, 2008; Schirripa and Colbert, 2006). The main limitation to achieve good forecasts, from a data analysis perspective is the sparse and 'noisy' nature of the available data (Fernandes et al., 2010; Francis, 2006).

A further problem is that data about some of the factors that can be controlling recruitment directly (e.g. food availability, larval growth), may be more laborious to obtain, than the recruitment estimate itself (Irigoien et al., 2009; Zarauz et al., 2008, 2009). Based on a simplified approach, fisheries management has been moving towards the use of environmental relationships using oceanographic data. These are collected routinely, as proxies of recruitment conditions (Bartolino et al., 2008; Borja et al., 2008; De Oliveira et al., 2005). Nevertheless, the problem remains difficult because the mechanisms behind such relationships are often poorly understood; this in turn, makes it difficult to determine the forecast estimation robustness, leading to the failure of some proposed relationships, methods and performance estimations, when new data became available (Myers et al., 1995). Such failures may be related to new controls, which were not considered previously (Myers et al., 1995; Planque and Buffaz, 2008), or to limitations in the available data (Schirripa and Colbert, 2006).

Recruitment forecast is a problem of high uncertainty (Mäntyniemi et al., in press). Machine-learning techniques have been proposed as an appropriate approach with some desirable properties to address such problems (Dreyfus-León and Chen, 2007; Dreyfus-León and Schweigert, 2008; Fernandes et al., 2010, 2013; Uusitalo, 2007). In this study, an update of a previously proposed machine-learning based framework (Fernandes et al., 2010) is applied to several North Atlantic species of commercial interest, which share spawning and nursing environment in the shelf break (Ibaibarriaga et al., 2007; Sagarminaga and Arrizabalaga, 2010). The main properties of this methodology are: (i) forecasts with its

* Corresponding author at: Plymouth Marine Laboratory, PL1 3DH Plymouth, UK.
E-mail address: jfs@pml.ac.uk (J.A. Fernandes).

http://dx.doi.org/10.1016/j.ecoinf.2014.11.004
1574-9541/© 2014 Elsevier B.V. All rights reserved.

uncertainty estimated; (ii) forecasts and scenarios easy to interpret; (iii) recruitment and factors boundaries, that can be interpreted easily; (iv) high stability of selected factors, using a '*leaving one out*' schema; (v) error balanced through all recruitment level; and (vi) robust, as well as honest performance estimation.

Within this context, this work has three aims: to identify factors for forecasting of North Atlantic species that share spawning and nursing area; (ii) to propose a novel model to modify the previous framework in order to produce more accurate probabilistic forecasts; and (iii) to provide a comparison between goodness-of-fit and generalization power, in order to assess the reliability of the final forecasting models. This comparison is necessary since the used methods are non-parametric and might over-fit the data. The three objectives are crucial to produce reliable forecasts that can be used for decision taking in fisheries management of those species that share spawning and nursing area.

2. Methods

2.1. Target species

The species recruitment time series analysed for the North East Atlantic that share the shelf break as spawning and nursing area are summarized below: 1) The *anchovy recruitment mixed time-series (ARM)* is a combination of two anchovy recruitment time-series; the long a*nchovy recruitment index* time-series (ARI; Borja et al., 1996) established from the percentage of age 1 in the landings (40 years) and the *Anchovy Recruitment* (AR; ICES, 2008a; 23 years). The resulting time-series contains 45 years of data (1964–2008). The reason for establishing this combined time-series is that data-mining or machine-learning methods can benefit from the availability of more data. 2) The n*orthern hake recruitment time-series (HR)* covers a period of 29 years of data (1978–2006; ICES, 2008b). 3) *Sardine recruitment time-series (SR)* covers a period of 30 years (1978–2007; ICES, 2008c). 4) The *albacore recruitment time-series (ALR)* covers a period of 56 years (ICCAT, 2007). However, since most of the environmental variables have only data available for the last 39 years, these years have been used to learn the model (1967–2005). 5) The *blue whiting recruitment time-series (BWR)* covers a period 27 years (1981–2007; 2007a). 6) The *northeast mackerel recruitment time-series (MR)* covers a period of 36 years of data (1972–2007; ICES, 2008d). 7) The *western horse mackerel recruitment time-series (HMR)* covers a period of 26 years (1982–2007; ICES, 2008d).

2.2. Variables

The dataset of environmental variables used in this study has been obtained from the 2007 Workshop on 'Long-term Variability in SW Europe' (ICES, 2007); this consists mainly of northern hemisphere atmospheric indexes. In addition, other environmental indexes have been added, such as wind data for the area of the North East Atlantic and temperature anomalies. The annual mean of these variables has been used, except when the index has an associated time period (e.g. Upwelling Index, along the French and Spanish coasts from March to July). Finally, the spawning stock biomass (SSB) of each species has also been considered as a variable candidate for recruitment forecasting. A list of the indexes selected by the methodology applied and their description is provided in Table 1.

