Vol. 255: 249–258, 2003 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published June 24

Predicting demersal fish species distributions in the Mediterranean Sea using artificial neural networks

Christos D. Maravelias[1,*], John Haralabous[2], Costas Papaconstantinou[3]

[1]University of Thessaly, Dept. of Agriculture, Animal Production and Aquatic Environment, Fitoko Street, 38446 Magnesia, Greece
[2]Institute of Marine Biology of Crete, PO Box 2214, 71003 Crete, Greece
[3]National Centre for Marine Research, Agios Kosmas, 16604 Athens, Greece

ABSTRACT: Predicting the occurrence of economically important demersal fish in a multispecies marine environment can be of considerable value to fisheries management and protection of biodiversity. Here, 2 predictive modelling principles were utilised, artificial neural network (ANN) and discriminant function analysis (DFA), to develop presence/absence models for 3 species (anglerfish *Lophius budegassa*; hake *Merluccius merluccius*; red mullet *Mullus barbatus*) in the Mediterranean Sea. ANN-based models of demersal fish distribution outperformed conventional models and attained better recognition and prediction performance. Results indicated the ability of ANN's to predict presence more accurately than DFA when tested against independent field data. More precisely, sensitivity values obtained using DFA were 62.1% for anglerfish, 5.8% for hake and 59.8% for red mullet whereas using ANN were 75, 71 and 72.9% respectively. The accuracy of test data was 79.6% for anglerfish, 49.5% for hake and 83.3% for red mullet using DFA and 83.7, 83.3 and 85.6% respectively using a back-propagation ANN. After learning from a set of selected patterns, the neural network (NN) models displayed a relatively high demersal fish classification accuracy, which was consistent with present understanding of the aggregating effects of the examined variables on these species' distribution. Predicting presence or absence was found to be easier for red mullet and anglerfish than for hake. The present results also suggested that the main processes modulating the occurrence of anglerfish, hake and red mullet in the NE Mediterranean Sea can be approximated by linear functions only to a limited extent. Due to their ability to mimic non-linear systems, ANNs proved far more effective in modelling the distribution of these species in the marine ecosystem. The main results and the ANN potential to predict suitable habitat profiles and structural characteristics of species assemblages are discussed.

KEY WORDS: ANN · Anglerfish · Hake · Red mullet

INTRODUCTION

Increasing focus on global and regional patterns of biodiversity necessitates reliable models of species presence/absence. In a multispecies fishery, such as in the Mediterranean Sea, binary classification methods of economically important demersal fish, based on commonly measured quantitative biotic and abiotic factors, are of major significance due to the role certain fish species play in conduct of research and practice of commercial fisheries. Likewise, reliable habitat models of target species may help conservation planning by reducing by-catch of non-target species; thus, protection of biodiversity may be an added benefit.

In fisheries, a wide range of multivariate techniques have been used to this end, including several methods of ordination and canonical analysis, univariate and multivariate linear, curvilinear and logistic regressions (Mastrorillo et al. 1997). Most of the models used to predict species distribution assume that relationships are smooth, continuous and either linear or simple polynomials (Shepherd et al. 1984, James & McCulloch 1990, Mann 1993). However, in real nature, any changes in distributional boundaries or location cen-

*Email: cmaravel@uth.gr

tres of a fish stock are unlikely to be either monotonic or linear. Traditional methods of statistical analysis (namely linear regression models, multiple or not) may therefore be inadequate for detecting and successfully quantifying such changes (Maravelias & Reid 1997).

This work utilises the ability of artificial neural networks (ANNs) to recognise and learn the complex non-monotonic and non-linear relationships between biotic and abiotic aspects of the marine environment that can be used to correctly predict the presence or absence of demersal fish species. Recently, neural networks (ANNs) have been used in various disciplines of aquatic ecology, e.g. fish school species classification (Haralabous & Georgakarakos 1996), prediction of phytoplankton production (Scardi 1996) and freshwater fish biomass (e.g. Baran et al. 1996, Lek et al. 1996, Mastrorillo et al. 1997, Brosse et al. 1999a,b). ANN models were initially intended to mimic the neural activity in the human or animal brains (Garson 1991, Goh 1995, Stern 1996). ANNs are a form of artificial intelligence that is composed of a network of connected nodes (Rumelhart et al. 1986). Density estimation (also referred to as 'unsupervised learning'), classification and regression (both often referred to as 'supervised learning') are 3 broad types of statistical problems that can be successfully modelled by ANNs. Based on a source of training data, the aim of supervised learning is to produce a model of the process from which the data were generated to allow the best predictions to be made for new data.

