Hydrobiologia (2011) 670:241–266
DOI 10.1007/s10750-011-0673-9

Prediction of marine species distribution from presence–absence acoustic data: comparing the fitting efficiency and the predictive capacity of conventional and novel distribution models

A. Palialexis · S. Georgakarakos · I. Karakassis · K. Lika · V. D. Valavanis

Published online: 25 March 2011

Abstract The accurate representation of species distribution derived from sampled data is essential for management purposes and to underpin population modelling. Additionally, the prediction of species distribution for an expanded area, beyond the sampling area can reduce sampling costs. Here, several well-established and recently developed habitat modelling techniques are investigated in order to identify the most suitable approach to use with presence–absence acoustic data. The fitting efficiency of the modelling techniques are initially tested on the training dataset while their predictive capacity is evaluated using a verification set. For the comparison among models, Receiver Operating Characteristics (ROC), Kappa statistics, correlation and confusion matrices are used. Boosted Regression Trees (BRT) and Associative Neural Networks (ASNN), which are both within the machine learning category, outperformed the other modelling approaches tested.

Keywords Species distribution models · Species distribution predictions · Habitat modelling · Models comparison · Geostatistics · Spatial autocorrelation

Guest editors: Graham J. Pierce, Vasilis D. Valavanis, M. Begoña Santos & Julio M. Portela / Marine Ecosystems and Sustainability

A. Palialexis (✉) · I. Karakassis · K. Lika
Department of Biology, University of Crete, Vassilika Vouton, P.O. Box 2208, 71 409 Heraklion, Crete, Greece
e-mail: andreaspal@her.hcmr.gr

S. Georgakarakos
Department of Marine Sciences, University of the Aegean, University Hill, 81 100 Mytilini, Lesvos, Greece

A. Palialexis · V. D. Valavanis
Marine GIS Lab, Hellenic Centre for Marine Research, P.O. Box 2214, 71 003 Heraklion, Crete, Greece

Introduction

Knowledge of species ecological and geographical distribution is essential for conservation planning and forecasting (Ferrier et al., 2002) as well as for assessing evolutionary determinants of spatial patterns of biodiversity (Graham et al., 2006). Several techniques have been developed for the identification of species distribution using sampling data. These may be categorized as species distribution models (SDM), which are statistical models that relate surveyed data on species distribution with the associated environmental and geographical characteristics of the surveyed locations (Elith & Leathwick, 2009). In the literature, SDMs are variously described as resource selection functions (RSFs), habitat models and ecological niche models (see Elith & Leathwick, 2009). Approaches based on SDMs have only relatively recently been applied to marine species and several novel modelling methods have been proposed (Leathwick et al., 2005, 2006a; Phillips et al., 2006; Palialexis et al., this issue). SDMs have

been also used to study relationships between environmental variables and species presence (Amara et al., 2004; Giannoulaki et al., 2008; Lefkaditou et al., 2008; Martin et al., 2008), identifying species essential habitats (Planque et al., 2007) and forecasting how species distribution may be affected by climate changes (Siapatis et al., 2008). Easy access to satellite data which cover extended geographical areas is one reason for the increasingly wide use of SDMs. Presence–absence data derived from several sampling strategies are commonly used with SDMs. Zaniewski et al. (2002) argued that modelling based on presence–absence data is more likely to reflect the present natural distribution of a species, i.e. the realized niche, whereas presence-only methods are more likely to predict potential distributions, more closely resembling the fundamental niche.

The evolution of computer science and statistics is reflected in the novel methods proposed to model species distribution. Presence-only models [e.g. Bioclim, Envelope Score, Ecological Niche Factor Analysis (ENFA)] were initially applied to terrestrial species, taking advantage of data from natural history museums (Ready et al., 2010). Environmental envelope-based models are related to niche theory, where a suitable environmental 'envelope' that favours species occurrence is estimated. Compared to presence-only models, presence–absence approaches to SDM perform more efficiently in terms of prediction (Brotons et al., 2004) since they exploit the additional information about unsuitable environmental conditions for species occurrence. Regression models, such as Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) are widely used to model presence–absence data (Olivier & Wotherspoon, 2005; Leathwick et al., 2006b). Recently, several modelling techniques were developed utilizing the evolution of methods in computer science, like Boosted Regression Trees (BRT; Leathwick et al., 2006a) and Associative Neural Networks (ASNN; Tetko, 2002a, b) combining different algorithms in order to optimize the predictive capacity of the models. Additionally, the most widely used models, like GAMs, have been further developed (e.g. BRUTO, Hastie & Tibshirani, 1990) or modified (e.g. MARS, Leathwick et al., 2006a) to meet additional requirements identified from experience with model building.

Presence–absence models are generally easier to develop since the training data have a binomial distribution while abundance models require more complicated distributions (e.g. Poisson, Gaussian) and thus, several assumptions are necessary. The validation process and error assessment for presence–absence models is correspondingly more straightforward than is the case for abundance models. Several methods have been developed to assess the quality of model predictions (Fielding & Bell, 1997; Boyce et al., 2002). Receiver Operating Characteristics (ROC), Kappa statistics and confusion matrices (Fielding & Bell, 1997) combined with omission and commission errors are widely used to estimate model performance and to compare different methods (Elith et al., 2006).

Modelling the distribution of marine species is now common component of scientific research projects and applied management. The generality of the developed models is essential for accurate predictions of species distribution over extended spatial and temporal scales. On the other hand, precision and accuracy are essential for realistic representation of species distribution and essential fish habitat identification. There is a plethora of modelling techniques suitable for fisheries and acoustic data. However, in the Mediterranean Sea, most of the studies predicting marine species distribution have utilized GAMs (Giannoulaki et al., 2008; Martin et al., 2008; Siapatis et al., 2008), Maximum Entropy Models (MAXENT) (Lefkaditou et al., 2008) and Discriminant Function Analysis (DFA) (Tsagarakis et al., 2008).

The SDMs literature is rapidly expanding, reflecting the rapid evolution of SDMs and their contribution to ecological studies. During the last decade several reviews on SDMs were published (e.g. Guisan & Zimmermann, 2000; Redfern et al., 2006; Richards et al., 2007; Schröder, 2008; Valavanis et al., 2008; Elith & Leathwick, 2009). Other studies addressing essential issues of the development of SDMs include topics such as methods of assessment of prediction errors (Fielding & Bell, 1997; Boyce et al., 2002), effects of spatial autocorrelation in SDMs (Dormann et al., 2007), SDMs and ecological theory (Guisan & Thuiller, 2005), new approaches to SDMs (Leathwick et al., 2005, 2006a) and extended SDMs comparisons to identify their efficiency and applicability for use with several data types (Caruana & Niculescu-Mizil, 2006; Elith et al., 2006; Leathwick et al., 2006a; MacLeod et al., 2008; Palialexis et al., 2009; Aertsen et al., 2010).

This study contributes to the latter two issues by comparing several well-established and recently developed techniques and by introducing the use of ASNN. We compare SDMs that are developed under different statistical principles (machine learning, regression models and others) and represent the most commonly used techniques in order to develop advice for the selection of suitable modelling approaches based on presence–absence acoustic data. The performance of models in terms of goodness of fit to a training dataset is initially tested in order to identify the ability of each approach to accurately represent species distribution. Second, a verification dataset is used to evaluate the predictive performance of each method and to contribute to a better understanding of SDM performance with an independent dataset. Finally, SDMs were applied to high resolution predictors in order to generate distribution maps for small pelagic species. Several criteria were used in order to compare the efficiency of SDMs.

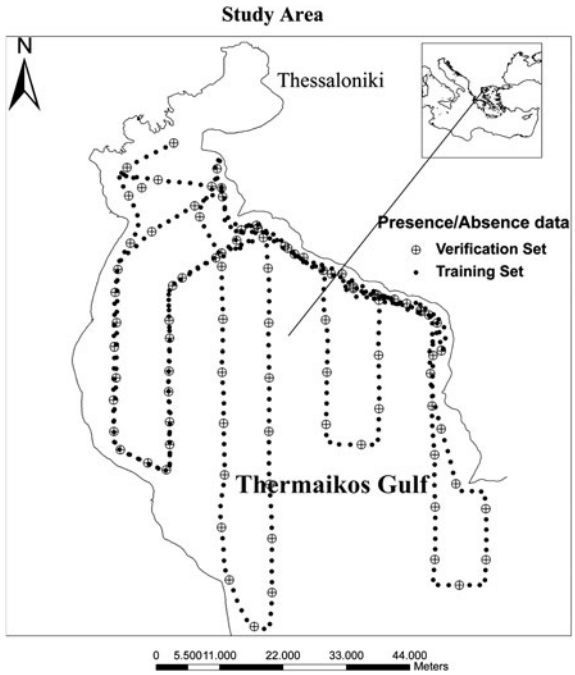

Fig. 1 Study area and sampling transects. *Black dots* represent the training dataset; *circles* represent the verification dataset

Materials and methods

Study area and data

The study area (Fig. 1) is Thermaikos Gulf in the North Aegean, Northeastern Mediterranean Sea. Thermaikos Gulf is a semi-enclosed basin, relatively productive, because of the influence of four major rivers. As a result, bottom relief is smooth due to the continuous sediment input. Thermaikos Gulf forms a wide continental shelf, which smoothly extends to the south into the 1,400 m deep Sporades Basin. Water mass circulation is predominantly cyclonic (Poulos et al., 2000). Aegean water masses enter the gulf from deeper layers along the eastern coast and move counterclockwise towards the gulf of Thessaloniki. This circulation produces a gyre in the area, which is obvious in satellite imagery and affects the life-history of pelagic marine species consisting an identified recruitment habitat (Somarakis et al., 2002). Riverine waters usually move to the south along the western coast forming nutrient-rich water masses.

Acoustic data were collected during April–May 1998 in Thermaikos Gulf using a calibrated 38 kHz SIMRAD EK 500/BI 500 system (Bodholt et al., 1989). The echograms were scrutinized, allocating the nautical area-scattering coefficient (s_A, m^2 n mi^{-2}, NASC, MacLennan et al., 2002) to the target pelagic species. The integration values, with a horizontal resolution of 1 nautical mile, have been transformed to presence–absence data (Fig. 1). Acoustic data have been not converted to biomass in order to avoid the propagation of uncertainty in species composition and length distribution from the trawl sampling in the response variable (Walline, 2007). Species identification based on biological sampling as well as concurrent catch data indicated that the majority of the target species were *Sardina pilchardus* (~55%), *Engraulis encrasicolus* (~25%) and *Trachurus* spp. (<10%). Thus, the SDMs will essentially depict the distribution of sardine and anchovy in the study area. Life-history information on these species was used to inform several parts of this study, such as variable selection, explanation of the distribution maps and comparison of the SDMs with other related studies.

The acoustic dataset has been divided into two parts. The first one is the training set, including 80% of the initial data (black dots in Fig. 1). The second one is used as the verification set (20% of the sampling data—circles with crosses in Fig. 1) in order to justify the predictive capacity of the SDMs.

Table 1 Data used in models and their sources

Data variable	Abbreviation	Data type/sensor	Archive source
Acoustic data	s_A	Total pelagic NASC (nautical area-scattering coefficient), ESDU = 1 n mi	SIMRAD EK500/BI500 system on April/May 1998 in Thermaikos Gulf
Sea surface temperature	SST	Grid/Aqua MODIS	German Aerospace Agency (DLR)
Chlorophyll-*a* concentration	CHL	Grid/Aqua MODIS	Distributed Active Archive Center (NASA)
Photosynthetically available radiation	PAR	Grid/SeaWiFS	Distributed Active Archive Center (NASA)
Sea level anomaly	SLA	Grid/Merged Jason-1, Envisat, ERS-2, GFO, T/P	AVISO
Bathymetry	DEP	Grid/Processed ERS-1, Geostat and historical depth soundings	Laboratory for Satellite Altimetry (NOAA)
Coastline	Coast	Cover/digitisation of nautical charts and aerial photography	Hellenic Ministry of Environment
Distance to coast	DCoast	Grid and cover	Extracted from coastline
Temperature slope (thermal fronts)	SSTsl	Grid	Extracted from SST grid
Longitude and latitude of stations	LON, LAT	Cover in decimal degrees and metres	SIMRAD EK500/BI500 system on April/May 1998 in Thermaikos Gulf
Day-dark-night-dawn categorical factor	DDND	Cover and grid	Based on sampling date and hour
Depth slope	DEPsl	Grid	Extracted from bathymetry grid

The partitioning of data was based on Tetko et al. (1995) whereby successive points are separated to construct the training and the verification sets. In this case, of every five sequential points, the first four are included in the training set and the fifth assigned to the verification set. The latter set thus consists of sampling points that cover the whole sampling area and are spaced $\sim$5 Nm apart. The selection of this distance, which determines the proportion of the training and verification datasets was based on the fact that, at this distance, no spatial autocorrelation is detected. Semi-variograms (Matheron, 1971) using ESRI's ArcGIS Geostatistical Analyst Software (GAS) and Auto-correlation Function Estimation plots in R statistical software (R Development Core Team, 2005) were used to identify spatial patterns in the raw data, the training and the verification sets and model residuals. Moran's *I* spatial autocorrelation statistics (Moran, 1950) was also used to estimate the spatial pattern of the two datasets. Furthermore, the homoscedasticity of the residuals was tested by plotting response and explanatory variables against the residuals. Since the training and the verification dataset were not normally distributed non parametric test namely Mann–Whitney and Kolmogorov–Smirnov were used to compare the two datasets. By these processes, the suitability of the verification set for testing SDM predictions was verified.