2.3. Supervised classification based methodology

The methodology proposed in Fernandes et al. (2010) has been applied, which consists of a sequential pipeline or group of state-of-art supervised classification methods. A high dimensional dataset (hundreds of factors) is provided as input and a model with a trade-off between simplicity and high forecast power is produced by means of strong validation. This final model consists in a *naive Bayes classifier* where a small subset of factors as been selected and the factors as well as the recruitment values are simplified in two or three categories (low, medium, high). The establishment of the boundaries of these recruitment categories can be provided by experts or by the methodology itself.

Table 1
Abbreviation and description of variables that appear through the text.

Variable abbreviation	Variable description
EA	East Atlantic pattern.
AA_Index	Sun geomagnetic activity index.
AMO	Atlantic Multidecadal Oscillation.
Central England temperature	Hadley Centre Central England temperature (HadCET).
CLI1	First PCA component of climatic detrended indices.
CurlSurfaceWind_40N10W	FNMOC Curl of surface wind stress (40°N, 10°W).
CurlSurfaceWind_45N2W	FNMOC Curl of surface wind stress (45°N, 2°W).
CurlSurfaceWind_45N3W	FNMOC Curl of surface wind stress (45°N, 3°W).
EkmanTransportNS_45N2W	FNMOC North–south component of Ekman Transport (45°N, 32°W).
E_W_Wind_45N3W	FNMOC East–west wind (45°N, 3°W).
E_W_WindStress_43N11W	FNMOC East–west wind stress (45°N, 11°W).
EP_NP	Eastern Pacific/North Pacific Pattern.
Global_Tanom	Hadley Centre global SST anomaly data set (HadSST2).
MMF_GSB_ 48.5 N9.5 W	Meridional Momentum Flux at Great Sole Bank.
MMF_PB_ 52.5 N11.5 W	Meridional Momentum Flux at Porcupine.
Natlantic.average	North Atlantic SST average (NOAA ERSST V2 SST).
N_S_Wind_45N2W	FNMOC North–south wind (45°N, 2°W).
N_S_WindStress_45N2W	FNMOC North–south wind stress (45°N, 2°W).
N_S_Wind_45N3W	FNMOC North–south wind (45°N, 3°W).
N_S_WindStress_45N3W	FNMOC North–south wind stress (45°N, 3°W).
POL	Polar/Eurasia Pattern
POLE	Poleward index from geostrophic winds (43°N, 11°W).
SSB	Spawning stock biomass.
SST_4311	Mean sea surface temperature (43°N,11°W; °C).
SSTP	Mean sea surface temperature Portugal (39.5°N, 9.5°W; °C).
SunSpot	Number of sun spots.
TempAnom N	Temperature anomaly for the area 55–60°N, 15–10°W.
UIBs_4502	Upwelling index Basque coast (45°N, 2°W; March-July mean).
Uim_4311	Upwelling index from geostrophic winds (43°N, 11°W).

The methodology is based in supervised classification methods, i.e. methods which consider an objective: in this study the forecasting of three recruitment levels for each species (e.g. Fayyad and Irani's method (1993) discretization method or Hall's CFS multivariate factors subset selection method (2000)). Data re-sampling methods are used during the model building steps in order to ensure more robust (stable) recruitment levels by means of *bootstrapping* (Efron, 1979) and selected factors (reduce spurious links) using *leaving-one out* (Francis, 2006; Mosteller and Tukey, 1968). Finally, after factor discretization and selection a Bayesian network classifier, probabilistic model, is learned such the *naive Bayes classifier* (NBC). In Fernandes et al. (2010) several classification model paradigms where compared without outperforming the NBC for recruitment forecasting of two fish species.

Bayesian networks (BNs) are a modelling framework based on probability theory and graph theory (Buntine, 1991; Jordan, 1998), adequate for domains of high uncertainty such as recruitment forecasting for fisheries management purposes. BNs provide a probability distribution of the different recruitment levels instead of only a forecast of one level or value as the most probable or the forecasted. This additional information of the uncertainty associated to a forecast is crucial for decision making. The *naive Bayes classifier* (NBC; Duda and Hart, 1973; Langley et al., 1992) is a BN model where independence between factors is assumed and the recruitment is the parent of all the factors. These assumptions allow building a model that needs few parameters (more robust with few data) and a competitive performance.

The aim of this work is to extend previous work (Fernandes et al., 2010) to more species that share spawning and nursing area. In