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/

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.

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).

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.

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

addition, *flexible naive Bayes classifier* (FNBC) is used in order to improve estimated probabilities of each possible outcome. Previously proposed NBC is commonly applied by discretization of all the factors. The use of discretized factors has the advantages that there is the property of non-parametric assumptions as well as high comprehensibility of the model. However, in some cases the discretization of variables might lead to same information loss. This information loss can be avoided without losing most of the advantages of a discretized NB replacing it by a FNB, where the recruitment is discrete and predictors continuous.

A '*flexible naive Bayes*' classifier consists of the '*multinomial naive Bayes*' classifier, supported by the '*kernel-based Bayesian network*' paradigm proposed in Pérez et al. (2009), which are based upon a non-parametric density estimation technique, '*kernel density estimation*' (Silverman, 1986). This means that the classifier is built by aggregating a mixture of kernel avoiding any assumption such as normality. The factorisation of the generalized joint distribution represented by a '*flexible naive Bayes*' classifier defined over the factors $(V_1, \ldots, V_k)$ and recruitment (C) is given by:

$$\rho(\mathrm{V}_1, \ldots, V_k, C) = p(c) \prod_{i=1}^{k} f(v_i|\mathrm{c}).$$

where $p(C)$ is the maximum likelihood estimator of recruitment priors, and $f(v_i|c)$ is the estimated conditional density function of the *ith* variable, given the recruitment, using '*kernel density estimation*':

$$f(v_i|\mathrm{c}) = \frac{1}{\mathrm{N_c}} \sum_{j=1}^{N_c} K_{h_c}\left(\frac{v_i - v_i^{(j)}}{\sigma_i}\right)$$