The remotely sensed and topographical data that have been used for SDM development are listed in Table 1. Initially, a large number of explanatory variables were collected that could potentially be related to species distribution. These could be classified as environmental variables (i.e. sea surface salinity, current speed and direction), spatial data (i.e. distance to coast), temporal data (i.e. date and hour of sampling) and oceanographic features (i.e. productivity hotspots, Valavanis et al., 2004). After an extensive exploration process (see Palialexis et al., this issue), only the non-correlated explanatory variables were used in order to avoid any biased estimations. Small pelagic species life-history indicates that most of the variables used do greatly influence species distribution (Daskalov et al., 2003; Santos et al., 2004; Ruiz et al., 2006; Planque et al., 2007). Additionally, certain oceanographic features like upwelling, gyres and river outflows (Garcìa & Palomera, 1996; Bakun, 2001) affect small pelagic

species distribution and are easily identified in remotely sensed data (sea surface temperature, chlorophyll-*a* concentration) (Valavanis et al., 2004). Spatial resolution of the explanatory variables varied from 0.01 to 0.04 decimal degrees. For modelling purposes, all datasets were interpolated to the lowest resolution. A dataset with the final selected explanatory variables that covers the study area at a resolution of 0.01 decimal degrees was applied in the SDMs in order to produce high resolution prediction maps of small pelagic species. Since the response variable is well explored, functionally relevant predictors were selected in order to cover both the environmental and geographical space (Elith & Leathwick, 2009). The multi-dimensional nature of the environmental predictors raises spatial autocorrelation issues in the SDMs, which are discussed later.

SDM

The SDM methods that were used are listed in Table 2. The selection of the explanatory variables used in each model was based on each method's parameter selection process or on the statistically meaningful contribution of the parameters to models. Approaches that were developed by the same method were compared regarding their predictive capacity and the trade-off between explained variation and model complexity. The one that performed better was finally selected. Less complex nested models developed by the same approach reduced predictive capacity and increased prediction errors compared to those finally selected. Additionally, an increase in model complexity beyond a certain threshold lead to an increase in prediction errors and was penalized. Akaike's Information Criteria (AIC) (Akaike, 1974), Cross-validation, Root-Mean-Square-Errors and Mean-Absolute-Error were used to assess the trade-off between model complexity and model predictive capacity, depending on the technique. In this way, it was confirmed that no unnecessary complexity was added in the models. The documentations and software used for each SDM are also listed in Table 2. All SDMs were developed as proposed by the authors mentioned in Table 2. A number of SDMs were implemented using several combinations of variables but only the model with the best fit and predictive capacity has been used in the comparison process.

GAMs, Generalized Additive Mixed Models (GAMMs) and Multivariate Analysis and Regression Splines (MARS) belong to the family of regression methods while MAXENT, BRT, ASNN, Artificial Neural Networks Ensemble (ANNE) and Support Vector Machines (SVM) are machine learning models. Bioclim Envelope Model (BIOCLIM) and Envelope Score (EnvScore) are envelope style methods using environmental data to define bioclimatic envelopes. Environmental Distance is a two-distance based method that makes use of a generic algorithm based on environmental dissimilarity matrices. Finally, Genetic Algorithm for Rule-set Prediction (GARP) uses a genetic algorithm that creates ecological niche models for species.

GAMs are generalized models involving a sum of smooth functions of covariates (Hastie & Tibshirani, 1990; Wood, 2006). GAMMs are also used, complementary to GAMs, in order to deal with spatial autocorrelation, which could lead to biased models and predictions. GAMs are the most frequently used approach in habitat modelling field (Valavanis et al., 2008) and several recent modifications and applications have increased their utilization (Leathwick et al., 2006b; Wood, 2006). The selection of the GAMs' smoothing predictors followed the method proposed by Wood & Augustin (2002), using the 'mgcv' library (Wood, 2008) in the R statistical software (R Development Core Team, 2005). The degree of smoothing was selected based on the observed data and the Generalized Cross Validation (GCV) method (Wood, 2006). First order interactions among the explanatory variables were also added in several GAMs. The best-fitting model was selected by using AIC and a stepwise forward selection was applied to restrict collinearity among the explanatory variables. The binomial family was applied using a logistic link function. GAMMs were developed based on the final GAM model with the assumption that a specific correlation structure exists among all sampled points in the study area. This structure was modelled by using the binomial distribution.

Multivariate Adaptive Regression Splines (MARS) (Leathwick et al., 2005) is an alternative regression-based method used for fitting non-linear responses but it differs from GAM because it utilizes piecewise linear fits instead of smoothers. In particular, MARS is a technique in which non-linear responses between a species and an environmental predictor are described

Table 2 Resource selection functions applied, variables and software used

Model	Explanatory variables	Software	Reference
Generalized Additive Models (GAM)	SST, SLA, DEP, DDND	R [18], library: mgcv	Wood (2006), Hastie & Tibshirani (1990)
Generalized Additive Mixed Models (GAMM)	SST, SLA, DEP, DDND	R, library: mgcv, geoR, spatstat, spdep	Wood (2006), Hastie & Tibshirani (1990)
Boosted Regression Trees (BRT)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	R, library: gbm	Leathwick et al. (2006a)
Multivariate Analysis and Regression Splines (MARS)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	R, library: mda	Leathwick et al. (2005)
Maximum Entropy (MAXENT)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	Maxent software for species habitat modelling	Phillips et al. (2004)
Support Vector Machines (SVM and SNM-Nu)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	openModeller Desktop	Cristianini & Shawe-Taylor (2000)
Genetic Algorithm for Rule-set Prediction (GARP)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	openModeller Desktop	Stockwell (1999)
Envelope Score (EnvScore)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	openModeller Desktop	Nix (1986), Piñeiro et al. (2007)
Bioclim Envelope Model (BIOCLIM)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	openModeller Desktop	Nix (1986)
Environmental Distance (EnvDist and EnvDistChe)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	openModeller Desktop	Carpenter et al. (1993)
Associative Neural Networks (ASNN)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	Associative Neural Network software by http://www.vcclab.org	Tetko (2002a, b)
Artificial Neural Network Ensemble (ANNE)	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl, DDND	Associative Neural Network software by http://www.vcclab.org	Tetko (2002a, b)

by a series of linear segments of differing slope, each of which is fitted using a basis function as was described by Friedman (1991). Breaks between segments are defined by a knot in a model that initially over-fits the data and is then simplified using a backwards/forwards stepwise cross-validation procedure to identify terms to be retained in the final model. MARS is capable of fitting complex, nonlinear relationships between species and predictors and in one of its implementations can be used to fit a model describing relationships between multiple species and their environment (Leathwick et al., 2005). MARS is much faster than GAMs in model development and is easily utilized with GIS applications to generate species distribution maps. In this study, MARS was developed using the 'mda' library (Hastie et al., 1994) in R statistical software (R Development Core Team, 2005) and the 'MARS public function 3.1' by Leathwick & Elith (per. comm.). The selection of the explanatory variables was based on their contribution to model goodness of fit. Several models were developed including models with interactions among the explanatory variables. The final model was selected by comparing the predictive performance using Receiver Operating Characteristics and the Area Under Curve (ROC–AUC).

MAXENT estimates a target probability distribution by finding the probability distribution of maximum entropy (i.e. the most spread out or closest to uniform), subject to a set of constraints that represent incomplete information about the target distribution (Phillips et al., 2006). MAXENT is a general-purpose machine learning method with a simple and precise mathematical formulation and it has a number of characteristics that make it well-suited for species distribution modelling. It is based on the maximum-entropy principle developed by Jaynes (1957). Maximum Entropy Species Distribution Modelling software version 3.3.1 was used for model development. Several models were developed and the one with the lowest ROC–AUC and containing highly contributing variables was finally selected. MAXENT is user-friendly software, which provides outputs containing all the essential information about the models developed. Additionally, MAXENT's data output can be easily inserted in GIS for further analysis and generation of distribution probability maps.

In BRT (Leathwick et al., 2006a), each of the individual models consists of a simple Classification and Regression Tree (CART). The boosting algorithm uses an iterative method for developing a final model in a forward stage-wise way, progressively adding trees to the model by re-weighting the data in order to emphasize cases that are poorly predicted by the previous trees. Advantages offered by a BRT model include its ability to accommodate different types of predictor variables and missing values, its immunity to the effects of extreme outliers and the inclusion of irrelevant predictors and its facility for fitting interactions between predictors (Friedman & Meulman, 2003). BRT models were constructed using the BRT functions version 2.8, as developed by Leathwick & Elith (pers. comm.) for R statistical software (R Development Core Team, 2005), and the 'mda' library. The best performing model was selected according to the area under the Receiver Operating Characteristic curve.

Associative Neural Networks (ASNN) is a method with improved predictive abilities compared to traditional neural networks techniques, including combination of feed-forward neural networks and a *k*-nearest neighbour technique. This method uses the correlation between ensemble responses as a measure of distance of the analyzed cases for the nearest neighbour technique. This provides an improved prediction ability by correcting the bias of the neural network ensemble. An ASNN has a memory that can coincide with the training set. If new data become available, the network further improves its predictive ability and provides a reasonable approximation of the unknown function without the need to retrain the neural network ensemble. This feature of the method dramatically improves its predictive ability over traditional neural networks and *k*-nearest neighbour techniques. Here, an Artificial Neural Network Ensemble (ANNE) was developed using one hidden layer with three neurons. The number of the nearest neighbour, k, and parameter σ for the Parzen-window regression represent smoothing parameters of ASNN in order to minimize the ASNN error for the training set. ASNNs were initially applied in chemistry (Tetko et al., 1995), providing more accurate predictions than ANNE. More detailed information on ASNN development can be found in Tetko (2002a, b). Both ANNE and ASNN were developed in order to compare ASNN performance to traditional ANNE and to other modelling approaches. ANNE and ASNN models were selected based on processes

including the training algorithm, the number of neurons and hidden layers and the iterations and number of ensembles. The Early Stopping over Ensemble (ESE) method was used to train the neural networks (Bishop, 1995; Tetko & Tanchuk, 2002). Models presenting the lowest Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were finally selected (Tetko et al., 2008).

BIOCLIM is an 'envelope' method that implements a bioclimatic envelope algorithm (Nix, 1986; Busby, 1991). Environmental envelopes are conceptually closely related to niche theory as they strive to delineate the hyper-surface (or envelope) that best circumscribes suitable conditions within the niche hyper-space defined by the environmental variables. The algorithm finds the mean and standard deviation for each environmental variable (assuming normal distribution) associated with the occurrence of surveyed species presence points. Besides the envelope, each environmental variable has additional upper and lower limits taken from the maximum and minimum values related to the set of occurrence points. In this model, any point can be classified as: Suitable (when all associated environmental values fall within the calculated envelopes), Marginal (when one or more associated environmental value falls outside the calculated envelope but still within the upper and lower limits) or Unsuitable (when one or more associated environmental value falls outside the upper and lower limits). BIOCLIM's categorical output is mapped to probabilities of 1.0, 0.5 and 0.0, respectively. OpenModeller software (Muñoz et al., 2009) was used for BIOCLIM development.

