Comparing novel approaches used for prediction of species distribution from presence/absence acoustic data

A. Palialexis[1,2*], S. Georgakarakos[3], K. Lika[2] and V. D. Valavanis[1]

[1]Marine GIS Laboratory, Institute of Marine Biological, Resources, Hellenic Centre for Marine Research, Thalassocosmos 71003, Heraklion Crete, Greece

[2]Department of Biology, University of Crete, Vassilika Vouton, P.O. Box 2208, 71409 Heraklion Crete, Greece

[3]Sonar Laboratory, Department of Marine Sciences, University of the Aegean, University Hill, 81100 Mytilini, Lesvos, Greece

*Corresponding author: E-mail: andreaspal@her.hcmr.gr, Tel: +30 2810 337817, Fax: +30 2810337822

Abstract

Accurate modelling and prediction of fish spatial distributions, based on sampled data, provide essential information for management purposes and stock monitoring. This study compares current and novel modelling techniques, in order to justify their suitability and accuracy on acoustic data. Ten different Resource Selection Functions were tested, and Receiver Operation Characteristic and Area Under Curve indicated that Boosted Regression Trees and Generalized Additive Models appear to fit acoustic data more efficiently. The corresponding probability maps also indicated that these functions produce accurate species distribution patterns when used with presence/absence data.

Keywords: models' comparison; Resource Selection Function; Receiver Operation Characteristic; environmental data; acoustic data.

1. INTRODUCTION

Species ecological and geographic distribution is essential for conservation planning and forecasting [1], and for evolutionary determinants of spatial patterns of biodiversity [2]. Several approaches have been developed for the identification of species distribution using sampling data. Most of these approaches are grouped as Resource Selection Functions (RSF) and are statistical models defined to be proportional to the probability of use of a resource unit. Approaches based on RSFs have only been applied recently on marine species special characteristics, and several novel modelling methods have been proposed [3]. RSFs have been also used to study relationships between environmental parameters and species presence [4, 5], identifying essential species habitats [6] and forecasting species distribution with corresponding climate changes [7]. Easy access to satellite data, covering extended geographic areas, comprise an essential occasion of the wider use of RSF. Presence/absence data type, derived from related sampling strategies, are commonly used with RSFs and as Zaniewski et al. [8] argued, presence/absence modelling is more likely to reflect the present natural distribution derived from realized niche of the species, whereas presence-only methods are more likely to predict potential distributions that more closely resemble the fundamental niche.

In this study, alternative and novel approaches are used to improve model implementation of some well established modelling techniques. The RSFs that have been applied include Generalized Additive Models (GAMs) and Mixed Models (GAMMs), Maximum Entropy models (MAXENT), Boosted Regression Trees (BRTs), Environmental Distance, Genetic Algorithm for Rule-set Prediction (GARP), Support Vector Machines (SVMs), Bioclim, Environmental Distance, Envelope Score and Multivariate Adaptive Regression Splines (MARS). Each model's output consists of the final selected model, the probability map of species distribution, the predictive capacity and other specific characteristics that describe the "quality" of each approach. Receiver Operation Characteristic (ROC) and Area Under Curve (AUC) used mainly for the RSFs comparison. The aim of this study is to identify the most accurate method that is able to predict the potential small pelagic fish distribution based on presence/absence data in Thermaikos Gulf (NE Mediterranean). Additionally, the best fitted function for acoustic and satellite data is inquired. The advantages and disadvantages of each technique are discussed.

2. MATERIALS AND METHODS

2.1 Study area & data

The study area (Figure 1) is the Thermaikos Gulf, North Aegean, Northeastern Mediterranean Sea. Thermaikos Gulf is a semi-enclosed basin, relatively productive, because of the influence of four major rivers (Axios, Aliakmon, Loudias and Gallikos). As a result, bottom relief is smooth due to the continuous sediment input. The Thermaikos Gulf forms a wide continental shelf, which extends to the south into the 1400 m deep Sporades Basin. Water mass circulation is predominantly cyclonic [23]. Aegean water masses entrain the gulf from deeper layers along the eastern coast and move counterclockwise towards the gulf of Thessaloniki. Riverine waters usually move to the south along the western coast.

