Use of GIS, remote sensing and regression models for the identification and forecast of small pelagic fish distribution

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: +302810337822

Abstract

Accurate techniques that are able to identify potential distributions of small pelagic fish in any spatial or temporal scale are essential tools for fisheries management purposes. Additionally, knowledge on small pelagic species distribution could be used for the proper sampling strategy designation and decision making for effective management. In this work, acoustic data corresponding to small pelagic species concentrations are modelled with environmental parameters, to produce predictions of small pelagic species distribution in Thermaikos Gulf, North Aegean Sea. Generalized Additive and Mixed Models are used for the modelling development and prediction of species distribution, while GIS routines are used for the mapping. Specific environmental-fish distribution patterns are also identified and discussed.

Keywords: GIS; remote sensed data; small pelagic species; GAMs; prediction.

1. INTRODUCTION

The prediction of species spatial distribution is an important research theme to a variety of applications in ecology, evolution and conservation science [1]. Many studies are based on models development for the temporal and spatial prediction of species distribution and identification of species-environment relations [2, 3]. GIS and remote sensed data provide essential tools for decision making and management purposes, especially when used with Resource Selection Functions (RSF) [4]. RSFs are statistical models defined to be proportional to the probability of use of a resource unit. Habitat modelling is an applied science and many different techniques have been developed and evaluated for the acquisition of realistic and accurate distribution maps. Among modelling techniques, Generalized Additive Models (GAMs) [5, 6] are perhaps the most common, and well developed and documented [7, 8]. Generalized Additive Mixed Models (GAMMs) are more complicated than GAMs, although they provide specialized approaches to deal with modelling limitations. These tools offer accurate species distribution maps based on sampling areas, as well as predictions of species distribution in wider areas than the sampling, or in a different temporal extend. GAMs and GAMMs are able to identify potential species distribution–environment relations, while the latter could deal with spatial autocorrelation structures in the predicted distribution [9, 10].

In this study, acoustic data have been used in a model development, along with remote sensed environmental data and metadata, derived from GIS techniques. An important number of abiotic variables were grouped, in order to describe potential species distribution–environment associations, and to acquire an accurate model for species distribution prediction. GAMs and GAMMs were used for the model development and the prediction of species distribution, while GIS techniques were used for the mapping of the predicted species distribution. Special issues

concerning modelling techniques and acoustic data such as spatial autocorrelation and model specificity in contradiction to generality are also discussed. Finally, depending on the application of the predicted distribution maps, specific models have been proposed.

2. MATERIALS AND METHODS

2.1 Study area and acoustic data

The study area (Figure 1) is the Thermaikos Gulf, North Aegean, Northeastern Mediterranean Sea.

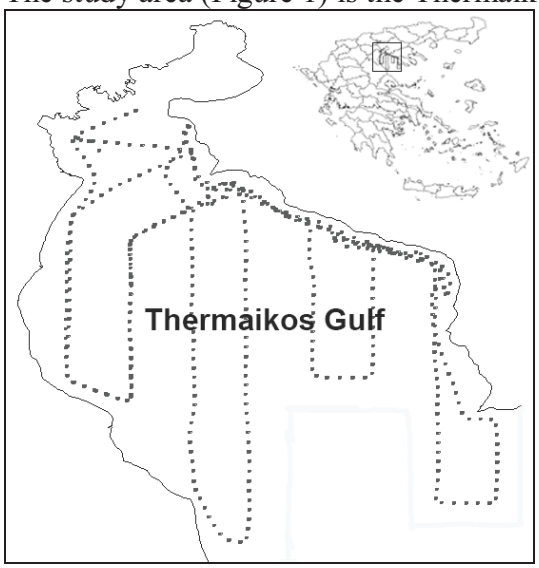

Figure 1. Study area and sampling transects

In this study, acoustic fish density data (Sa: area backscattering coefficient) recorded through SIMRAD EK500/BI500 system on April/May 1998 in Thermaikos Gulf are used. The insonification system, operated at 38 kHz, was calibrated with standard spheres [20]. 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%). Acoustic transects are also shown on Figure 1.