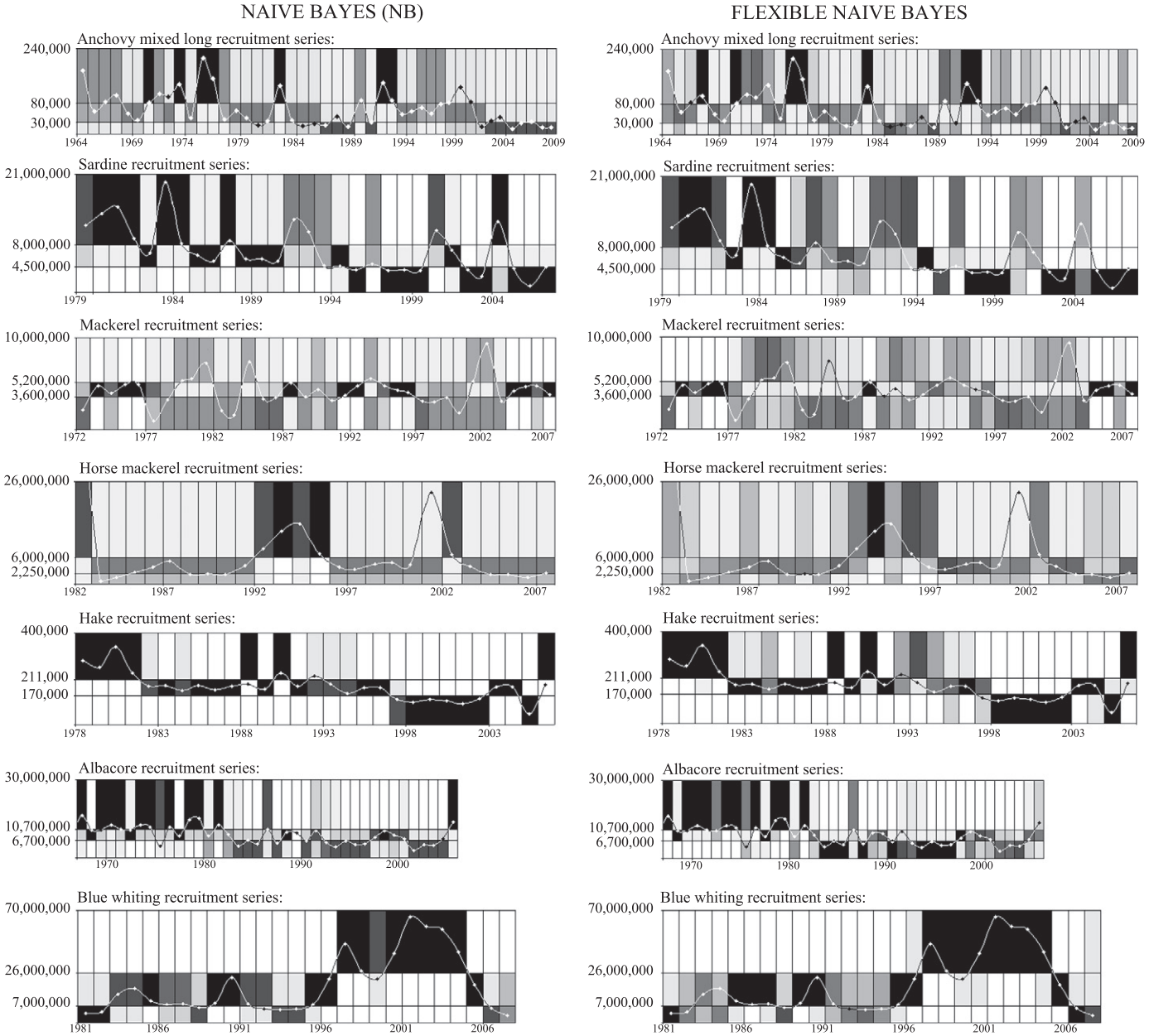

Fig. 1. Goodness of fit for the seven fish species in terms of '*Brier score*'.

Table 2
Fitting using Brier score measure. The classifier is learned with all the data and all recruitment interval likelihoods calculated for all the years.

	Brier score: fitting		Brier score: generalization		Accuracy (%): generalization	
	NBC-pipeline	*FNBC-pipeline*	*NBC-pipeline*	*FNBC-pipeline*	*NBC-pipeline*	*FNBC-pipeline*
Anchovy	0.22	0.18	0.24 ± 0.05	0.11 ± 0.05	46.1 ± 8.9	47 ± 7.9
Hake	0.10	0.04	0.28 ± 0.07	0.28 ± 0.07	56.7 ± 10.2	49.9 ± 5.3
Sardine	0.14	0.10	0.16 ± 0.04	0.28 ± 0.05	70.0 ± 4.7	23.3 ± 5.9
Mackerel	0.20	0.17	0.20 ± 0.05	0.06 ± 0.05	31.3 ± 6.8	35.3 ± 5.7
Horse mackerel	0.21	0.20	0.29 ± 0.05	0.22 ± 0.06	40.9 ± 4.9	44.7 ± 11.2
Albacore	0.12	0.09	0.19 ± 0.04	0.09 ± 0.07	58.1 ± 5.8	34.6 ± 5.1
Blue whiting	0.11	0.09	0.26 ± 0.04	0.17 ± 0.08	51.3 ± 7.6	43.9 ± 8.1

where N_c is the number of cases or years for which $C = c$, $v_i^{(j)}$ in the j^{th} case which take the value $C = c$, h_c is the smoothing degree that it is computed using the normal rule:

$$h_c = \left(\frac{4}{(n+2)N_c}\right)^{\frac{1}{n+4}}$$

and K_h is a Gaussian kernel function given by:

$$K_{h_c}(v_i) = \frac{1}{h_c\sqrt{2\pi}} \exp\left(\frac{v_i^2}{2h_c^2}\right).$$

Independently from this heuristic used to compute the smoothing degree, the '*kernel density estimator*' is a non-parametric estimator, which avoids parametric assumptions. For further details, the reader should consult Pérez et al. (2009).

FNBC cannot be used with missing data. Therefore, the '*supervised missing imputation*' as been applied by means of the method '*Cmean*'. This simple method has proved to be very effective, consists in imputing the mean for continuous variables, or the most repeated for categorical variables (Delavallade and Dang, 2007; Little and Rubin, 2002). In the supervised variant of '*Cmean*', the imputed values are the mean of the values, in those cases that have the same recruitment level. In this work, the previous proposed set of methods (pipeline) in Fernandes et al. (2010) is named '*NBC-Pipeline*' *which* consists on factor discretization, factor selection and a *naive Bayes classifier*. This pipeline is compared with a novel one that consists in missing data imputation, no discretization, factor selection and a *flexible naive Bayes classifier*, which is named '*FNBC-Pipeline*'. We keep the feature selection with leaving-one-out scheme in all the pipeline of methods. Notice that when we validate in the cross-validation we are validating not only the model, but also the pre-processing by doing the train-test split at the beginning of the pipeline of methods.

2.4. Performance estimation

The reliability in a model needs of its performance assessment. Therefore, to know the fit of the model to the data is needed ('*goodness of fit*' or data descriptive power). However, that a model fits well the data does not mean that the model has a good predictive power ('*generalization*'). Therefore, both must be tested.

The '*goodness of fit*' is achieved by learning a classifier with all the data and testing with each year (Fig. 1 and Table 2). The '*generalization*' has been evaluated using '*10-times repeated 5-fold cross-validation*' scheme. '*Cross-validation*' consists of performing data partition, in *k* parts or folds, using one fold for evaluation and the remainder for learning a model *k* times, being the estimated performance the mean of the *k* models. '*Repeated cross-validation*' consists of repeating the '*cross-validation*' scheme several times. Finally, instead of split in folds just before learning the classifier, the split is performed before the whole pipeline is applied, for an honest validation (Francis, 2006; Reunanen, 2003; Statnikov et al., 2005). This is necessary because all the pre-processing steps are supervised, i.e. they use the recruitment values to perform their optimisation task.