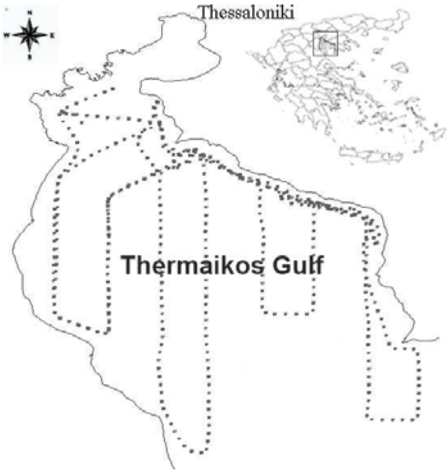

Figure 1. Study area and sampling transects

Acoustic fish density data (Sa: area backscattering coefficient) recorded through SIMRAD EK500/BI500 system on April/May 1998 in Thermaikos Gulf are transformed to presence/absence data (Figure. 1). 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%). The remotely sensed and topographic data that have been used for RSFs' development are presented in Table 1. Only non correlated parameters were used that could be able to interfere to small pelagic species distribution, based on their life-history characteristics. The resolution of each parameter is 0.01 degree (~1km).

Table 1. Data and their sources

Data Variable	Abbreviation	Data type / sensor	Archive Source
Acoustic data	Sa	Total acoustic integration (Area backscattering coefficient Sa per ESDU=1nm),	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	PAR	Grid / SeaWiFS	Distributed Active Archive

Available Radiation			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 & 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 & meters	SIMRAD EK500/BI500 system on April/May 1998 in Thermaikos Gulf
Current speed & direction	CURSP & CURDR	Grid & cover / NEMO (OPA9 + LIM)	Mercator operational oceanography
Day-night information	DN	Cover & grid	Based on sampling date & hour
Depth slope	DEPsl	Grid	Extracted from bathymetry grid

2.2 Modelling techniques and models' development

The RSF methods that were used are presented in Table 2. The selection of the explanatory variables that was used in each model was based on each method's parameter selection process, or parameter contribution information. Models that were developed by the same method were compared for their predictive capacity and the one that performed better was selected for that method's comparison. The documentations and software, used for each RSF, are also presented in Table 2. All RSFs were developed as proposed by the authors in Table 2. A number of RSFs were implemented in more than one way, but only the model with the best predictive capacity has been used in the comparison.

Table 2. Resource Selection Functions applied and variables used

Model	Explanatory Variables	Software	Reference
Generalized Additive Models, GAM	SST, CHL, PAR, SLA, DEP, SSTsl	R [18], library: mgcv	[9, 10]
Generalized Additive Mixed Models, GAMM	SST, CHL, PAR, SLA, DEP, SSTsl, DN	R, library: mgcv	[9, 10]
Boosted Regression Trees, BRT	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl	R, library: gbm	[3]
Multivariate Analysis and Regression Splines, MARS	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl	R, library: mda	[11]
Maximum Entropy, MAXENT	SST, CHL, PAR, SLA, DEP, SSTsl	Maxent software for species habitat modeling	[12]
Support Vector Machines, SVM	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl	openModeller Desktop	[13]
Genetic Algorithm for Rule-set Prediction, GARP	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl	openModeller Desktop	[14]
Environmental Distance, DOMAIN	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl	openModeller Desktop	[15]
Bioclim Envelope Model, Bioclim	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl	openModeller Desktop	[16]
Envelope score	SST, CHL, PAR, SLA, DEP, SSTsl, DCoast, DEPsl	openModeller Desktop	[16]

GAM, GAMM and MARS belong to regression approaches, while MAXENT, BRT and SVM are developed within the machine learning community. Bioclim and Envelope Score are envelope style methods, using environmental data to define bioclimatic envelopes. DOMAIN makes use of a generic algorithm, based on environmental dissimilarity matrices, and finally, GARP is using a genetic algorithm that creates ecological niche models for species.