Envelope Score (EnvScore) is analogous to the BIOCLIM approach and implements a Bioclimatic Envelope Algorithm. For each given environmental variable, the algorithm finds the minimum and maximum at all occurrence sites. The Envelope Score algorithm is equivalent to the inclusive 'OR' implementation of Bioclim described in Piñeiro et al. (2007). EnvScore models were developed using OpenModeller software (Muñoz et al., 2009).

Climate Space is a principle components-based algorithm developed by Neil Caithness (http://openModeller.sf.net). The component selection process in this algorithm implementation is based on the Broken-Stick cut-off, whereby any component with an eigenvalue less than n standard deviations above a randomised sample is discarded (see also Muñoz et al., 2009). The original Climate Space Model was written as series of Matlab functions.

Environmental Distance (EnvDist) uses a generic algorithm based on environmental dissimilarity metrics. When combined with the Gower metric (Gower & Legendre, 1986) and maximum distance is set to one, this algorithm should produce the same result as the algorithm known as DOMAIN (Carpenter et al., 1993). DOMAIN is a distance-based method that assesses new sites in terms of their environmental similarity to sites of known presence by transforming the known occurrences into an environmental space and computing the minimum distance in environmental space from any cell to a known presence of the species. EnvDistChe was developed using Chebyshev distance instead of the Gower metric. Chebyshev distance is a metric defined in a vector space such that the distance between two vectors is the greatest of their differences along any coordinate dimension. Both models were developed using OpenModeller software (Muñoz et al., 2009).

Support Vector Machines (SVMs) consists of a set of related supervised learning methods that belong to the family of generalized linear classifiers. They could be considered as a special case of Tikhonov regularization (Tychonoff & Arsenin, 1977). SVMs simultaneously minimize the empirical classification error and maximize the geometric margins. The model produced by support vector classification depends only on a subset of the training data because the cost function for building the model does not take into account training points that lie beyond the margin (Vapnik, 1995; Schölkopf et al., 2000). SVMs are able to represent nonlinear effects and interactions between variables by projecting the explanatory variables into a higher dimensional feature space where the prediction problem has a linear solution (Moguerza & Muñoz, 2006). Two SVM were developed (SVM-Nu and SVM) because of their different performances in relation to fitting efficiency and predictive capacity. SVM-Nu differs to SVM in the degrees used in Kernel function. Both approaches were developed using openModeller software.

The GARP uses a genetic algorithm to select a set of rules (e.g. adaptations of regression and range specifications) that best predicts the species distribution (Stockwell & Peters, 1999). The Genetic Algorithm used in GARP is based on the basic concept developed by Holland (1975). GARP creates

ecological niche models for species, identifying where the environmental conditions could maintain populations. For input, GARP uses a set of point localities where the species are known to occur and a set of geographic layers representing the environmental parameters that might limit the species' capabilities to survive. In this study, based on openModeller software (Muñoz et al., 2009), the algorithm applies the Best Subsets procedure using the new openModeller implementation in each GARP.

Comparison

There are several techniques to validate a model or to compare the accuracy of prediction among different models. Kappa statistics, ROC–AUC, *k*-Fold cross validation, confusion matrices and classification tables are well described by Boyce et al. (2002) for presence–absence data.

Models that predict the presence or absence of a species are normally judged by the number of prediction errors. There are two types of prediction errors: false positive (FP) and false negative (FN). The performance of a presence–absence model is normally summarized in a confusion or error matrix (Table 3) that cross-tabulates the observed and predicted presence–absence patterns. Morrison et al. (1992) refer to FP errors as type I and FN errors as type II errors. FP or commission error leads to an over-prediction while FN or omission error leads to an under-prediction. Generally, omission error could be characterized as 'hard'—true error while commission might or might not be a true error. Commission error can relate to unsuitable areas (true error), suitable areas with no sampling effort (species may be there), or suitable areas where historical (barriers, dispersal capability) or biotic (competition, predation) factors have impeded occupation by the species or caused it to go extinct. An accurate presence–absence model should be characterized by low omission error. On the other hand, low commission error indicates that the model over-fits the training data while high commission error indicates that the model over-predicts the training set. Specificity and sensitivity are terms analogous to omission and commission errors, although they refer to correctly predicted presence and absence instead of the errors. Specificity is the proportion of observed negatives correctly predicted and reflects a model's ability to predict an absence given that a species actually does not occur at a location. Sensitivity is the proportion of observed positives correctly predicted and reflects a model's ability to predict a presence given that a species actually occurs at a location.

Table 3 Confusion matrix summarizes observed and predicted presence/absence values

Confusion matrix	Predicted present	Predicted absent
Actually present	True positive	False positive (error type I)
Actually absent	False negative (error type II)	True negative

In this study, SDM comparison was achieved using the best representative models derived using each function. ROC–AUC (Fielding & Bell, 1997) was used because in contrast to other model evaluation methods (Kappa statistics, confusion matrices and classification tables, see Boyce et al., 2002), it avoids the problem of threshold value selection (Lehmann et al., 2002). ROC-plots and the Area Under the Receiver Operating Characteristic Curve measure the ability of a model to discriminate between those sites where a species is present and those where it is absent, and they have been widely used in the species distribution modelling literature (Elith et al., 2006). ROC–AUC values range from 0 to 1, with 1 standing for perfect discrimination, 0.5 for predictive discrimination close to a random guess and values <0.5 indicating performance worse than random (Boyce et al., 2002; Elith et al., 2006).

The correlation (COR) between the observation in the presence–absence dataset (a dichotomous variable) and the prediction is known as the point biserial correlation, and it can be calculated as a Pearson correlation coefficient (Zheng & Agresti, 2000). It is similar to ROC–AUC but carries extra information: instead of being rank based, it takes into account the difference between the prediction and the observation. This gives further insight into the distribution of the predictions and provides information on the model's discrimination (Murphy & Winkler, 1992).

The Kappa statistic (Cohen, 1960) summarizes all the available information in the confusion matrix. Kappa measures the proportion of correctly classified units after accounting for the probability of chance

agreement. Kappa, which is a chance-corrected measure of agreement, is commonly used in ecological studies with presence–absence data (Boyce et al., 2002). It requires a threshold to be applied to the predictions in order to convert them to presence–absence predictions. Kappa provides an index that considers both omission and commission errors. In this study, a maxKappa is used for each model generated by using the 'PresenceAbsence' library of the R statistical software (R Development Core Team, 2005).

Confusion matrices for the modelling approaches were formulated for both predictions on the training and the verification set. Omission and commission errors, sensitivity and specificity as well as Kappa statistics were estimated from the confusion matrices. Since Kappa is threshold-dependent, in order to avoid threshold selection, the maxKappa was used (Liu et al., 2005). The correlation coefficient between predicted and observed values in both datasets was also estimated. ROC–AUC was also used to classify the accuracy of the predictions. Finally, the probability maps/grids that were generated from each modelling technique were compared for their spatial similarities, using ESRI's ArcInfo correlation function for grids.

The Akaike Information Criterion (AIC) was also used for model selection. However, for many adaptive, nonlinear techniques, estimation of the effective number of parameters and consequently the AIC calculation is very difficult (Hastie et al., 2009). For this reason, all critical comparisons were mainly based on cross-validation techniques and ROC–AUC, whilst the AIC was used for the best candidate model within a given model family, taking into account the trade-off between model complexity and predictive capacity. Cross-validation is also preferable for theoretical reasons. Hastie et al. (2009) found, in simulation experiments, that the AIC can greatly overestimate the prediction error (>30%) compared to the cross-validation procedure. Nevertheless, the ordinary cross-validation procedure does not work well when the data are autocorrelated, resulting in underestimation in error prediction and consequentially in biased model selection (Hastie et al., 2009). This is the case in the observed spatio-temporal autocorrelation in hydroacoustic abundance records, which is a property of the biomass structure, not of the measurement processes (Simmonds & MacLennan, 2005). Their spatial characteristics, estimated for instance in geostatistics as nugget and range parameters, are affected by the selected acoustic integration unit (ESDU). Presence of a spatial structure in the errors, causes, among other things, underestimated standard errors of the slopes in the regression model and it represents a serious shortcoming for hypothesis testing and prediction (Ostrom, 1990). In this study, the spatial autocorrelation of the training and the verification dataset was estimated and the residuals of each modelling technique were checked for potential spatial patterns. In GAMMs the spatial autocorrelation pattern was inserted in the model, while in other methods several covariates were used to absorb the autocorrelated errors (Elith & Leathwick, 2009).

Initially, the predictive efficiency of each method was tested on the training set. In this case, the best performing techniques are considered to model the sampling data more accurately and thus, they describe species distribution more accurately. On the other hand, this does not necessarily reflect the predictive capacity of the methods, which is better presented by the predictions on the verification set, which is an unknown dataset to the training process of SDMs. The process of model evaluation is crucial in the SDM field, though there are diverse opinions on what properties of a model are important and how to test them appropriately (Elith & Leathwick, 2009). During SDM development to explain patterns or biological relationships, statistical tests of model fit and comparison with existing knowledge are generally used. In the case where an SDM is developed to predict species distribution in time or space, the predictive capacity is evaluated using either resampling techniques (cross-validation, bootstrapping) or an independent dataset. In this study, both the fitting efficiency and the predictive capacity of different SDMs were compared. For these processes, ROC–AUC was used as a threshold-independent index that quantifies the predictive performance of the models while omission and commission errors were used as prediction quality indices with respect to over- and under-prediction and over-fitting. MaxKappa was also used to complement ROC–AUC, as a chance-corrected measure of agreement, and COR was used to estimate the similarities between observed and predicted values.

Results

Spatial patterns and comparison among training and verification datasets

The selected validation dataset contains measurements omitting, in each step, h units (h equals at least 5 n mi), where h is chosen according to the empirical variogram of both validation set and prediction residuals. The empirical variograms (Fig. 2) and the autocorrelation function plots revealed a low autocorrelation, even in distances below the h limit, because the nugget was at least 65% of the sill and the range was larger than 10 km (5.4 n mi). The verification set presents no autocorrelation, as shown in Fig. 2. The selected distance among the points in the verification set reduces the spatial autocorrelation (Moran's I Index = 0.03), while in the training set the index was equal to 0.14. The index indicates that the verification set presents some clusters that might be due to random chance; however, there is less than 1% likelihood that the clusters in the training are due to chance alone. Residuals of the SDMs were also tested for spatial patterns, but no significant patterns were observed, presumably due to the adequate set of predictors used and the appropriately specified model fit (Elith & Leathwick, 2009). These results confirm that the verification set is not spatially autocorrelated, while there is an amount of autocorrelation in the training set (Fig. 2).

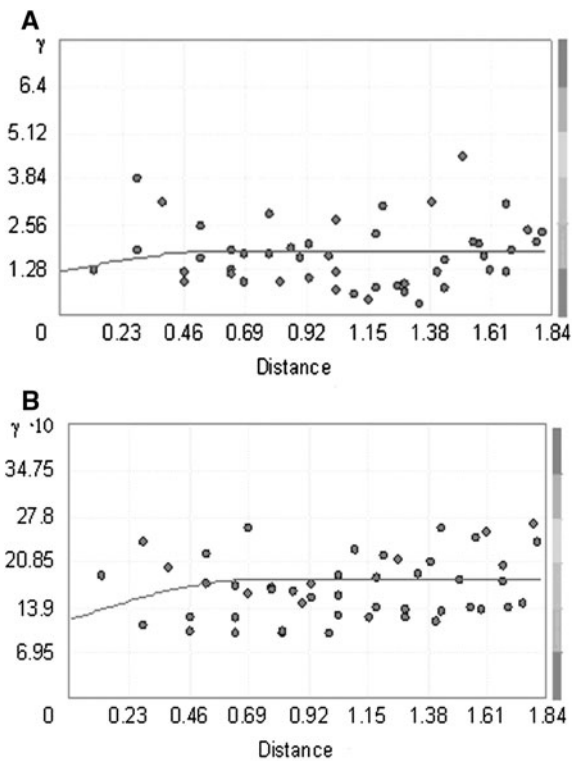

Fig. 2 Empirical variograms and autocorrelation function plots of the acoustic density that correspond to the validation dataset (**A**) and the training dataset (**B**)

Although the training and the validation datasets differed in their autocorrelation pattern, it was verified using the Mann–Whitney test that the null hypothesis was not rejected ($U = 1550$, P-value $= 0.48$) and both dataset refer to the same 'population' as derived by the definition of the test. Additionally, the Two-sample Kolmogorov–Smirnov test indicated that both dataset have the same distribution since the null hypothesis was also rejected ($Z = 0.56$, P-value $= 0.87$). The mean and standard deviation of the validation set (196 s_A and 220) was relatively higher than the training set (193 s_A and 215, respectively). Despite the similarities among the two datasets and the fact that they correspond to the same 'population' they could not be characterized as identical since their vector of values were not significantly correlated (Pearson's correlated coefficient $= -0.202$).