'*Accuracy*', '*Brier score*' and '*true positive rate*' performance measures have been used to assess the generalization power. '*Accuracy*' measures model performance without considering the estimated probability (win/loss measure), whereas the Brier score considers these estimated probabilities to each possible outcome (Brier, 1950; van der Gaag and Renooij, 2001; Yeung et al., 2005). Finally, '*true positive rate*' measures the error distribution between different recruitment levels. '*Accuracy*' and '*true positive rate*' are measured between 0% and 100%, with the objective of the highest values indicating better results; whereas '*Brier score*' lies between 0 and 1, with the lowest values indicating the best results. The '*Brier score*' has been used for measuring both, '*generalization*' and '*goodness-of-fit*':

$$\frac{1}{2}\sum_{i=1}^{m}(p_i - Y_i)^2$$

where *m* is the number of recruitment intervals, and *pi* is the predicted probability for each recruitment value. The *Yi* value is 1, if *i* is the observed value of the recruitment; and 0 otherwise.

All of the above steps have been implemented using several established API machine-learning software: Weka (Witten and Frank, 2005); and Elvira (www.ia.uned.es/~elvira/index-en.html) with an

Table 3
'*Accuracy*' comparison between the pipeline without '*missing imputation*' as well as with a '*multinomial naive Bayes*' classifier, with '*missing imputation*' as well as with a '*multinomial naive Bayes*' classifier and with '*missing imputation*' as well as replacing the '*multinomial naive Bayes*' by a '*flexible naive Bayes*'.

Accuracy (%) generalization	NB-pipeline	MIS + NB-pipeline	MIS + FNB-pipeline
Anchovy	46.1 ± 8.9	47.8 ± 7.7	47 ± 7.9
Hake	56.7 ± 10.2	53.7 ± 11.1	49.9 ± 5.3
Sardine	70.0 ± 4.7	70.3 ± 5.1	23.3 ± 5.9
Mackerel	31.3 ± 6.8	33.1 ± 8.6	35.3 ± 5.7
Horse mackerel	40.9 ± 4.9	42.9 ± 7.8	44.7 ± 11.2
Albacore	58.1 ± 5.8	61.4 ± 6.9	34.6 ± 5.1
Blue whiting	51.3 ± 7.6	52.5 ± 7.8	43.9 ± 8.1

Table 4
Brier score comparison between the pipeline without '*missing imputation*' as well as with a '*multinomial naive Bayes*' classifier, with '*missing imputation*' as well as with a '*multinomial naive Bayes*' classifier and with '*missing imputation*' as well as replacing the '*multinomial naive Bayes*' by a '*flexible naive Bayes*'.

Brier score generalization	NB-pipeline	MIS + NB-pipeline	MIS + FNB-pipeline
Anchovy	0.24 ± 0.05	0.22 ± 0.04	0.11 ± 0.05
Hake	0.28 ± 0.07	0.28 ± 0.07	0.28 ± 0.07
Sardine	0.16 ± 0.04	0.15 ± 0.04	0.28 ± 0.05
Mackerel	0.20 ± 0.05	0.17 ± 0.07	0.06 ± 0.05
Horse mackerel	0.29 ± 0.05	0.28 ± 0.07	0.22 ± 0.06
Albacore	0.19 ± 0.04	0.25 ± 0.06	0.09 ± 0.07
Blue whiting	0.26 ± 0.04	0.24 ± 0.07	0.17 ± 0.08

adaptation to incorporate '*flexible Bayesian network*' classifiers by Pérez et al. (2009). Reproducibility is ensured by a Java programming language implementation of all the methodology, available from the ISG group webpage (www.sc.ehu.es/ccwbayes/members/jafernandes/ or at www.azti.es). The kernel estimators are provided as a library. However, these are documented in detail in Pérez et al. (2009) and their implementation is available on request from this author.

3. Results

3.1. Selected factors

3.1.1. Pipeline comparison

The missing imputation can also be applied to the '*NBC-Pipeline*'; however, no significant improvement was observed. This result was expected since NBC can be learned with missing data and there was no factor with high levels of missing values.

Both classifiers, NB and FNB classifiers, show good-fit for most of the considered species (Fig. 1). The '*MIS + FNB-Pipeline*' produces the best fitting for the seven species (Table 2). The most interesting property of this fitting for fisheries management is that the lowest and highest recruitment levels are associated with high probability estimations (Fig. 1). In years where recruitment was close to the boundaries between both recruitment levels, the probabilities are better distributed between the two levels using FNB. The FNB shows also higher fitting and generalization power if estimated probabilities are considered using the Brier score measure (Tables 2 and 4). However, a good-fit does not guarantee a good generalization power. While, mackerel and horse mackerel show a good-fit (Table 2), they show the worst generalization power in '*accuracy*' terms for the '*NB-Pipeline*' (Tables 3, 4 and 5). In terms of '*Brier score*', mackerel shows better results comparable with other species; however, horse mackerel does not, except using the FNB. Hake, that shows higher '*accuracy*' than mackerel and horse mackerel, shows the worst results in terms of '*Brier score*' in the three pipelines.