2.2 Remote sensed data

All the available parameters that could explain the variance of acoustic data were used in GAMs development. Two dataset were extracted from each variable. The first represents the sampling points across transects and the other represents each point of the grid that covers the sampling area in a resolution of 0.01 degree (~1km). These variables as well as their sources are shown in Table 1.

Table 1. Remote sensed data, metadata 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 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
Precipitation	PRE	Grid	Mediterranean Oceanic Database (MODB)
Sea Surface Salinity	SSS	Grid / CARTON-GIESE SODA and CMA BCC GODAS models	Mercator operational oceanography

Bathymetry	DEP	Grid / Processed ERS-1, Geostat and historical depth soundings	Laboratory for Satellite Altimetry (NOAA)
Wind stress & direction	WS & WD	Grid & cover	Mercator operational oceanography
Coastline	Coast	Cover / Digitisation of nautical charts and aerial photography	Hellenic Ministry of Environment
Distance to coast	DCoast	Grid & cover	Extracted from coastline
Depth slope	DEPsl	Grid	Extracted from bathymetry grid
Temperature slope (thermal fronts)	SSTsl	Grid	Extracted from SST grid
Marine Productivity Hotspots	MPH	Grid	According to Valavanis et al. 2004
Mesoscale thermal fronts	MTF	Cover	According to Valavanis et al. 2005
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 categorical factor	DN	Cover & grid	Based on sampling date & hour
Date	DT	Cover	Based on sampling date

GIS routines (ArcInfo, version 8.0.2, ESRI) were utilized for the conversion of satellite images into grids and for the extraction of the environmental values at each sampling point. Before the GAM development, an analytic data exploration was performed to acquire a better understanding of the data, and to avoid any assumptions' violation of Generalized Additive Modelling. The basic GAM assumptions consist of outliers, extreme values and collinearity between explanatory variables [11]. Additionally, the exploration process showed potential relationships between variables and prospective variable transformations.

2.3 Model development

Generalized Additive Model is a generalized linear model with a linear predictor, involving a sum of smooth functions of covariates [5, 6]. The main advantage of GAMs over traditional regression methods is their capability to model non-linearities using non-parametric smoothers [5, 6]. GAMs are selected among several resource selection functions, because of their ability to provide biologically interpretable relations between the response and explanatory variables. In addition, Generalized Additive Mixed Models are used complementally to GAMs, in order to deal with spatial autocorrelation. Spatial autocorrelation on acoustic data could lead to biased models and predictions. The total acoustic integration has been used as a response variable, which has been transformed with the natural logarithm. The appropriate transformation method has been selected by using Quantile-Quantile plots (QQ-plots) [12]. The selection of the GAMs' smoothing predictors followed the method proposed by Wood & Augustin [13], using the 'mgcv' library in the R statistical software [14]. The degree of smoothing has also been selected based on the observed data and the Generalized Cross Validation method [6]. The best-fitted model has been selected by using Akaike's Information Criterion (AIC) [15] and a stepwise forward selection method was applied to restrict collinearity among the explanatory variables. The Gaussian family has been selected with identity as a link factor. The variables that did not show significant correlation in the exploration process have been used as explanatory factors. Depending on the corresponding QQ-plots, some of the explanatory variables have also been transformed. The GAMM has been developed based on the final GAM model, by the assumption that a specific correlation structure exists between all points of the study area. This structure has been modelled by using the Gaussian distribution.

2.4 Model comparisons and predictions

The 'predict' function of mgcv library was applied on final models using the R statistical software

[14]. Initially, predictions were acquired from the sampling points by inserting the explanatory variables that were used for models' training. Following that, a new extended dataset containing the models' explanatory variables was used, in order to get a prediction of the entire grid of the sampling area, in greater resolution (0,01 degree). The predictions were inserted in Geographic Information Systems (GIS), where predicted covers and grids were generated. Pearson's correlation was used to compare models' outputs with the initial acoustic data from the sampling points.

3. RESULTS AND DISCUSSION

3.1 Model outputs

The modeling development process ended up with two Generalized Additive Models (GAM8 & GAM83) and one Generalized Additive Mixed Model (GAMM21). Table 2 presents the variables that compose each model and some quality characteristics.

Table 2. Final models and their characteristics. Level of significance was set at 0.05. The ''**:**'' sign denotes interaction. Dev. Exp. = Deviance Explained, Res. d.f = residual d.f., R_a^2 = adjusted R^2, AIC = Akaike Information Criterion value, P-value (chi-square) = significance values, s = denotes smooth function of predictors.