Fitting efficiency

ROC–AUCs and the associated standard deviations of all models are presented in Fig. 3A. Models with the highest ROC–AUC and the lowest standard deviation provide the best fitted SDMs. This is depicted in Fig. 3A (upper right). BRTs, EnvDist, EnvDistChe and SVM-Nu out-perform the other approaches achieving ROC–AUC greater than 0.9. Regression models (GAM, GAMM, MARS) as well as ASNN and SVM also acheived a high ROC–AUC (0.86–0.9). ANNE, GARP and MAXENT had AUCs in the range of 0.81–0.76. BIOCLIM, EnvScore and ClimSpace did not perform so well, achieving ROC–AUCs less than 0.64 while ClimSpace's AUC was 0.52.

The COR, which indicates the similarities between observed and predicted values, and the maxKappa are presented in Fig. 3B. Generally, the clusters (in relation to performance) of the modelling techniques are analogous to those indicated by the ROC–AUC,

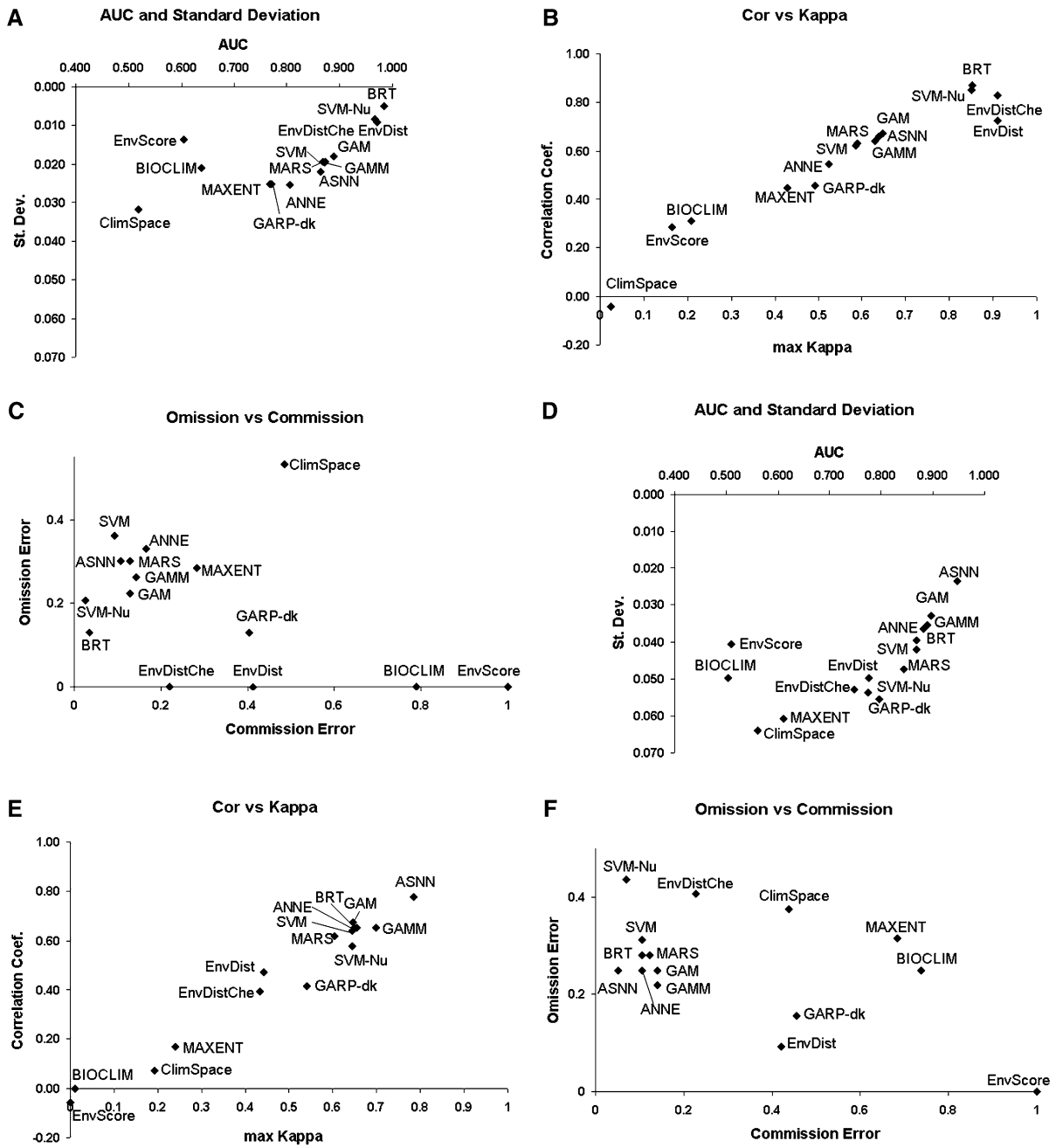

Fig. 3 Comparison of fitting efficiency (**A**, **B**, **C**) and predictive capacity (**D**, **E**, **F**) between the SDMs. ROC–AUC and the associated standard deviation scored by modelling approaches applied on the training set (**A**) and verification set (**D**). Correlation Coefficient and maxKappa scored by modelling approaches applied on the training (**B**) and verification set (**E**). Omission and Commission errors of modelling approaches applied on the training (**C**) and verification set (**F**)

though some differences in the classification are apparent for maxKappa and COR. BRTs and SVM-Nu model predictions were most highly correlated with the training dataset (0.87 and 0.85, respectively). However, EvnDist and EnvDistChe achieved the highest maxKappa (0.91). ASNN, GAM and GAMM performed almost equally (COR: 0.67–0.64 and maxKappa: 0.65–0.62). MARS performs equally to

SVM (COR: 0.63–0.62 and maxKappa: 0.59–0.58). ANNE achieved COR of 0.55 and maxKappa 0.52 while MAXENT and GARP present similar COR (0.45 and 0.46, respectively) but maxKappa is higher for GARP than for MAXENT (0.49 over 0.43). Bioclim, EvnScore and especially ClimSpace obviously failed to fit the training dataset presenting COR 0.31, 0.29 and 0.04, respectively, and maxKappa 0.2, 0.16 and 0.02, respectively.

As mentioned above, omission and commission errors reflect the quality of the predicted values with respect to over- and under-prediction and over-fitting. Figure 3C depicts the omission and commission errors of the different modelling techniques applied. High omission values indicate poor fitting efficiency (e.g. ClimSpace). Zero omission error combined with high commission error indicates over-prediction of the potential species distribution (mainly EnvScore and BIOCLIM). Zero omission error combined with no commission error indicates that predicted values over-fit the training values. So, it is expected that EnvDistChe tends to over-fit the training set more than EnvDist. The probability maps (are discussed later) that were generated by relatively high omission and commission errors SDMs provide a visual interpretation of the biased predicted patterns.

Predictive capacity

Comparison of the observed values with predicted values, derived by a dataset 'unknown' to SDMs, indicates the predictive capacity of the techniques applied and provides additional evidence of the performance of SDMs (i.e. additional to information derived from the fitting process and its diagnostics). ROC–AUC and the associated standard deviation from the application of SDMs to the validation set are presented in Fig. 3D. Models with the highest ROC–AUC and the lowest standard deviation could be characterized as those with the highest predictive capacity (Fig. 3D, upper right). Additionally, techniques that performed relatively efficiently in predicting the training dataset, but failed to accurately predict the verification set, probably tend to over-fit the training data and thus suffer decreased generality. According to the ranking in Fig. 3D, ASNN clearly out-performs the other approaches, achieving ROC–AUC close to 0.96. Regression models (GAM, GAMM, MARS) as well as BRT, SVM and ANNE also achieved high ROC–AUC (0.84–0.9). SVM-Nu, EnvDist, EnvDistChe and GARP scored ROC–AUC from 0.75 to 0.8. MAXENT and ClimSpace had ROC–AUC values of 0.61 and 0.56, respectively. BIOCLIM and EnvScore did not perform well, achieving AUC 0.50 and 0.51, respectively.

The COR values, which indicates the similarities between observed and predicted values on the verification set, as well as the maxKappa values are presented in Fig. 3E. Generally, the resulting groups of the modelling techniques are analogous to those arising from the ROC–AUC, though some differences are seen between maxKappa and COR. ASNN shows the highest predictive capacity (0.78 COR, 0.78 maxKappa). A distinct cluster described by the COR range of 0.58–0.67 and maxKappa of 0.6–0.7 includes GAM, GAMM, ANNE, BRT, SVM, SVM-Nu and MARS. GARP presents a COR of 0.42 and maxKappa of 0.54. EvnDist and EnvDistChe scored almost equally according to maxKappa (0.44) but differ in the correlation of their predictions with the verification set (0.47 and 0.39, respectively). MAXENT, ClimSpace, BIOCLIM and EvnScore present lower predictive capacities, achieving COR from 0.17–0.00 and maxKappa 0.24–0.00.

In contrast to Fig. 3C (which refers to the trained models), Fig. 3F summarizes the predictive ability of the models in relation to the verification set in terms of the associated omission and commission errors. High omission error indicated model weakness in terms of identifying species occurrence, while high commission error indicates a model's inability to distinguish unsuitable habitats. ASNN, GAM, GAMM, ANNE, BRT and SVM provide the less erroneous predictions regarding the independent verification set. MAXENT, BIOCLIM and EvnScore are the models with the highest commission error while SVM-Nu, EnvDistChe and ClimSpace have highest omission error.

Probability maps

Models developed by each approach were applied to grids of predictor variables in order to generate the corresponding species distributions maps. When presence–absence data are used, the maps generated are actually probability maps that denote the probability of species occurrence. Acoustic data used in this study were converted to presence–absence data. Presence corresponds to high acoustic density, indicating suitable

species habitats, while absence corresponds to low acoustic density indicating species absence or low fish density. Environmental variable grids were used for generating maps at a spatial resolution of 0.01 decimal degrees. In addition to predicting the fine scale distribution of small pelagic species, these maps should be helpful to identify potential habitat heterogeneity. Figure 4 depicts the probability maps derived from