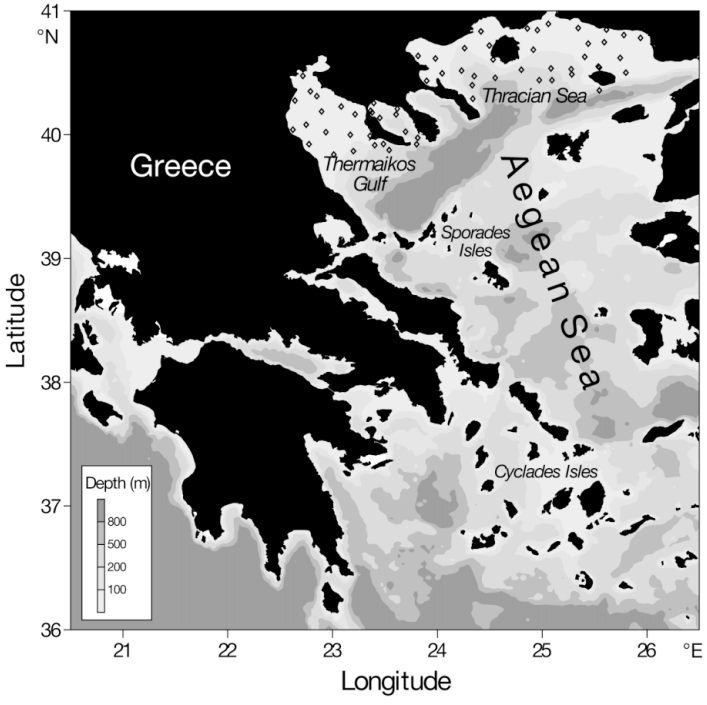

Fig. 1. Studied area. Sampling stations (◇)

The most commonly used ANNs with supervised learning are the multilayered perceptrons also known as 'back-propagation' ANNs after their training algorithm (Rumelhart et al. 1986). Such ANNs have the ability to learn patterns of relationships in data from being shown a given set of inputs (including combinations of descriptive and quantitative data), generalise or abstract results from imperfect data, and be insensitive to minor variations in input (such as noise in the data, missing data or a few incorrect values). ANNs model the physical environment (habitat) system on the basis of a set of hidden input/output examples, as is available in existing fisheries data, without any prior knowledge or assumptions about the underlying distribution function. General references to ANNs can be found in Rumelhart et al. (1986), Garson (1991), Ripley (1994), Goh (1995) and Stern (1996).

The goal of the present study was to determine the predictive capacity of ANN models for estimating presence/absence of 3 commercially important fishes, the European hake *Merluccius merluccius*, the red mullet *Mullus barbatus* and the anglerfish *Lophius budegassa*, from 5 predictor variables (biomass/abundance ratio, depth of the water column, geographical position, i.e. latitude and longitude, and sampling month) in the NE Mediterranean Sea. The objective was to learn more about the factors that might modulate the spatio-temporal aggregation patterns of these heavily exploited demersal species. Finally, the application of non-linear ANNs to empirical modelling was compared with a conventional linear approach, i.e. discriminant function analysis (DFA).

MATERIALS AND METHODS

The fish data for all 3 species examined were collected in the north Aegean Sea (NE Mediterranean) and more precisely in the Thermaikos Gulf and Thracian Sea regions (Fig. 1). Sampling was performed on a seasonal basis, i.e. every 3 mo, from April 1996 to January 1998 using experimental trawl surveys. Trawls had a cod-end mesh size of 14 mm from knot to knot. Sampling took place only during daylight hours. The duration of each haul was 1 h with a vessel speed of 2.5 knots. Since in every survey the boat was at sea for the same length of time, from 06:00 to 18:00 h, the number of individuals caught by trawling hour and their corresponding weight were considered as units of relative abundance and biomass, respectively. A total of 675 stations were sampled, covering a depth range from 30 to 500 m and the sampling design adopted was simple random. The ratio of the natural logarithms of biomass against abundance (B/A ratio), both incremented by one (i.e. data + 1) was used as a biological

index for each of the 3 studied species. Low values of the B/A ratio in a specific location (i.e. station) suggest relatively higher abundance and lower biomass values of the examined species in that sampling station. On the contrary, high B/A ratio values indicate relatively higher biomass and lower abundance values. All specimens caught were measured recording total length to the nearest mm and weighed to the nearest g.

In the present study, an advanced (i.e. ANN) and a conventional (i.e. DFA) discrimination technique were implemented. Both methods were employed on a random subset of available data (training set) and then applied to the remaining data (validation or testing set). The training set consisted of random 4/5 of the available data set (540 cases) with the remaining random 1/5 data (135 cases) to consist the testing set. This holdout partitioning technique (Kohavi 1995) was repeated 5 times to statistically compare the methods' ability to correctly predict unknown cases. The following performance criteria were evaluated: (1) sensitivity, i.e. the percentage of true presences correctly identified; (2) specificity, i.e. the percentage of true absences correctly identified; and (3) accuracy, i.e. the total fraction of the sample correctly identified. When applied to training sets, the accuracy provides a measure of the recognition performance; whereas, when applied to testing sets, it gives a measure of prediction performance. Since a large number of cases (20%) was used for testing, each iteration could not be considered an independent replicate because it was likely that many of the same data had been used in each iteration. Because of that and in order to relax any distributional assumptions of the studied variables, a non-parametric approach was used to compare the 2 methods, i.e. the Wilcoxon signed rank test.