The different pipeline comparisons reveal that there is no significant improvement (corrected paired *t*-test $p < 0.10$; Nadeau and Bengio, 2003) in '*accuracy*' (Table 3), between '*NB-Pipeline*' and '*MIS-NB-Pipeline*'. Similarly, the use of the '*MIS-FNB-Pipeline* does not show significant improvements in '*accuracy*' terms and it decreases in the case of sardine ($p < 0.05$). Although, the differences are not significant, the Brier score is usually inferior in '*MIS-NB-Pipeline*' than in '*NB-Pipeline*' (Table 4). Finally, the use of '*flexible naive Bayes*' classifier reduces the '*Brier score*' for all species (Table 5) except in sardine ($p < 0.05$); however, the differences for most of the species remain not significant. This superiority in most species of '*Brier score*' using FNB ($p < 0.05$) and the tie in '*accuracy*' ($p < 0.05$) is observed using a statistical test to compare multiple algorithms over multiple datasets (García and Herrera, 2008; Fig. 2).

A low stability of selected variables produce large variable sets that can be effectively reduced using the '*Markov blanket*' (Table 5), with minor variation of recruitment forecast estimates. In Silverman (1986) the limitation to three variables is recommended for data size lower than 223 samples (years). Fig. 3 presents the effect of the different selected variables on the recruitment of each species. In general, recruitment appears to be associated to transport or temperature parameters.

Table 5
True positive rate comparison between the pipeline without '*missing imputation*' as well as with a '*multinomial naive Bayes*' classifier, with '*missing imputation*' as well as with a '*multinomial naive Bayes*' classifier and with '*missing imputation*' as well as replacing the '*multinomial naive Bayes*' by a '*flexible naive Bayes*'.

TP rate (%) generalization	NB-pipeline	MIS + NB-pipeline	MIS + FNB-pipeline
Anchovy	39.9, 40.1, 45.0	54.0, 31.1, 45.3	64.0, 22.0, 12.0
Hake	44.7, 52.3, 49.0	48.2, 41.2, 52.1	14.0, 66.0, 18.0
Sardine	65.0, 55.0, 68.0	82.8, 64.3, 55.3	10.0, 12.0, 60.0
Mackerel	27.1, 17.9, 25.2	54.2, 54.2, 11.0	42.0, 54.0, 00.0
Horse mackerel	19.9, 32.3, 11.5	26.8, 33.3, 14.5	32.0, 38.0, 16.0
Albacore	64.1, 35.4, 70.3	65.2, 39.7, 76.3	72.0, 08.0, 19.0
Blue whiting	34.7, 56.7, 43.8	36.8, 46.2, 54.4	48.0, 24.0, 16.0

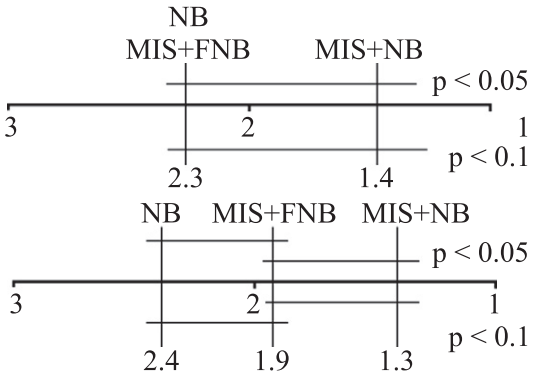

Fig. 2. Rank comparison between the three pipelines. The horizontal line joins algorithm where there is not a significant difference at the specified level ($p < 0.05$ or $p < 0.1$).

4. Discussion

The main contribution of this work is the application of the methodology developed in Fernandes et al. (2010), to a broad set of species using a global set of variables. The forecast estimates of each species can be improved by applying more specific knowledge (more specific environmental data), to each species. However, the results show that, even using a global approach, useful information can be obtained using machine learning techniques applied to the recruitment forecasting problem. The data is time-series, but we do use them as independent observations. This is due to the nature of the problem of fish recruitment of pelagic species that varies highly between years depending on environmental conditions. Otherwise we would have used classical stock-recruitment functions or the temporal Dynamic Bayesian Networks. However, those relationships do not hold for pelagic species which are species of short life and which reproduction strategy is to spawn a large number of eggs.