Model's code	Explanatory Variables	Dev. Exp.	Res. d.f.	R_a^2	AIC	p-value
GAM8	s(SLA:CHL) + s(CURSP) + s(PAR) + s(SST)	44,8%	39,786	0.397	494,87	<<0.05
GAM83	s(SLA:CHL) + s(CURSP) + s(PAR) + s(SST) + as.factor(DN)	47.9%	42.795	0.424	477.6857	<<0.05
GAMM21	s(SLA:CHL) + s(CURSP) + s(PAR) + s(SST) + as.factor(DN)	NA	34.034	0.394	570.2206	<<0.05

The decision of using three models, instead of one, to predict species distribution, was based on several model's characteristics that are related to their predictive efficiency, generality and biological interpretability. SLA contributes the most to model's deviance explained, and CURSP, PAR, CHL, SST, DN categorical factor, are following in a descending way. GAM8 is simpler than the others and thus more general in its predictive capacity. The difference between GAM8 and GAM83 is the DN categorical factor, which is used to explain the variance of the data that corresponds to behavioral variation of small pelagic species between day and night. The use of DN factor was based on the life history characteristic of small pelagic species that tend to vertically migrate during the night and surface during the day [18]. The mixed model was developed under the assumption that the sampling data follow a certain correlation structure.

GAMs and GAMMs are able to identify specific relationships, between the response and the explanatory variables [4]. In this case, high acoustic backscattering is related to a combination of high CHL and extreme SLA, average values of CURSP, high PAR and low SST values. The above environmental conditions are generally found in upwelling areas, where it is well documented that small pelagic species are concentrated.

An overview of models' characteristics indicates that all models are statistically significant. According to AIC (the lower the better), GAM83 overcomes GAM8 and GAMM21 in fitting the training data. Additionally, the deviance explained (not available in GAMM21) and the adjusted R^2, suggested that GAM83 explains higher proportion of the response variance, than the others. In conclusion, GAM83 performs better in underlying the relationship between acoustic data and environmental factors.

3.2 Predictions in and outside the sampling area

Each model was trained by using the 442 points of sampled data. In Figure 2, predictions referring to acoustic information distribution are presented. Each map includes the output of predicted values

based on sampling points (pie charts), and those based on the whole grid (grayscale area). In each pie chart, sampling (white) and predicted (black) values of each model are also depicted. White patches on the grid correspond to high acoustic information, which indicate high species concentration. ArcInfo GIS GRID software version 8.0.2 was used [19] for the generation of prediction maps.