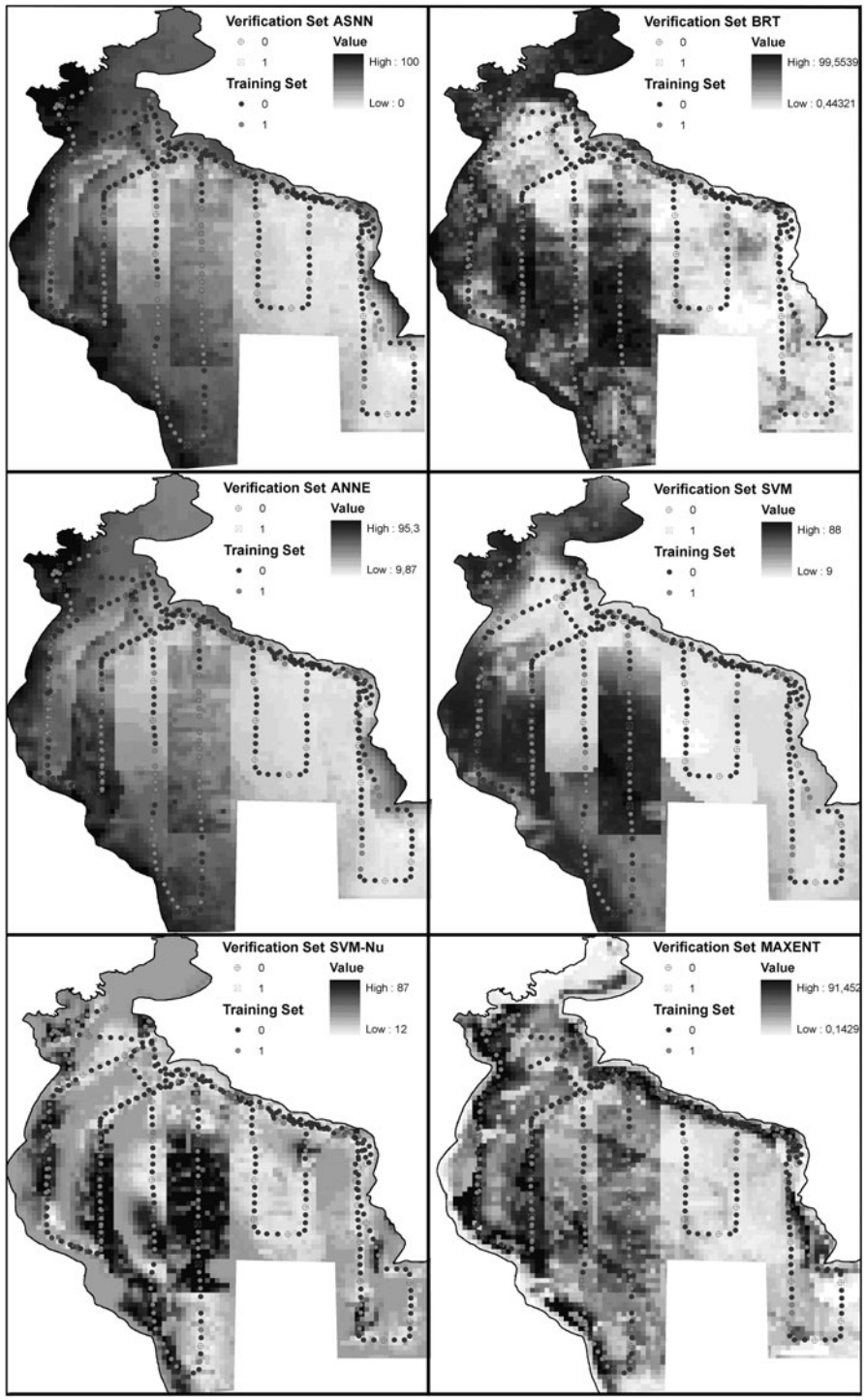

Fig. 4 Predicted probability distribution maps generated by machine learning approaches

machine learning approaches (BRT, ASNN, ANNE, MAXENT, SVM). The probability maps of the regression approaches are presented in Fig. 5, while the probability maps that arise from the envelope style approaches are presented in Fig. 6. Neither EnvScore nor BIOCLIM supports absence data. Finally, probability maps generated by EnvDist, EnvDistChe, ClimSpace and GARP are depicted in Fig. 7. Among the latter methods, only GARP supports the use of absence data.

Table 4 presents the Pearson correlation coefficients among SDMs. The upper-right part corresponds to the correlation among predictions on the validation set. The lower-left part corresponds to the correlation among the predicted grids as estimated using ESRI's ArcInfo correlation function for grids.

Discussion

Spatial structure in data

Patterns of spatial autocorrelation are common in species and biomass abundance or other ecological records (Legendre, 1993). Consequently, standard statistical models based on such data may violate the basic assumption that residuals are independent. Possible causes of spatial autocorrelation are categorized in three groups: the nature of the biological processes involved, the absence of important explanatory variables in the model and the linear modelling of a process that in reality is non-linear (Legendre & Legendre, 1998).

Commonly used methods to deal with the problem of spatial structure in the errors, are based on:

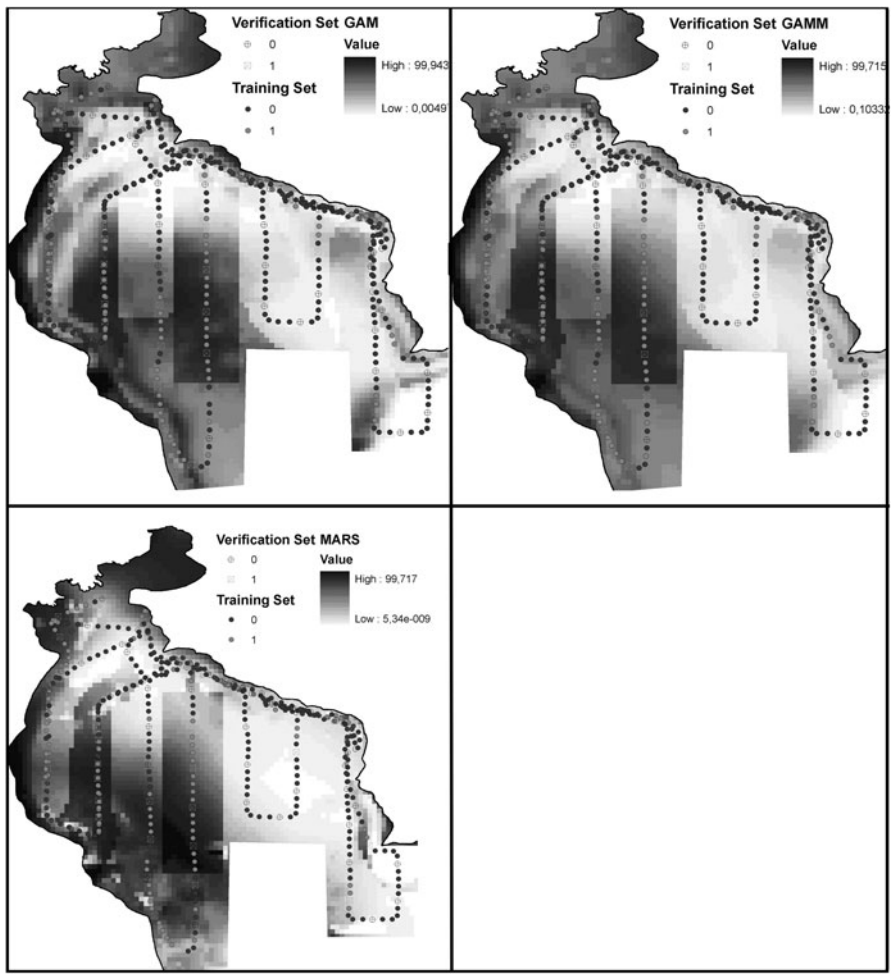

Fig. 5 Predicted probability distribution maps generated by regression approaches

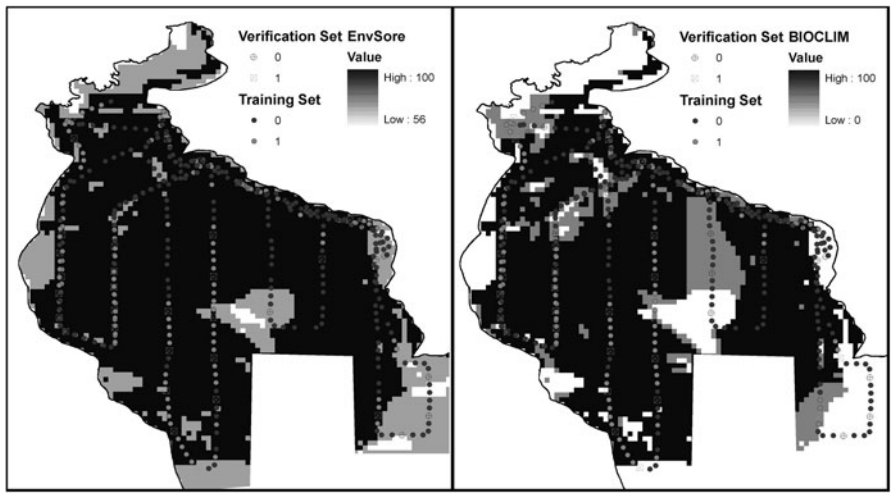

Fig. 6 Predicted probability distribution maps generated by envelope style approaches

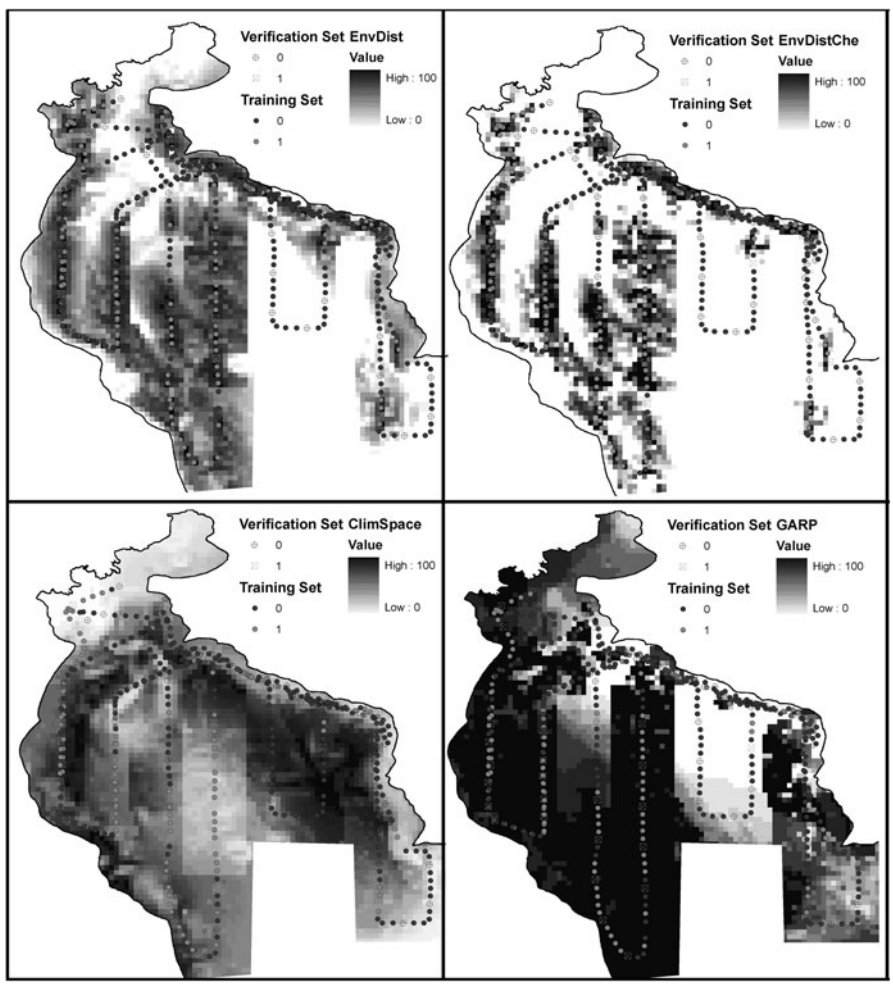

Fig. 7 Predicted probability distribution maps generated by Environmental Distance, GARP and climate space approaches

Table 4 Pearson's correlation coefficient among SDMs' predictions on the validation set (upper right) and among predicted grids (lower left)

Pearson R	GAM	GAMM	BRT	MARS	BIOCLIM	ClimSpace	EnvDist	EnvDistChe	EnvScore	GARP-dk	SVM	SVM-Nu	MAXENT	ANNE	ASNN
GAM		0.98	0.87	0.90	0.17	−0.04	0.64	0.53	0.10	0.63	0.82	0.70	0.49	0.90	0.85
GAMM	0.97		0.87	0.93	0.19	−0.02	0.61	0.49	0.12	0.69	0.85	0.70	0.48	0.93	0.85
BRT	0.84	0.87		0.85	0.18	−0.03	0.66	0.60	0.14	0.63	0.84	0.83	0.48	0.85	0.88
MARS	0.87	0.91	0.86		0.21	−0.04	0.60	0.47	0.14	0.65	0.80	0.64	0.47	0.89	0.80
BIOCLIM	−0.05	−0.02	−0.01	0.00		0.38	0.40	0.32	0.87	0.14	0.10	0.07	0.46	0.17	0.10
ClimSpace	−0.23	−0.20	−0.24	−0.26	0.41		−0.08	−0.07	0.30	0.17	0.04	0.08	−0.08	0.01	0.00
EnvDist	0.41	0.43	0.43	0.42	0.34	−0.06		0.87	0.24	0.28	0.49	0.54	0.69	0.60	0.69
EnvDistChe	0.30	0.29	0.32	0.28	0.26	−0.05	0.80		0.21	0.10	0.40	0.57	0.57	0.45	0.60
EnvScore	−0.11	−0.10	−0.10	−0.11	0.85	0.40	0.26	0.21		0.11	0.10	0.09	0.30	0.13	0.02
GARP-dk	0.55	0.60	0.56	0.55	0.16	−0.05	0.39	0.24	0.05		0.69	0.58	0.30	0.73	0.56
SVM	0.78	0.82	0.82	0.80	0.00	−0.23	0.50	0.34	−0.10	0.65		0.80	0.33	0.91	0.83
SVM-Nu	0.57	0.55	0.59	0.51	0.15	−0.14	0.63	0.65	0.09	0.46	0.67		0.20	0.75	0.82
MAXENT	0.31	0.34	0.36	0.33	0.43	−0.06	0.65	0.55	0.38	0.39	0.38	0.44		0.43	0.41
ANNE	0.73	0.79	0.71	0.76	−0.05	−0.16	0.39	0.23	−0.13	0.72	0.78	0.39	0.35		0.89
ASNN	0.73	0.78	0.72	0.76	−0.93	−0.20	0.36	0.21	−0.16	0.66	0.75	0.35	0.33	0.97	

(a) adding covariates, which can absorb the autocorrelated errors (see review by Elith & Leathwick, 2009), (b) choosing an appropriate ESDU, which reduces the autocorrelation for a given range, (c) applying wavelet-based methods for removing autocorrelation effects (Gudrun & Kühn, 2008), (d) extending the model in order to include the autocorrelation (e.g. extending GAM as GAMM) (Dray et al., 2006) or using the autocorrelation itself for interpolation purposes (Rossi et al., 1992; Simmonds & MacLennan, 2005).