2.3 Models' Comparison

RSFs' comparison was achieved using the best representative of each function. The method used to evaluate RSFs' predictive capacity, was the Receiver Operating Characteristic (ROC) [17], because in contrast to other models' evaluation methods (Kappa statistics, confusion matrices and classification tables [19]), ROC avoids the problem of threshold value selection [20]. ROC-plots and the Area Under the Receiver Operating Characteristic curve (AUC) measure the ability of a model to discriminate between those sites, where a species is present, and those where it is absent, and have been broadly used in the species' distribution modelling literature [21]. 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 indicate performance worse than random [19, 21].

3. RESULTS AND DISCUSSION

3.1 Models' outputs

The output information of RSFs depends on the function that has been used. Distribution prediction maps were generated by all methods utilized, and the AUC was estimated for all models. The variables that contributed to each RSF are shown in Table 2. Several models' characteristics were also used to evaluate their predicted ability and their fit to the data. Omission and commission errors describe the false predicted absence (underprediction) and false predictive presence (overprediction), respectively. Additionally, sensitivity is the proportion of the observed positives correctly predicted, and reflects a model's ability to predict a presence, given that a species actually occurs at a location. Specificity, on the other hand, is the proportion of the observed negatives correctly predicted, and reflects a model's ability to predict an absence, given that a species does not actually occur at a location. Both sensitivity and specificity are used for ROC-plots creation. The above characteristics and the predicted maps were used for the comparison process. RSFs using presence/absence data generate maps illustrating the probability of species' presence at each point on the grid.

3.2 Comparison results

The predicted ability of the RSFs, estimated with AUC, is presented in Figure 2. Environmental Distance, BRTs, GAMs and SVMs have generated relatively high values, while Bioclim and Envelope Score have predicted values almost equal to a random guess.

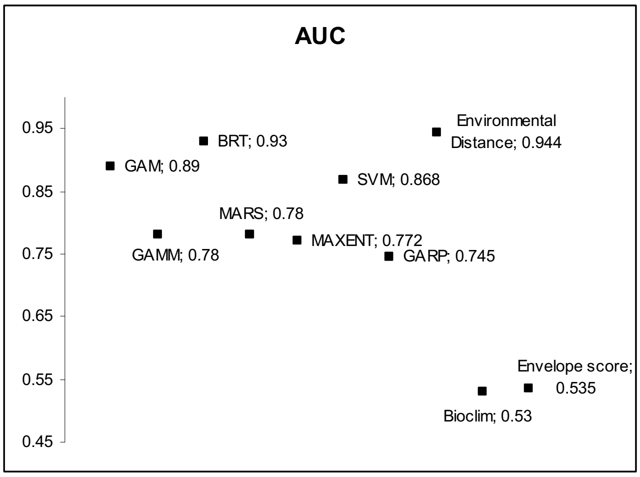

Figure 2. RSFs' comparison using AUC

It is well known that species distribution modelling is only as good as the data used [22]; in addition RSF's performance depends on the number of the samples that are used to train the model. The ranking of RSF's provides a general idea, regarding the methods' performance, though it concerns the specific dataset and area that have been used in this study. Different data types (e.g. abundance, only-presence data, and richness) could produce different RSFs' ranking [21]. Other studies comparing RSFs support the high predicted ability of BRTs and the low one of Bioclim [21]. The small AUC difference between MARS and GAMs is also observed by Leathwick et al. [11]. Environmental Distance overperformed in this study, in contrast to the study by Elith et al. [21], probably due to the fact that this function overfits the training data according to its resulted commission index.

3.3 Best performed RSFs

BRT, GAM and Environmental Distance probability maps are illustrated in Figure 3.