It has been argued to what extent the use of machine-learning, or similar techniques, can be applied in practice to recruitment forecasting (Francis, 2006). This issue is particularly relevant for species where the '*accuracy*' is low. However, such an argument does not consider the information provided by the estimated probabilities. Indeed, it is in the provided scenarios probabilities, where the modelling approach is superior to using random, average or recent recruitment values, as reflected in the Brier score. This is similar to the '*accuracy paradox*' (Fernandes et al., 2010), i.e. if the most frequent recruitment level is always predicted and the rest of levels have low frequency, the global '*accuracy*' would be high, but without contributing additional information. As such, different performance measures should be used to evaluate the models. To be useful a model must comply with certain properties, e.g. error balanced between different recruitment levels (true positive rate). Another possibility is the use of the previous year or recent recruitment values. On the one hand, this approach might provide on average a good '*accuracy*'. However, it fails to detect changes in the recruitment level. These are the most important to be reliably forecasted because of their economic and biological implications.

The use of recruitment '*a posteriori*' probabilities instead of win/losses for a specific scenario is an approach that fits well with the Bayesian models being incorporated into fisheries management (e.g. Ibaibarriaga et al., 2008). Although the environment can influence recruitment, other factors affect it. Therefore, the scenarios' probability can provide an '*a priori*' of whether or not the environmental conditions are favourable for each recruitment level. It is important to realize that the forecast estimates have a double reading, or interpretation. As an

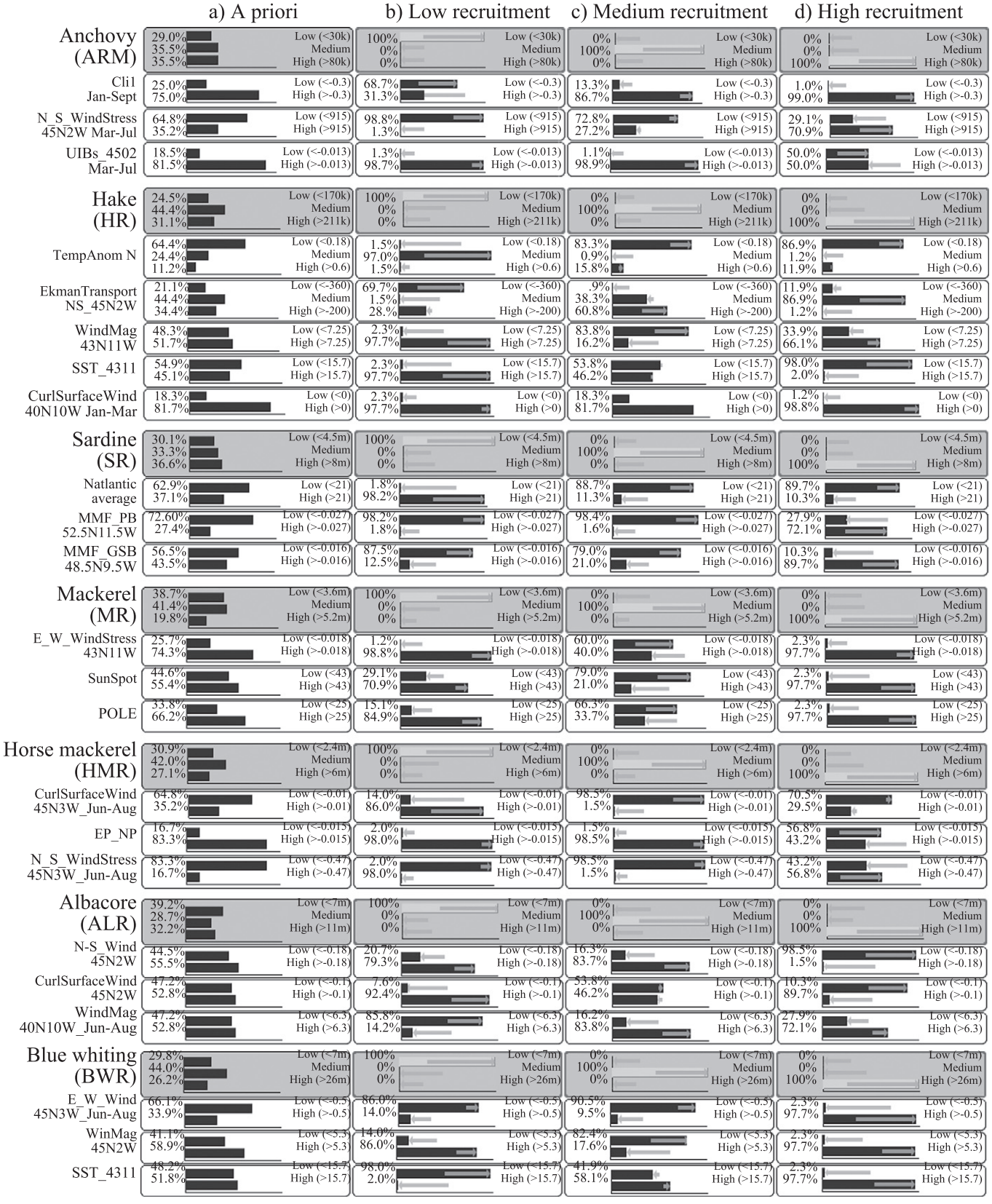

Fig. 3. Recruitment scenarios using the classifier learned with the '*NB-pipeline*' for seven species of North East Atlantic.

example, in a forecast of 25% low recruitment, 60% medium and 15% high, the most likely scenario is that the recruitment for that year is medium. However, there is another interpretation of this model output; in 60% of years in which there were the same conditions, the recruitment was medium, i.e. these conditions are more favourable for a medium recruitment.