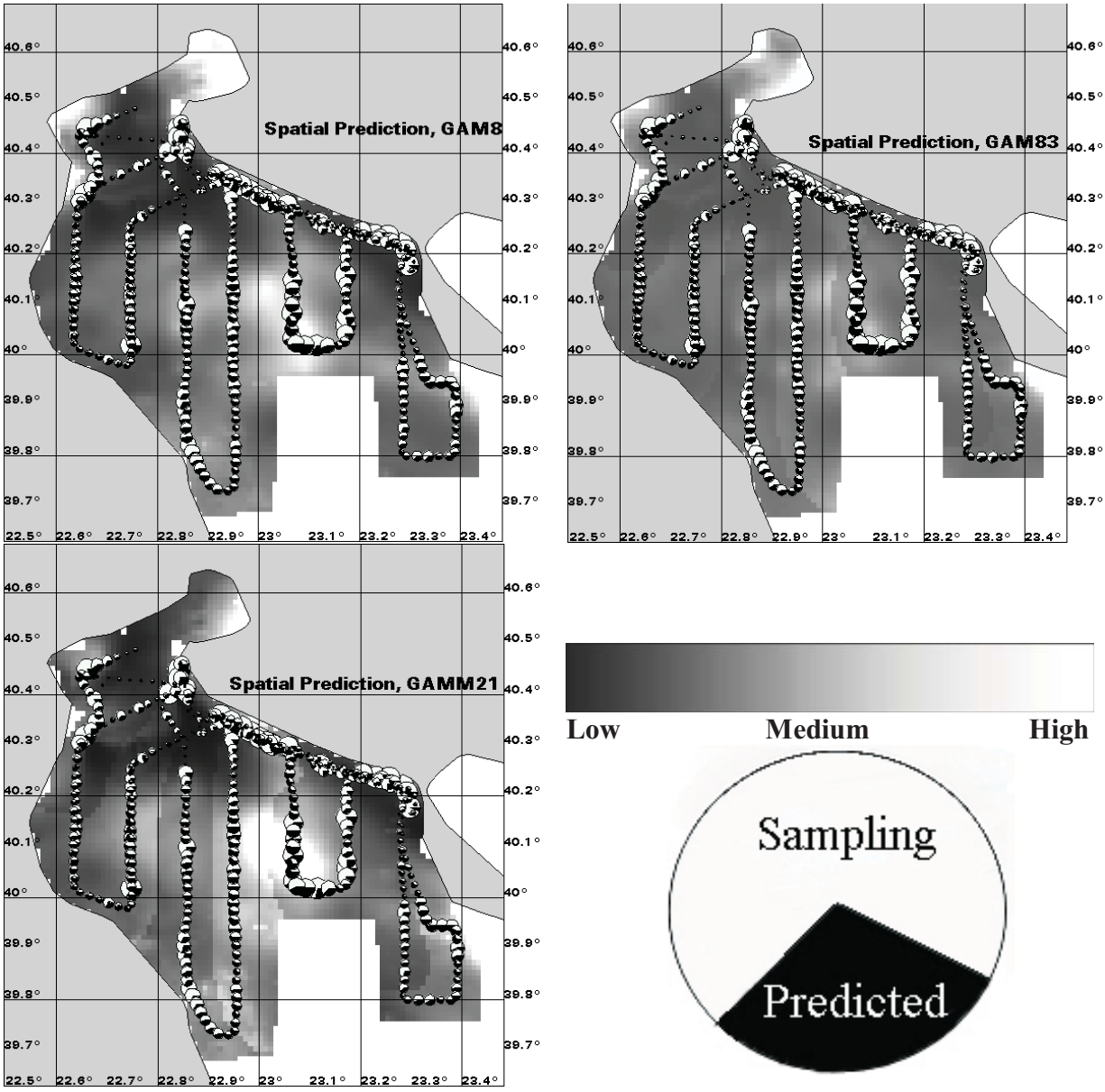

Figure 2. Predicted maps and pie charts presenting prediction-Sa portion of sampling points

The grid predictions reveal common distinct areas, where high acoustic information is concentrated, although the extents of these areas vary among the models. In GAMM21 and GAM83 models, that include the DN factor, an obvious stratification of the predicted values is observed, derived from the impact of that factor in the models. Even if DN contributes to the better understanding of the acoustic variance, it seems to generate an artifact, when applied to the predicted acoustic distribution maps.

3.3 Model evaluation and comparison

Pearson's correlation coefficients among the acoustic measurements and the predicted values of the sampling points are shown in Table 3, while Table 4 contains the correlations between the grid predictions.

Table 3. Pearson's Correlation Coefficients among point predictions and Sa

Correlation	Sa	GAM8	GAM83	GAMM21
Sa	1	0.459	0.494	0.442
GAM8	0.459	1	0.954	0.926
GAM83	0.494	0.954	1	0.933
GAMM21	0.442	0.926	0.933	1

Table 4. Pearson's Correlation Coefficients for among grid predictions

Correlation	GAM8	GAM83	GAMM21
GAM8	1	0.783	0.689
GAM83	0.783	1	0.359
GAMM21	0.689	0.359	1

A visualization of Table 3 is also presented per sampling point in Figure 2, by the use of pie charts. The correlation coefficients confirm that the model with the best quality characteristics (AIC, deviance explained), correlated more with the initial acoustic data (Sa). On the other hand, GAMM21 has the lowest correlation coefficient. With respect to the predicted grids, there is no reference variable and thus, the correlations do not demonstrate predicted capacity, but relations among the grids. In Table 3 all predicted values are highly correlated (>0.92), although in grid predictions (Table 4) there is a significant variance between the correlation coefficient. GAMM21, which includes the spatial autocorrelation pattern, differs from GAM83, but not from GAM8. This could be an indication of a substantial spatial autocorrelation pattern, inserted by the DN factor and modeled in GAMM21. As mentioned above, the DN factor complicates the predicted grids, though the correlations between GAM8 and the models including DN are relatively high (0.783 & 0.689). The DN factor contributes significantly to GAM83 and GAMM21, increasing the deviance explained by these models, however DN is decreasing models' generality and prediction capacity.