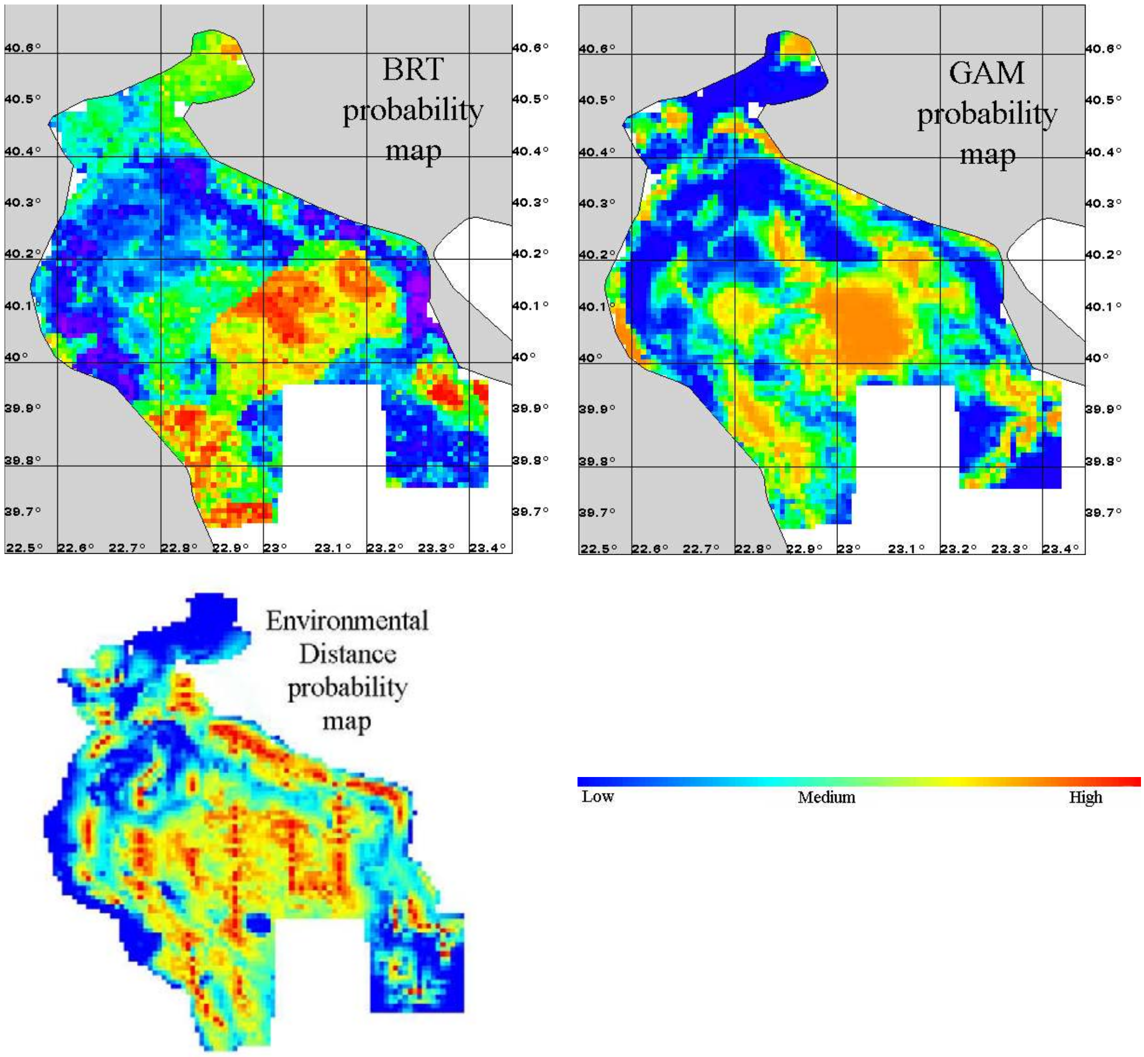

Figure 3. Probability distribution maps of small pelagic species occurrence

The maps in Figure 3 correspond to the three RSFs that generated the highest AUC values. Environmental Distance clearly overfits presence data, since high probabilities match sampling species occurrence. BTR and GAM produced quite similar maps, indicating common high probability areas. However, there seems to be a difference in the range of probabilities among the maps, caused by the model's different fit on training data. Probability maps that were generated by the rest of the RSFs offer relatively poor predictions of species' presence.

4. CONCLUSIONS

Many functions have been developed to model species distribution using presence-only or presence/absence data and environmental satellite images as predictors. Depending on data quality, number of sampling records, extend of area and data type, these functions perform differently. However, there are approaches like BRTs and GAMs that usually generate more accurate distribution maps, compared to other methods. In this study, ten RSFs were evaluated using AUC and the above mentioned methods overachieved the comparison process and the probability map justification. Environmental Distance also had high AUC value, though other model characteristics and the corresponding probability map indicated that this method overfitted the training data. We propose GAMs and BRTs to be the most appropriate approaches to handle acoustic presence/absence data and to provide accurate distribution probability maps; however different study cases might require more analytical method selection.