Another contribution, in comparison with the previous study undertaken in Fernandes et al. (2010), is the use of '*flexible classifiers*', which

Table 6
Fitting values comparison between NB-pipeline and MIS + FNB-pipeline for anchovy recruitment time-series. Bold represents the observed value in the data.

	ARM values		*NBC-pipeline*			*FNBC-pipeline*		
Year	Non-discretized	Discretized	Low (%)	Medium (%)	High (%)	Low (%)	Medium (%)	High (%)
1964	176934	More_80000	12	41	**47**	6	53	**41**
1965	58016	30000–80000	12	**41**	47	21	**72**	7
1966	85953	More_80000	12	41	**47**	7	67	**26**
1967	104729	More_80000	1	40	**59**	5	5	**90**
…	…	…	…	…	…	…	…	…

permit having more precise forecast estimates, even if there is not a significant increment in '*accuracy*', except for some of the species. The use of a '*naive Bayes*' with discretized variables, returns always the same forecast estimates as long as the influencing variables remain in the same interval (e.g. year 1966 and 1967; Table 6). It does not consider the distance of these environmental variable values to the boundaries between different levels. In contrast, a '*flexible naive Bayes*' can consider this. It returns '*smoother*' estimations that can be more precise, as reflected in the Brier score. In particular, the FNB provides a better distribution of probabilities between two recruitment levels for those years where the recruitment is close to a recruitment boundary (e.g. year 2006; Table 6). In addition, there are classifiers that cannot be learned with missing data, as is the case of the used implementation of '*flexible naive Bayes*'. The '*missing imputation*' can be performed without negative effects in performance and with some performance improvements in some species.

Table 7
Comparison between the set of variables that is returned by the methodology before and after applying the '*Markov blanket*' property to reduce the number of selected variables. Stability (stab.) has been calculated, as the total number of times the two most repeated subset of variables has been selected in a '*leaving one out scheme*' divided by the number of available data years.

Variables	Before '*Markov blanket*'	SUS
Anchovy	CLI1_Jan–Sept	0.296
(0.33 stab.)	N_S_WindStress_45N2W_Mar–Jul	0.239
	UIBs_4502_ Mar–Jul	0.234
Hake	TempAnom N	0.577
(0.17 stab.)	EkmanTransportNS_45N2W	0.473
	WindMag_40N10W	0.406
	SST_4311	0.362
	WindMag_43N11W	0.344
	CurlSurfaceWind_40N10W_Jan–Mar	0.323
	AA_Index	0.301
Sardine	Natlantic.average	0.499
(0.5 stab.)	Central England temperature	0.446
	AMO	0.442
	MMF_PB_ 52.5 N11.5 W	0.436
	Uim_4311	0.388
	MMF_GSB_ 48.5 N9.5 W	0.338
Mackerel	E_W_WindStress_43N11W	0.35
(0.36 stab.)	SunSpot	0.283
	N_S_wind_45N2W	0.265
	POLE	0.251
Horse Mackerel	CurlSurfaceWind_45N3W_Jun–Aug	0.427
(0.5 stab.)	EP_NP	0.329
	N_S_WindStress_45N3W_Jun–Aug	0.329
Albacore	N_S_Wind_45N2W	0.15
(0.45 stab.)	SSTP	0.135
	N_S_Wind_45N3W	0.135
	CurlSurfaceWind_45N2W	0.128
	WindMag_40N10W_Jun–Aug	0.123
Blue whiting	E_W_Wind_45N3W_Jun–Aug	0.466
(0.14 stab.)	TempAnom N	0.447
	Global_Tanom	0.422
	WindMag_45N2W	0.417
	CurlSurfaceWind_45N3W	0.354
	SST_4311	0.301

In most cases, environmental factors influencing recruitment were related to temperature and transport (Table 7), whilst the performance was similar between the different species. However, the case of horse mackerel species is a good example of the limitations and precautions to be taken when using a data-mining approach. The time series of horse mackerel recruitment is relatively long, but there are only three peaks of high recruitment; most of the recruitments are low or medium. It would be enough to predict always a medium recruitment to obtain a high accuracy; however, this would miss the capacity to predict any change. Furthermore, it is not sufficient to have a long time-series to predict high recruitment events, if those events occur very rarely during the observed period. To be useful, a data-mining approach needs the distribution of the events to be learnt to be relatively equally distributed. Otherwise, high accuracy can be achieved but with little predictive capacity. This is why the use of robust approaches is needed for its application in fisheries management.