Several modifications of ordinary cross-validation have been published to address the training-verification data dependence issue (Burman et al., 1994; Racine, 2000), omitting some data from the point of prediction and its neighbours within *h* units and using the remaining data for both model estimation and prediction. In this study, the selected validation dataset contains measurements omitting, in each step, *h* units (*h* equals at least 5 n mi), where *h* is chosen according to the empirical variogram of both validation set and prediction residuals. The empirical variograms and the autocorrelation function plots revealed a low autocorrelation, even in distances below the *h* limit. The verification set presents no autocorrelation, as shown in Fig. 2. A similar spatial structure has been encountered in previous surveys (October 1996, May 1997) even in different seasonal conditions (Georgakarakos & Kitsiou, 2008). Results from a comparative study using series of acoustic survey data from five different locations in Europe suggested that the spatial organization of the stock would be more dependent on environmental parameters than on fish abundance (Petitgas et al., 2001). This result disagrees with the general notion which relates stock size to spatial organization, at least for higher values of fish abundance (MacCall, 1990). In a similar study, school cluster characteristics (e.g. dimension, nb of schools) were correlated with total population school number but not with total population biomass (Muiño et al., 2003).

The autocorrelation characteristics of biomass are in agreement with the school clustering tendency of the biomass in previous surveys (1996, 1997) as this is estimated by the distance between two schools in a cluster (Petitgas et al., 2001). The estimated school and cluster descriptors from these surveys (average school number per km in the clusters, average ratio for summed school lengths/cluster length and maximum distance between two schools in a cluster) indicate a small aggregative scale in the biomass spatial structure.

Furthermore, the observed high representation in the species composition of sardine and the strong mixed aggregation with anchovy in the biological sampling did not allow development of alternative models utilizing the species composition as regressor variables. For the same reason the authors decided to work on acoustic density data without any transformation into biomass in order to avoid the propagation of the variability from the trawl sampling in the response variable. As a result, the distribution maps reflect the distribution of small pelagic species in the study area. In most SDM studies distribution patterns refer to a single species. In this case, the distribution maps correspond to small pelagic species in general (but mostly sardine and anchovy). The variable selection, the development of the models and the generation of the distribution maps were all carried out bearing in mind that the models refer to multispecies distribution; thus common features of their life-history were utilized. According to Stergiou & Lascaratos (1997), the distributions of these species are affected by environmental parameters, fishing activity, inter- and intra-specific competition. The use of small pelagic species as a group instead of a specific species includes the between-species competition and thus reduces the biotic parameters that affect species distribution. On the other hand, the distribution maps that are derived from this study are unsuitable to identify specific species–environment relations.

Verification process

The use of an independent well-structured presence–absence verification set is proposed as the optimal method to verify the predictive performance of the SDMs (Elith et al., 2006). On the other hand, the use of entirely independent datasets carries the comparing of different sampling strategies instead of evaluating a model (Lehmann et al., 2002). Alternatively, cross-validation (Jaberg & Guisan, 2001) and jackknife (Lehmann et al., 2002) are also proposed for model validation, especially where there are not sufficient data to be partitioned into a training and a validation set. According to Lehmann et al. (2002) and Jaberg & Guisan (2001), cross-validation,

bootstrapping and jackknife validation approaches are generally more practical, because they create relatively independent random data subsets and allow the use of all available data in the modelling process. These approaches are very useful in cases where insufficient data are available to be partitioned in a training set and a validation set that is not used during development of SDMs. However, Fu et al. (2005) and Simon et al. (2003) observed that cross-validation and, especially, leave-one-out cross-validation could lead to underestimation of prediction errors. In this study, a validation set was selected, as described earlier, since there were sufficient sampling records to formulate both datasets. The specific selection of the validation set overcomes the underestimation of predicted errors that could caused by cross-validation approaches, especially when acoustic data are spatially autocorrelated (Hastie et al., 2009). Additionally, there is no risk of comparing different sampling strategies, since the verification set is a sub-set of the raw data. These were also verified by the comparison of the training and the verification dataset. Generally, the SDM validation process is of great concern among species distribution modellers (Elith & Leathwick, 2009), while SDM evaluation would benefit from identifying useful techniques in other fields.

SDMs comparison

Among the machine learning techniques, SVM and MAXENT do use only presence data. Brotons et al. (2004) showed that predictions based on presence–absence data generally perform better than those based on presence-only data. Presence-only models can perform almost as well as presence–absence approaches, especially when survey coverage is evenly and widely distributed (MacLeod et al., 2008) but they contain no mechanism to control for biased sampling. In the present study, MAXENT under-performed compared to other machine learning approaches whereas SVM performed equally well with approaches that use presence–absence data.

The probability maps corresponding to machine learning models present notable similarities, identifying high probabilities of species occurrence near the coast, especially the west coast, and in the centre-to-south of the study area. In SVM-Nu and MAXENT, however, the predicted probability of occurrence along the west coast is lower than was the case for the other models. Additionally, high probabilities seem to overlap with the sampling transects, which could be an indication of over-fitting. Both approaches are characterized by relatively high omission error and thus do not predict the observed data accurately. SVM-Nu shows remarkable fitting efficiency for a presence-only model, having the second best ROC–AUC and COR scores (after BRT). However, its predictive capacity is relatively poor (as was indicated by the moderate ROC–AUC in total, and the highest omission error) and SVM-Nu over-fits the training dataset. This was not the case for SVM, making it the best modelling approach among those that do not support use of absence data. SVM performs at a similar level to the regression models.

BRTs, ASNN and ANNE were among the best performing models. In particular, BRT presents the best-fitting efficiency while its predictive capacity is relatively high compared to the other models. ASNN presents the best predictive capacity, and is characterized by satisfactory fitting efficiency (ROC–AUC 0.86). ANNE performs relatively well compared to other approaches, especially regarding its predictive capacity. However, performs less well than ASNN, as might be expected given its relationship to ASNN (Tetko, 2002a, b). ASNN achieves ROC–AUC, COR and maxKappa values that are markedly higher than the second best approach (GAM).

GAM, GAMM and MARS generate very similar probability maps, which confirms the similarities in fitting efficiency and predictive capacity. The comparison among the regression models indicates that GAM performs slightly better than GAMM while GAMM performs slightly better than MARS. The similarity in their performance was expected due to their common statistical origin. Even if there are other approaches that out-perform the regression models, either in fitting efficiency or in the predictive capacity, GAM, GAMM and MARS achieve relatively high values in the criteria used for both comparisons. Thus, the widespread use of regression models compared to that of other traditional approaches, like envelope style methods, GARP and MAXENT to predict species distributions is justified by their stability and performance.

The envelope style models failed to predict species distribution, achieving the worst ROC–AUC, COR and maxKappa values. They were characterized by high commission error (1 for EnvScore and 0.79 for

BIOCLIM) and, as shown in the probability maps, both approaches over-predict the training set. Envelope style approaches were initially developed to model data on terrestrial species from natural history museums and are probably unsuitable to model high resolution and density species occurrence data and to predict any habitat heterogeneity.

EnvDist and EnvDistChe performed relatively accurately in fitting the training data (ROC–AUC 0.97 for both). Only BRT and SVM-Nu achieved higher ROC–AUC, while EnvDist and EnvDistChe achieved the highest maxKappa (0.91) and relatively high COR values (0.73 and 0.83, respectively). Both models are characterized by zero omission error and EnvDistChe has almost half of the commission error of EnvDist (0.22 compared to 0.41). The zero omission error combined to the low commission error for EnvDistChe indicate that the model over-fits the training dataset and this fact is confirmed by the probability map where high probability regions are concentrated around the sampling transects. EnvDist shows a tendency to over-fit the training set, although not as much as EnvDistChe. This is also confirmed by the predictive capacity of EnvDist and EnvDistChe. Since the latter over-fits the training set, it is unable to accurately predict the independent set, presenting lower ROC–AUC, COR and maxKappa values than EnvDist. Compared to the other approaches, both models seem less effective in their predictive capacity than regression models and most of the machine learning techniques. Among methods that do not support absence data, EnvDist and EnvDistChe perform relatively well, but not as well as SVM.

ClimSpace failed to fit the training data or to predict the independent dataset. It had the worst ROC–AUC, COR and maxKappa values. Even if several ClimSpace models were developed (the best performing is presented here), none would succeed in modelling the training set. Thus, ClimSpace seems inappropriate to predict species distribution using acoustic data.

GARP shows moderate performance in both fitting efficiency and predictive capacity. Results and errors indicate that GARP corresponds well to the variables used, though the output, which reflects the environmental conditions where species could maintain populations is relatively coarse compared to the other approaches. The inability to generate more detailed species distribution maps makes GARP less efficient than the approaches that support presence–absence data, even if generally GARP's output grid is in agreement with the high probability spatial pattern that is identified by the most accurate SDMs.

The predicted grids for Thermaikos Gulf identify two distinct areas where small pelagic species are concentrated: first the west coastline from north to south and the east coastline of the Gulf, which are characterized by the presence of riverine waters, and, second, the central study area, which is related to gyre formation (Somarakis et al., 2002). These areas are characterized as nutrient-rich resulting in aggregations of small pelagic species. The areas identified are in agreement with other studies on small pelagic species (Somarakis et al., 2002; Giannoulaki et al., 2008; Tsagarakis et al., 2008). Correlations among grids are generally in agreement to correlations among predictions of SDMs on the validation dataset. Generally, equally performing SDMs generate grids that are significantly correlated, such as MARS and GAMM. Additionally, grids that generated by SDMs resulting from similar approaches are also highly correlated, such as regression models and neural networks. In order to evaluate the predicted spatial patterns of small pelagic species distribution, we refer to both these correlations among grids and to previous studies in the area.

Evaluation of models

It is well known that species distribution modelling is only as good as the data used (Hirzel & Guisan, 2002); in addition, SDM performance depends on the number of samples that is used to train the model. Different data types (e.g. abundance, presence-only data and richness) could produce different SDM rankings (Elith et al., 2006). Generally, predictions based on presence–absence data perform better than those based on presence-only data (Brotons et al., 2004), while presence–absence models generally perform better than abundance models (Francis et al., 2005). Presence-only models can perform equally well when survey coverage is evenly and widely distributed (MacLeod et al., 2008). In principle, abundance models should be more informative, however, their poor performance in practice is related to the fact that real abundance data rarely conform to standard distributions thus, violating model assumptions. The assumptions associated with

presence–absence data (binary distribution) are more easily met. Additionally, the validation process for presence–absence models (ROC–AUC, Kappa, Confusion matrix) is well developed and more informative compared to the validation techniques used in abundance models (*k*-fold cross-validation, models calibration, correlation), since it is easier to interpret presence–absence (binary distribution) models than abundance models (other distributions, e.g. Gaussian, Poisson). Finally, presence–absence models make less bold predictions about species distribution and are thus less likely to be proved wrong.