References

1. Ferrier S. et al., 2002. Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New SouthWales experience. I. Species-level modelling. *Biodiv. Conserv.*, **11**, 2275-2307.
2. Graham C. H., Moritz C., Williams S. E., 2006. Habitat history improves prediction of biodiversity in a rainforest fauna. *Proc. Natl. Acad. Sci. USA*, **103**, 632-636.
3. Leathwick J. R., Elith J., Francis M. P., Hastie T., Taylor P., 2006. 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.
4. Giannoulaki M., Valavanis V. D., Palialexis A., Tsagarakis K., Machias A., Somarakis S., Papaconstantinou C., 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.
5. Lefkaditou E., Politou C-Y., Palialexis A., Dokos J., Cosmopoulos P., Valavanis V. D., 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.
6. Planque B., Bellier E., Lazure P., 2007. Modelling potential spawning habitat of sardine (Sardina pilchardus) and anchovy (Engraulis encrasicolus) in the Bay of Biscay Fisheries Oceanography. **16 (1)**, 16-30.
7. Siapatis A., Giannoulaki M., Valavanis V. D., Palialexis A., Schismenou E., Machias A., Somarakis S., 2008. Modelling potential habitat of the invasive ctenophore *Mnemiopsis leidyi* in Aegean Sea. *Hydrobiologia*, **612 (1)**, 281-295.
8. Zaniewski A. E., Lehman A., Overton J., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157,** 261–280.
9. Wood S.N., 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press.
10. Hastie T., Tibshirani R., 1990. *Generalized additive models*. Chapman and Hall, London.
11. Leathwick J. R., Rowe D., Richardson J., Elith J., Hastie T., 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, **50**, 2034–2052.
12. Phillips S. J., Dudík M., Schapire R. E., 2004. A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 655-662.
13. Cristianini N., Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

14. Stockwell D. R. B., 1999. *Genetic algorithms II*. 123-144, in Fielding A. H., editor. *Machine learning methods for ecological applications*. Kluwer Academic Publishers, Boston.
15. Carpenter G., Gillison A.N., Winter J., 1993. DOMAIN: A flexible modeling procedure for mapping potential distributions of animals and plants. *Biodiversity and Conservation*, **2**, 667-680.
16. Nix H.A., 1986. A biogeographic analysis of Australian elapid snakes. In: *Atlas of Elapid Snakes of Australia*. (Ed.) R. Longmore, pp. 4-15. *Australian Flora and Fauna* Series Number 7. Australian Government Publishing Service: Canberra.
17. Fielding A. H., Bell J. F., 1997. A review of methods for the assessment of prediction errors in conservation presence/ absence models. *Environmental Conservation*, **24**, 38–49.
18. R Development Core Team, 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R -project.org.
19. Boyce M. S., Vernier P. R., Nielsen S. E., Schmiegelow F. K. A., 2002. Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.
20. Lehmann A., J. Mc, C. Overton & J. R. Leathwick, 2002. GRASP: generalized regression analysis and spatial prediction. Ecological Modelling 157: 189–207.
21. Elith J., Graham C. H., Anderson R. P., Dudik M., Ferrier S., Guisan A., Hijmans R. J., Huettmann F., Leathwick J. R., Lehmann A., Li J., Lohmann L. G., Loiselle B. A., Manion G., Moritz C., Nakamura M., Nakazawa Y., Overton J. Mc C., Peterson A. T., Phillips S. J., Richardson K. S., Scachetti-Pereira R., Schapire R. E., Soberon J., Williams S., Wisz M. S., Zimmermann N. E., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.
22. Hirzel A. H., Guisan A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157,** 331–341.
23. Poulos S. E., Chronis G. T., Collins M. B., V. Lykousis, 2000. Thermaikos Gulf Coastal System, NW Aegean Sea: an overview of water/sediment fluxes in relation to airland-ocean interactions and human activities. *Journal of Marine Systems*, **25**, 47–76.