Acknowledgements

The research of Jose A. Fernandes and Nerea Goikoetxea is supported by a Doctoral Fellowship from the Fundación Centros Tecnológicos Iñaki Goenaga. This study has been supported by the following projects: Ecoanchoa (funded by the Department of Agriculture, Fisheries and Food of the Basque Country Government); the Saiotek and Research Groups 2007–2012 (IT-242-07) programs (Basque Government), TIN2008-06815-C02-01 (Spanish Ministry of Education and Science); COMBIOMED network in computational biomedicine (Carlos III Health Institute); the EU project UNCOVER; the EU FACT; and the EU VII Framework project MEECE (MEECE No 212085). Professor Michael Collins (SOES, University of Southampton, UK and AZTI-Tecnalia, Spain) is acknowledged for his comments on the manuscript and help with the English language. This is contribution 695 from the Marine Research Division (AZTI-Tecnalia).

References

Bartolino, V., Colloca, F., Sartor, P., Ardizzone, G., 2008. Modelling recruitment dynamics of hake, *Merluccius merluccius*, in the central Mediterranean in relation to key environmental variables. Fish. Resh. 92, 277–288.
Borja, A., Uriarte, A., Motos, L., Valencia, V., 1996. Relationship between anchovy (*Engraulis encrasicholus*) recruitment and the environment in the Bay of Biscay. Sci. Mar. 60 (Suppl. 2), 179–192.
Borja, A., Fontán, A., Saénz, J., Valencia, V., 2008. Climate, oceanography, and recruitment: the Bay of Biscay anchovy paradigm. Fish. Oceanogr. 17 (6), 477–493.
Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78 (1), 1–3.
Buntine, W., 1991. Theory refinement on Bayesian networks. Proceedings of the Seventh Conference on Uncertainty. Artificial Intelligence 91, pp. 52–60.
Cushing, D., 1971. The dependence of recruitment on parent stock in different groups of fishes. ICES J. Mar. Sci. 33 (3), 340.
De Oliveira, J.A., Uriarte, A., Roel, B., 2005. Potential improvements in the management of Bay of Biscay anchovy by incorporating environmental indices as recruitment predictors. Fish. Res. 75 (1–3), 2–14.
Delavallade, T., Dang, H.D., 2007. Using entropy to impute missing data in a classification task. Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE'07, London, UK, pp. 577–582.
Dreyfus-León, M., Chen, D.G., 2007. Recruitment prediction with genetic algorithms with application to the Pacific Herring fishery. Ecol. Model. 203 (1–2), 141–146.