Species distribution maps could be valuable for several decision-making and management purposes [4], but their accuracy will always be controversial. Additionally, a prospective variation of any environmental factor can be used to identify changes of species distribution. Generally, species distribution modelling is only as good as the data used [16]. In this case, three models have been developed using similar techniques produced acoustic distributions that identify quite comparable distribution maps, but not identical. The choice of the appropriate distribution map depends basically on its respective application. Generality, reality and precision are the features that group modeling techniques and only two out of the three can be achieved by a model each time [17]. Although GAM8 is a general model and could be used in a wide range of spatial and temporal predictions, GAM83 is the model that describes more accurately the variance of the acoustic data and it is more precise. On the other hand, GAMM21 is the only model that deals with spatial autocorrelation issues [9], which insert bias to modeling processes of acoustic data, even if it does not perform equally to GAM83.

4. CONCLUSIONS

In this study, remote sensing data and metadata have been modeled with acoustic data using GAMs and GAMMs, in order to identify and predict small pelagic fish distribution in Thermaikos Gulf. Several GIS routines have also been used for the acquisition of several metadata sets and prediction maps. Our results indicate that high acoustic information is related to a combination of high Chlorophyll-a, and extreme Sea Level Anomaly, average values of Current Speed, high Photosynthetically Available Radiation and low Sea Surface Temperature values. Three models have been developed and used for predictions, GAM83 is more precise in the simulation of the data, GAM8 is more general (thus suitable for prediction), and GAMM21 incorporates data's spatial autocorrelation. However, the choice of the appropriate model depends on its relative application.

References

1. 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.McC., 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.
2. 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.
3. Guisan A., Edwards J., Thomas C., Hastie T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100.
4. Valavanis V.D., Pierce G.J., Zuur A.F., Palialexis A., Saveliev A., Katara I., Wang J., 2008. Modelling of essential fish habitat based on remote sensing, spatial analysis and GIS. *Hydrobiologia*, **612 (1)**, 5-20.
5. Hastie T., Tibshirani R., 1990. *Generalized additive models*. Chapman and Hall, London.
6. Wood S. N., 2006. *Generalized Additive Models: An Introduction with R*. CRC Press, London.
7. Leathwick J.R., Elith J., Hastie T., 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, **199**, 188–196.
8. Lehmann A., J.Mc, Overton C., Leathwick J. R., 2002. GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling*, **157**, 189–207.
9. Keitt T.H., Bjornstad O.N., Dixon P.M., Citron-Pousty S., 2002. Accounting for spatial pattern when modelling organism–environment interactions. *Ecography*, **25**, 616–625.
10. Wagner H.H., Fortin M. J., 2005. Spatial analysis of landscapes: concepts and statistics. *Ecology*, **86,** 1975–1987.
11. Zuur A.F., Ieno E.N., Smith G.M., 2007. *Analysing Ecological Data*. Springer Series: Statistics for Biology and Health.
12. Cleveland W.S., 1994. *The Elements of Graphing Data*, Hobart Press ISBN 0-9634884-1-4
13. Wood S. N., Augustin N. H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling* **157**, 157–177.
14. R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.Rproject.Org
15. Akaike H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**,716–723.
16. Hirzel A.H., Guisan A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157,** 331–341.
17. Levins R., 1966. The Strategy of Model Building in Population Biology. *American Scientist* **54**, 421-431.
18. Giannoulaki M.**,** Machias A., Tsimenides N., 1999. Ambient luminance and vertical migration of the sardine *Sardina pilchardus*. *Marine Ecology Progress Series*, **178**, 29-38.
19. ESRI, 1994. ARC Macro Language. Environmental Systems Research Institute Inc., Redlands CA, USA: 3–37.
20. Foote K.G., 1987. Fish target strengths for use in echo integrator surveys. *Journal of the Acoustical Society of America*, **82**, 981–987.