Studies of presence–absence modelling methods suggest that several non-linear techniques (e.g. GAM, ANN and MARS) are comparable in terms of predictive ability and they are often superior to methods such as traditional single decision trees (Ferrier & Watson, 1997; Elith & Burgman, 2002; Moisen & Frescino, 2002; Muñoz & Fellicisimo, 2004; Segurado & Araujo, 2004). Here, the similar performances of GAM and MARS is confirmed but ANNE and, especially, ASNN show higher predictive capacity, not only compared to traditional ANN but also compared to other widely used approaches (e.g. GAMs). Elith et al. (2006) evaluated the predictions of 11 distinct models and 16 approaches that use presence-only data. They classified the models into three performance categories. The first, highest performing, group includes MARS, BRT, generalized dissimilarity (GDM and GDM-SS) and maximum entropy (MAXENT and MAXENT-T) models. A second group of methods includes most of the standard regression methods (GAM/BRUTO, GLM, MARS and GARP). A third group includes the methods that use presence-only data (BIOCLIM, DOMAIN and LIVES). This study supports the high predictive ability of BRTs and the low predictive ability of Bioclim. EnvDist performed better in the present study, compared to the study by Elith et al. (2006), probably due to the fact that this function over-fits the training data (especially EnvDistChe), according to omission and commission errors. The small difference in ROC–AUC between MARS and GAMs that was observed in this study has also been observed in other studies. In particular, Leathwick et al. (2006a, b) fitted GAM and MARS models to the distributions of fifteen freshwater fish species in relation to their environment and, based on ROC values, they found little difference in the performances of both models. The higher predictive capacity of ASNN and ANNE in models trained with abundance data is also shown in Palialexis et al. (this issue).

The uncertainty associated with SDM predictions requires attention, especially when models are developed for decision-making and management purposes. Uncertainty in SDMs results both from data deficiencies and from errors in specification of the models (Elith & Leathwick, 2009). Problems related to uncertainty are often ignored because they are difficult to deal with. However, uncertainty can be minimized by the selection of (a) functionally relevant predictors that could explain the variance of the response variable both in environmental and geographical space and (b) SDMs that incorporate complex species–environment relations and variable interactions. The 'black-box' nature of the machine learning techniques cannot be very informative of such interactions, although results indicated their high predictive capacity. There is, however, a trade-off between variation explained and model complexity.

The use of biotic interactions, related to species life history, as explanatory variables in SDMs, e.g. prey–predator relations and fishing activity could increase the variance explained of the response variable. As mentioned by Guisan & Thuiller (2005), very few studies include variables that describe biological interactions. Elith & Leathwick (2009) indicate the difficulties of utilizing biological interactions as predictors. Such variables though could complement the variation explained in environmental space and identify more complex relationships in ecological space. In practice, the variable selection process depends on (a) the availability and quality of data, (b) the ability of data to explain a quantity of the variance of the response variable, based on biological knowledge or data exploration processes and (c) the assumptions of the SDMs. The latter point could exclude use of explanatory variables that are crucial from a biological point of view. In this case other modelling approaches could be useful in order to exploit the available biological inferences. Since the aim of this study was the comparative performance of SDMs, only well known and explored data were used. Additionally, the selection of explanatory variables was contingent on the requirements of available modelling software with grid generation capabilities. Fishing activity, interspecific competition and predator–prey relationships are all likely to affect small pelagic species distribution

(e.g. Ramzi et al., 2006; Sabatés et al., 2006). Such information could potentially explain a part of the variance that is not explained by the use of solely abiotic variables in cases of identification of species interactions or of the characterization of species distribution that approaches the realized habitat (Planque et al., 2007).

Most of the SDMs were able to depict the basic species distribution pattern, which is also confirmed by other studies in the area. The relatively novel SDMs provided more detailed outputs and, potentially, can indicate habitat heterogeneity with a high spatial resolution. Among SDMs that performed equally, the different explanatory variables used varied in terms of the importance of their contribution. Each modelling technique is able to explain a quantity of the variance of the response variable. Even if the proportion of the variance explained is equal for two SDMs, the part of the variance explained might differ and this is reflected by the different weights of the explanatory variables in the SDMs. Issues as the above should be considered carefully, especially when SDMs are used to improve ecological understanding, or for conservation planning and management.

Conclusions

The comparison of 13 species distribution models incorporating 15 different statistical approaches indicated that approaches belonging to Machine Learning Techniques are generally more accurate in predicting species distribution, utilizing presence–absence data, derived from predetermined sampling transects and a sufficient number of high resolution explanatory variables. In particular, BRTs outperformed the other techniques in fitting the training data, while ASNN showed remarkable predictive capacity in comparison with the other methods. SVM was the best performing technique among the approaches that do not support absence data. The aforementioned approaches did not over-fit the training dataset. Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to change behaviour based on data. The evolution in computer science supports more complex data simulations and models as well as combinations of techniques that are more accurate and efficient in their performance. That is the case in ASNN, which is a combination of an ANNE and a *k*-nearest neighbour algorithm, and in BRTs, which combine the boosting algorithm and regression trees to create a regression trees ensemble. The use and the evolution of such techniques in species distribution prediction generate a new perspective of more realistic and applicable outputs, while their performance may exceed that of more conventional techniques (Elith & Leathwick, 2009).

Regression models are ranked relatively highly compared to other techniques, with respect to their fitting efficiency and predictive capacity, and flexibility in modelling several types of data. GAMs, MARS and GAMMs performed almost similarly, though GAM output was slightly better. GAMM is able to model spatial autocorrelation, which is certainly present in the training dataset used, but did not exceed the predictive capacity of GAM. This could be due to the restricted spatial autocorrelation as shown in the variogram of the training dataset and/or because the autocorrelation in the response variable was adequately explained (statistically at least) by autocorrelation in the geographical and environmental predictors (Elith & Leathwick, 2009). As a note of caution, it should not be assumed that this will always be the case. Nevertheless, regarding regression models, it is suggested to use GAM for species distribution predictions or MARS as a more user-friendly approach.

When absence data are available, the loss of information in presence-only models affects their fitting efficiency and predictive capacity. In this study, approaches like BIOCLIM, EnvScore, ClimSpace and MAXENT failed to generate competitive outputs as compared to the other approaches. EnvDist was the only exception, although there are indications that this approach tends to over-fit the training data.

Generally, the fitting efficiency and the predictive capacity that characterize a model are strongly depended on the quality of the training data. In presence–absence data, derived from predetermined sampling transects that were modelled with high resolution environmental satellite and geographic data, BRTs and ASNN are suggested as the most appropriate techniques. Machine learning approaches, with their extensive analytical capabilities, could be useful tools for species distribution predictions. However, different study cases and datasets might require different approaches.

References

Aertsen, W., V. Kint, J. van Orshoven, K. Őzkan & B. Muys, 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. Ecological Modelling 221: 1119–1130.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716–723.

Amara, R., K. Mahé, O. LePape & N. Desroy, 2004. Growth, feeding and distribution of the solenette *Buglossidium luteum* with particular reference to its habitat preference. Journal of Sea Research 51: 211–217.

Bakun, A., 2001. 'School-mix feedback': a different way to think about low frequency variability in large mobile fish populations. Progress in Oceanography 49: 485–511.

Bishop, M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Bodholt, H., H. Nes & H. Solli, 1989. A new echo sounder system. Proceedings of the Institute of Acoustics 11(3): 123–130.

Boyce, M. S., P. R. Vernier, S. E. Nielsen & F. K. A. Schmiegelow, 2002. Evaluating resource selection functions. Ecological Modelling 157: 281–300.

Brotons, L., W. Thuiller, M. B. Araujo & A. H. Hirzel, 2004. Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27: 437–448.

Burman, P., E. Chow & D. Nolan, 1994. A cross-validatory method for dependent data. Biometrika 81(2): 351–358.

Busby, J. R., 1991. BIOCLIM—a bioclimate analysis and prediction system. In Margules, C. R. & M. P. Austin (eds), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Australia: 64–68.

Carpenter, G., A. N. Gillison & J. Winter, 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of animals and plants. Biodiversity and Conservation 2: 667–680.

Caruana, R. & A. Niculescu-Mizil, 2006. An empirical comparison of supervised learning algorithms. Proceedings of International Conference on Machine Learning, 23rd, Pittsburgh, PA.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20: 37–46.

Cristianini, N. & J. Shawe-Taylor, 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, London.

Daskalov, G. M., D. C. Boyer & J. P. Roux, 2003. Relating sardine *Sardinops sagax* abundance to environmental indices in northern Benguela. Progress in Oceanography 59: 257–274.

Dormann, C. F., J. M. McPherson, M. B. Araujo, R. Bivand & J. Bolliger, 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30: 609–628.

Dray, S., P. Legendre & P. R. Peres-Neto, 2006. Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecological Modelling 196: 483–493.

Elith, J. & M. A. Burgman, 2002. Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In Scott, J. M. (ed.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington, DC: 303–314.

Elith, J. & J. R. Leathwick, 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40: 677–697.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. Mc, C. Overton, A. T. Peterson, S. J. Phillips, K. S. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz & N. E. Zimmermann, 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129–151.

Ferrier, S. & G. Watson, 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, Canberra [avialable on internet at http://www.deh.gov.au/biodiversity/publications/technical/surrogates/].

Ferrier, M. D., G. Manion & G. Watson, 2002. Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New SouthWales experience. I. Species-level modelling. Biodiversity and Conservation 11: 2275–2307.

Fielding, A. H. & J. F. Bell, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38–49.

Francis, M. P., M. A. Morrison, J. Leathwick, C. Walsh & C. Middleton, 2005. Predictive models of small fish presence and abundance in northern New Zealand harbours. Estuarine, Coastal and Shelf Science 64: 419–435.

Friedman, J. H., 1991. Multivariate adaptive regression splines. Annals of Statistics 19: 1–141.

Friedman, J. H. & J. J. Meulman, 2003. Multiple adaptive regression trees with application in epidemiology. Statistics in Medicine 22: 1365–1381.

Fu, J. W., R. J. Carroll & S. Wang, 2005. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics 21: 1979–1986.

Garcìa, A. & I. Palomera, 1996. Anchovy early life history and its relation to its surrounding environment in the Western Mediterranean basin. Scientia Marina 60(2): 155–166.

Georgakarakos, S. & D. Kitsiou, 2008. Mapping abundance distribution of small pelagic species applying hydroacoustics and co-kriging techniques. Hydrobiologia 612(1): 155–169.

Giannoulaki, M., V. D. Valavanis, A. Palialexis, K. Tsagarakis, A. Machias, S. Somarakis & C. Papaconstantinou, 2008. Modelling the presence of anchovy *Engraulis encrasicolus* in the Aegean Sea during early summer, based on satellite environmental data. Hydrobiologia 612(1): 225–240.

Gower, J. C. & P. Legendre, 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification 3(1): 5–48.

Graham, C. H., C. Moritz & S. E. Williams, 2006. Habitat history improves prediction of biodiversity in a rainforest fauna. Proceedings of the National Academy of Sciences USA 103: 632–636.

Gudrun, C. & I. Kühn, 2008. Analyzing spatial ecological data using linear regression and wavelet analysis. Stochastic Environmental Research and Risk Assessment 22: 315–324.

Guisan, A. & W. Thuiller, 2005. Predicting species distribution: offering more than simple habitat models. Ecological Letters 8: 993–1009.

Guisan, A. & N. E. Zimmermann, 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135(2–3): 147–186.

Hastie, T. & R. Tibshirani, 1990. Generalized Additive Models. Chapman and Hall, London.

Hastie, T. J., R. Tibshirani & A. Buja, 1994. Flexible discriminant analysis by optimal scoring. JASA 89: 1255–1270.

Hastie, T., R. Tibshirani & J. Friedman, 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics), 2nd ed. Springer, Berlin.

Hirzel, A. H. & A. Guisan, 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling 157: 331–341.

Holland, J. H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor.

Jaberg, C. & A. Guisan, 2001. Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. Journal of Applied Ecology 38: 1169–1181.

Jaynes, E. T., 1957. Information theory and statistical mechanics. Physics Revisions 106: 620–630.

Leathwick, J. R., D. Rowe, J. Richardson, J. Elith & T. Hastie, 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biology 50: 2034–2052.

Leathwick, J. R., J. Elith, M. P. Francis, T. Hastie & P. Taylor, 2006a. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology Progress Series 321: 267–281.