Dreyfus-León, M., Schweigert, J., 2008. Recruitment prediction for Pacific herring (*Clupea pallasi*) on the west coast of Vancouver Island, Canada. Ecol. Inf. 3 (2), 202–206.
Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. John Willey & Sons, New York, NY, USA.
Efron, B., 1979. Bootstrap methods: another look at the jacknife. Ann. Stat. 7 (1), 1–26.
Fayyad, U.M., Irani, K.B., 1993. Multi-interval discretization of continuous valued attributes for classification learning. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pp. 1022–1027.
Fernandes, J.A., Irigoien, X., Goikoetxea, N., Lozano, J.A., Inza, I., Pérez, A., Bode, A., 2010. Fish recruitment prediction, using robust supervised classification methods. Ecol. Model. 221 (2), 338–352.
Fernandes, J.A., Lozano, J.A., Inza, I., Irigoien, X., Pérez, A., Rodríguez, J.D., 2013. Supervised pre-processing approaches in multiple class variables classification for fish recruitment forecasting. Environ. Model Softw. 40, 245–254.
Francis, R.I.C., 2006. Measuring the strength of environment-recruitment relationships: the importance of including predictor screening within cross-validations. ICES J. Mar. Sci. 63 (4), 594–599.
García, S., Herrera, F., 2008. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. JMRL 9, 2677–2694.
Ibaibarriaga, L., Irigoien, X., Santos, M., Motos, L., Fives, J.M., Franco, C., Lago de Lanzós, A., Acevedo, S., Bernal, M., Bez, N., Eltink, G., Farinha, A., Hammer, C., Iversen, S.A., Miligan, S.P., Reid, D.G., 2007. Egg and larval distributions of seven species in northeast Atlantic waters. Fish. Oceanogr. 16 (3), 284–293.
Ibaibarriaga, L., Fernandez, C., Uriarte, A., Roel, B., 2008. A two-stage biomass dynamic model for Bay of Biscay anchovy: a Bayesian approach. ICES J. Mar. Sci. 65 (2), 191.
ICCAT, 2007. Report of the 2007 ICCAT albacore tuna stock assessment session (Madrid, July 5 to 12, 2007). Collect. Vol. Sci. Pap. ICCAT 62, pp. 697–815.
ICES, 2007. Report of the ICES/GLOBEC Workshop on Long-term Variability in SW Europe (WKLTVSWE), February 13–16, Lisbon, Portugal (ICES CM 2007/LRC: 02, 111).
ICES, 2008a. Report of the Working Group on the Anchovy, ICES Headquarters, June 13–16 (ICES CM 2008/ACOM:04).
ICES, 2008b. Report of the Working Group on the Assessment of Southern Shelf Stocks of Hake, Monk and Megrim (WGHMM), April 30–6 May 6, ICES Headquarters, Copenhagen (ICES CM 2008/ACOM:07, 613).
ICES, 2008c. Report of the ICES Advisory Committee 2008. Books 1–10.
ICES, 2008d. Report of the Working Group on Widely Distributed Stocks (WGWIDE), September 2–11, ICES Headquarters, Copenhagen (ICES CM 2008/ACOM:13, 231).
Irigoien, X., Fernandes, J.A., Grosjean, P., Denis, K., Albaina, A., Santos, M., 2009. Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment. J. Plankton Res. 31 (1), 1–17.
Jordan, M.I., 1998. Learning in Graphical Models. Kluwer Academic Pub., Norwell, MA, USA.
Langley, P., Iba, W., Thompson, K., 1992. An analysis of Bayesian classifiers. Proceedings AAAI-94Seattle, WA. AAAI Press and MIT Press (223–228 pp.).
Little, R., Rubin, D., 2002. Statistical Analysis with Missing Data. 2nd edition. John Wiley and Sons.
Mäntyniemi, S., Kuikka, S., Rahikainen, M., Kell, L.T., Kaitala, V., 2014. The value of information in fisheries management: North Sea herring as an example. ICES J. Mar. Sci. http://dx.doi.org/10.1093/icesjms/fsp206 (in press).
Mosteller, F., Tukey, J.F., 1968. Data analysis, including statistics. In: Lindzey, G., Aronson, E. (Eds.), Handbook of Social Psychology vol. II. Addison-Wesley, Reading, MA, p. 588.
Myers, R.A., Bridson, J., Borrowman, N.J., 1995. Summary of worldwide spawner and recruitment data. Can. Tech. Rep. Fish. Aquat. Sci. 2024.
Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Mach. Learn. 52 (3), 239–281.
Pérez, A., Larrañaga, P., Inza, I., 2009. Bayesian classifiers based on kernel density estimation: flexible classifiers. Int. J. Approx. Reason. 50 (2), 341–362.
Planque, B., Buffaz, L., 2008. Quantile regression models for fish recruitment–environment relationships: four case studies. Mar. Ecol. Prog. Ser. 357, 213–223.
Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. JMRL 3, 1371–1382.
Ricker, W.E., 1954. Stock and recruitment. J. Fish. Res. Board Can. 11, 559–623.
Rothschild, B., 2000. Fish stocks and recruitment: the past thirty years. ICES J. Mar. Sci. 57 (2), 191.
Sagarminaga, Y., Arrizabalaga, H., 2010. Spatio-temporal distribution of albacore (Thunnus alalunga) catches in the northeastern Atlantic: relationship with the thermal environment. Fish. Oceanogr. 19 (2), 121–134.
Schirripa, M.J., Colbert, J.J., 2006. Interannual changes in sablefish (*Anoplopoma fimbia*) recruitment in relation to oceanographic conditions within the California Current System. Fish. Oceanogr. 15 (1), 25–36.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. School of Mathematics. University of Bath, UK.
Statnikov, Q., Aliferis, C.F., Tsamardinos, I., Hadinn, D., Levy, S., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21 (5), 631–643.
Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Model. 203, 312–318.
Van der Gaag, L.C., Renooij, S., 2001. Evaluation scores for probabilistic networks. Proceedings of the 13th Belgium-Netherlands Conference on Artificial Intelligence, Amsterdam. Universiteit van Amsterdam, The Netherlands, pp. 109–116.
Witten, I.H., Frank, E., 2005. Data Mining—Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco.
Yeung, K.Y., Bumgarner, R.E., Raftery, A.E., 2005. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 21 (10), 2394–2402.
Zarauz, L., Irigoien, X., Fernandes, J.A., 2008. Modelling the influence of abiotic and biotic factors on plankton distribution in the Bay of Biscay, during three consecutive years (2004–06). J. Plankton Res. 30 (8), 857–872.
Zarauz, L., Irigoien, X., Fernandes, J.A., 2009. Changes in plankton size structure and composition, during the generation of a phytoplankton bloom, in the central Cantabrian sea. J. Plankton Res. 31 (2), 193–207.