Leathwick, J. R., J. Elith & T. Hastie, 2006b. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199: 188–196.

Lefkaditou, E., C.-Y. Politou, A. Palialexis, J. Dokos, P. Cosmopoulos & V. D. Valavanis, 2008. Influences of environmental variability on the population structure and distribution patterns of the short-fin squid *Illex coindetii* (Cephalopoda: Ommastrephidae) in the Eastern Ionian Sea. Hydrobiologia 612(1): 71–90.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74: 1659–1673.

Legendre, P. & L. Legendre, 1998. Numerical Ecology, 2nd English Edition. Elsevier Science BV, Amsterdam.

Lehmann, A., J. M. C. Overton & J. R. Leathwick, 2002. GRASP: generalized regression analysis and spatial prediction. Ecological Modelling 157: 189–207.

Liu, C., P. M. Berry, T. P. Dawson & R. G. Pearson, 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28: 385–393.

MacCall, A. D., 1990. Dynamic Geography of Marine Fish Populations. University of Washington Press, Seattle: 153.

MacLennan, D. N., P. G. Fernandes & J. Dalen, 2002. A consistent approach to definitions and symbols in fisheries acoustics. ICES Journal of Marine Science 59: 365–369.

MacLeod, D., C. L. Mandleberg, C. Schweder, S. M. Bannon & G. J. Pierce, 2008. A comparison of approaches for modelling the occurrence of marine animals. Hydrobiologia 612(1): 21–32.

Martin, P., N. Bahamon, A. Sabates, F. Maynou, P. Sanchez & M. Demestre, 2008. European anchovy (*Engraulis encrasicolus*) landings and environmental conditions on the Catalan Coast (NW Mediterranean) during 2000–2005. Hydrobiologia 612(1): 185–199.

Matheron, G., 1971. The Theory of Regionalized Variables and its Applications. Ecole Nationale Supe′rieure des Mines de Paris, Fontainebleau.

Moguerza, J. & A. Muñoz, 2006. Support vector machines with applications. Statistical Science 21(3): 322–336.

Moisen, G. G. & T. S. Frescino, 2002. Comparing five modelling techniques for predicting forest characteristics. Ecological Modelling 157: 209–225.

Moran, P. A. P., 1950. Notes on continuous stochastic phenomena. Biometrika 37: 17–23.

Morrison, M. L., B. G. Marcot & R. W. Mannan, 1992. Wildlife Habitat Relationships: Concepts and Applications. University Wisconsin Press, Madison, WI: 341.

Muiño, R., P. Carrera, P. Petitgas, D. J. Beare, S. Georgakarakos, J. Haralambous, M. Iglesias, B. Liorzou, J. Masse & D. G. Reid, 2003. Consistency in the correlation of school parameters across years and stocks. ICES Journal of Marine Science 60: 164–175.

Muñoz, M. E. S., R. Giovanni, M. F. Siqueira, T. Sutton, P. Brewer, R. S. Pereira, D. A. L. Canhos & V. P. Canhos, 2009. openModeller: a generic approach to species' potential distribution modelling. GeoInformatica. doi: 10.1007/s10707-009-0090-7.

Muñoz, J. & A. M. Fellicisimo, 2004. Comparison of statistical methods commonly used in predictive modeling. Journal of Vegetation Science 15: 285–292.

Murphy, A. H. & R. L. Winkler, 1992. Diagnostic verification of probability forecasts. International Journal of Forecasting 7: 435–455.

Nix, H. A., 1986. A biogeographic analysis of Australian elapid snakes. In Longmore, R. (ed.), Atlas of Elapid Snakes of Australia (Australian Flora and Fauna Series 7). Australian Government Publishing Service, Canberra: 4–15.

Olivier, F. & S. J. Wotherspoon, 2005. GIS-based application of resource selection functions to the prediction of snow petrel distribution and abundance in East Antarctica: comparing models at multiple scales. Ecological Modelling 189: 105–129.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, New York.

Palialexis, A., S. Georgakarakos, K. Lika & V. D. Valavanis, 2009. Comparing novel approaches used for prediction of species distribution from presence/absence acoustic data. Proceedings of the Second International Conference on Environmental Management, Engineering, Planning and

Economics (CEMEPE 09), June 21–26, 2009, Mykonos, Greece.

Palialexis, A., S. Georgakarakos, I. Karakassis, K. Lika & V. D. Valavanis, this issue. Fish distribution predictions from different points of view: comparing associative neural networks, geostatistics and regression models. doi: 10.1007/s10750-011-0676-6.

Petitgas, P., D. Reid, P. Carrera, M. Iglesias, S. Georgakarakos, B. Liorzou & J. Masse, 2001. On the relation between schools, clusters of schools, and abundance in pelagic fish stocks. ICES Journal of Marine Research 58: 1150–1160.

Phillips, S. J., M. Dudik & R. E. Schapire, 2004. A maximum entropy approach to species distribution modeling. Proceedings of the Twenty-First International Conference on Machine Learning: 655–662.

Phillips, S. J., R. P. Anderson & R. E. Schapire, 2006. Maximum entropy modelling of species geographic distributions. Ecological Modelling 190: 231–259.

Piñeiro, R., J. F. Aguilar, D. D. Munt & G. N. Feliner, 2007. Ecology matters: Atlantic-Mediterranean disjunction in the sand-dune shrub *Armeria pungens* (Plumbaginaceae). Molecular Ecology 16: 2155–2171.

Planque, B., E. Bellier & P. Lazure, 2007. Modelling potential spawning habitat of sardine (*Sardina pilchardus*) and anchovy (*Engraulis encrasicolus*) in the Bay of Biscay. Fisheries Oceanography 16(1): 16–30.

Poulos, S. E., G. T. Chronis, M. B. Collins & V. Lykousis, 2000. Thermaikos Gulf Coastal System, NW Aegean Sea: an overview of water/sediment fluxes in relation to air-land-ocean interactions and human activities. Journal of Marine Systems 25: 47–76.

R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria [available on internet at http://www.Rproject.org].

Racine, J., 2000. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. Journal of Econometrics 99: 39–61.

Ramzi, A., My. L. Hbid & O. Ettahiri, 2006. Larval dynamics and recruitment modelling of the Moroccan Atlantic coast sardine (*Sardina pilchardus*). Ecological Modelling 197: 296–302.

Ready, J., K. Kaschner, A. B. South, P. D. Eastwood, T. Rees, J. Rius, E. Agbayani, S. Kullander & R. Froese, 2010. Predicting the distributions of marine organisms at the global scale. Ecological Modelling 221(3): 467–478.

Redfern, J. V., M. C. Ferguson, E. A. Becker, K. D. Hyrenbach, C. Good, J. Barlow, K. Kaschner, M. F. Baumgartner, K. A. Forney, L. T. Ballance, P. Fauchald, P. Halpin, T. Hamazaki, A. J. Pershing, S. S. Qian, A. Read, S. B. Reilly, L. Torres & F. Werner, 2006. Techniques for cetacean-habitat modelling. Marine Ecology Progress Series 310: 271–295.

Richards, C. L., B. C. Carstens & L. Knowles, 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. Journal of Biogeography 34: 1833–1845.

Rossi, R. E., D. J. Mula, A. G. Journel & E. H. Franz, 1992. Geostatistical tools for modeling and interpreting ecological spatial dependence. Ecological Monographs 2: 277–314.

Ruiz, J., E. Garcia-Isarch, I. E. Huertas, L. Prieto, A. Juárez, J. L. Munõz, A. Sánchez-Lamadrid, S. Rodrìguez-Gálvez, J. M. Naranjo & F. Baldó, 2006. Meteorological and oceanographic factors influencing *Engraulis encrasicolus* early life stages and catches in the Gulf of Cádiz. Deep-Sea Research II 53: 1363–1376.

Sabatés, A., P. Martìn, J. Lloret & V. Raya, 2006. Sea warming and fish distribution: the case of the small pelagic fish, *Sardinella aurita*, in the western Mediterranean. Global Change Biology 12: 2209–2219.

Santos, A. M. P., A. Peliz, J. Dubert, P. B. Oliveira, M. M. Angélico & P. Ré, 2004. Impact of a winter upwelling event on the distribution and transport of sardine (*Sardina pilchardus*) eggs and larvae off western Iberia: a retention mechanism. Continental Shelf Research 24: 149–165.

Schölkopf, B., A. Smola, R. Williamson & P. L. Bartlett, 2000. New support vector algorithms. Neural Computation 12: 1207–1245.

Schröder, B., 2008. Challenges of species distribution modeling belowground. Journal of Plant Nutrition and Soil Science 171: 325–337.

Segurado, P. & M. B. Araujo, 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography 31: 1555–1568.

Siapatis, A., M. Giannoulaki, V. D. Valavanis, A. Palialexis, E. Schismenou, A. Machias & S. Somarakis, 2008. Modelling potential habitat of the invasive ctenophore *Mnemiopsis leidyi* in Aegean Sea. Hydrobiologia 612(1): 281–295.

Simmonds, E. J. & D. N. MacLennan, 2005. Fisheries Acoustics: Theory and Practice. Blackwell Science Ltd, Oxford.

Simon, R., M. D. Radmacher, K. Dobbin & L. M. McShane, 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 95(1): 14–18.

Somarakis, S., P. Drakopoulos & V. Filippou, 2002. Distribution and abundance of larval fishes in the northern Aegean Sea—eastern Mediterranean—in relation to early summer oceanographic conditions. Journal of Plankton Research 24: 339–357.

Stergiou, I. K. & A. Lascaratos, 1997. Climatic variability and the anchovy/sardine ratio in Hellenic waters. GeoJournal 41(3): 245–254.

Stockwell, D. R. B., 1999. Genetic algorithms II. In Fielding, A. H. (ed.), Machine Learning Methods for Ecological Applications. Kluwer Academic Publishers, Boston: 123–144.

Stockwell, D. & D. Peters, 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13: 143–158.

Tetko, I. V., 2002a. Associative neural network. Neural Processing Letters 16: 187–199.

Tetko, I. V., 2002b. Neural network studies. Introduction to associative neural networks. Journal of Chemical Information and Modeling 42: 717–728.

Tetko, I. V. & V. Y. Tanchuk, 2002. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. Journal of Chemical Information and Modeling 42: 1136–1145.

Tetko, I. V., D. J. Livingstone & A. I. Luik, 1995. Neural network studies. Comparison of overfitting and overtraining. Journal of Chemical Information and Modeling 5: 826–833.

Tetko, I. V., I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches & A. Varnek, 2008. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. Journal of Chemical Information and Modeling 48(9): 1733–1746.

Tsagarakis, K., A. Machias, S. Somarakis, M. Giannoulaki, A. Palialexis & V. D. Valavanis, 2008. Habitat discrimination of juvenile sardines in the Aegean Sea using remotely sensed environmental data. Hydrobiologia 612(1): 215–223.

Tychonoff, A. N. & V. Y. Arsenin, 1977. Solution of Ill-posed Problems. Winston and Sons, Washington. ISBN:0-470-99124-0.Valavanis, V. D., S. Georgakarakos, A. Kapantagakis, A. Palialexis & I. Katara, 2004. A GIS environmental modelling approach to Essential Fish Habitat Designation. Ecological Modelling 178: 417–427.

Valavanis V. D., A. Kapantagakis, I. Katara & A. Palialexis, 2004. Critical regions: A GIS-based model of marine productivity hotspots. Aquatic Sciences 66(1): 139–148.

Valavanis, V. D., G. J. Pierce, A. F. Zuur, A. Palialexis, A. Saveliev, I. Katara & J. Wang, 2008. Modelling of essential fish habitat based on remote sensing, spatial analysis and GIS. Hydrobiologia 612(1): 5–20.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. SpringerVerlag, New York.

Walline, P. D., 2007. Geostatistical simulations of eastern Bering Sea walleye pollock spatial distributions, to estimate sampling precision. ICES Journal of Marine Science 64: 559–569.

Wood, S. N., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC Press, Boca Raton.

Wood, S. N., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society: Series B 70(3): 495–518.

Wood, S. N. & N. H. Augustin, 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157: 157–177.

Zaniewski, A. E., A. Lehman & J. Overton, 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157: 261–280.

Zheng, B. & A. Agresti, 2000. Summarizing the predictive power of a generalized linear model. Statistics in Medicine 19: 1771–1781.