A schematic representation of a typical ANN's structure is given in Fig. 2. It consisted of 3 interconnected layers of 'nodes' or 'neurons': an 'input layer' containing 1 neuron per independent variable, an intermediate 'hidden layer' whose number of neurons optimised performance (Geman et al. 1992, Mastrorillo et al. 1997) and finally, an 'output layer' with 2 nodes (presence and absence). An empirical approach was employed to determine the network's configuration in the present study. Allowing the number of hidden nodes to vary from one half to the double of the number of input neurons (Scardi 2001) identified a 6-hidden neurons' configuration to be optimal. The neurons were connected to the next layer neurons with adjustable weights. When a neuron received input signals from others, these were weighted by different values and summed. Then, the neuron output the signal according to a nonlinear sigmoid function, $F(x) = 1/(1 + \exp[-x])$. The training of ANNs consisted in iteratively adjusting the weights in order to minimise the prediction error, i.e. difference between observed and predicted values, and thus, obtaining the maximum number of correctly classified cases. A form of gradient descent algorithm, i.e. the error back-propagation (EBP) algorithm, was used for that purpose (Rumelhart et al. 1986). The EBP requires the specification of the search step size (the learning rate); here, a unit learning rate was used in order to allow for a valid comparison between models. The 'momentum' term, an additional parameter optional for accelerating model's convergence, was set to 0.1. A random submission of input cases at each iteration (learning epoch) reduced the risk of memorisation of the presentation order of the training cases. Here, the early stopping strategy was followed to avoid overfitting and thus, to obtain a generalised model (Scardi 1996, 2001). This procedure consisted of terminating the training phase when the prediction error in the testing set (i.e. validation error) began to increase. Several authors have proposed methods allowing the determination of the impact of the ANN input variables (Garson 1991, Goh 1995, Lek et al. 1996) in a similar manner that DFA identifies the contribution of each studied variable in determining the output. Here, the Garson's algorithm modified by Goh (1995) was used to determine the relative importance of variables examined.

In DFA, linear combinations of the predictor variables were formed to produce discriminant scores. These served as the basis for assigning cases into 1 of the 2 groups (presence/absence). The discriminant function coefficients were estimated in such a way so as to maximise the ratio of the between groups to the

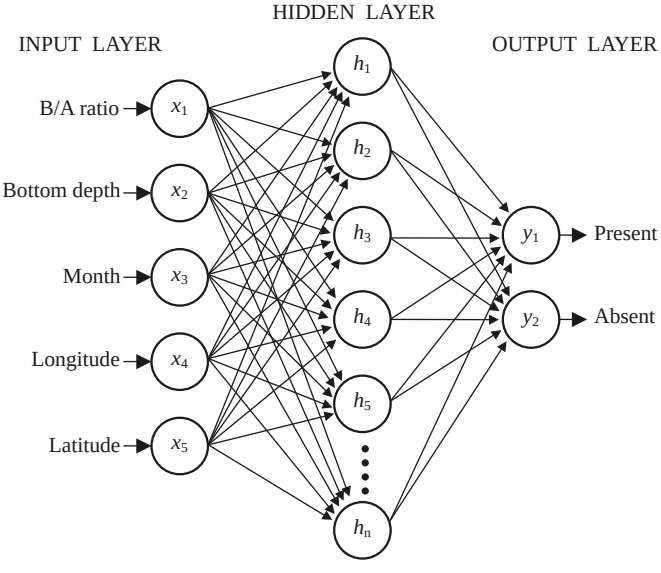

Fig. 2. An artificial neural network (ANN) consisted of an input layer with 5 nodes (predictor variables), a hidden layer and an output layer with 1 node per group (presence/absence) to be predicted

within groups sum of squares. When the covariance matrices of the 2 groups were equal, as observed for anglerfish and red mullet, the linear discriminant function was found to be adequate; while when the covariance matrices were unequal (hake), the quadratic discriminant function was preferred (Wahl & Kronmal 1977). The Box's *M*-test for equality of the group covariance matrices was used.

The effect of 5 predictors, i.e. sampling month, bottom depth, longitude, latitude and B/A ratio, on the presence/absence of each of the 3 studied species was also evaluated using these 2 modelling approaches, i.e. ANN and DFA. Finally, 2-dimensional probability density maps were produced in order to visualise the ANN-derived effect of the predictor variables on species' presence. The present analysis was performed using Splus 2000 (MathSoft), NeuroShell II (Ward Systems) and SPSS 9 (SPSS) software.

RESULTS

In Table 1 the classification results of DFA and ANN are given for testing sets only. In the training cases (not shown), ANN constantly performed better than DFA in terms of all 3 criteria, i.e. sensitivity, specificity and recognition performance. All differences observed between ANN and DFA in the training cases were found to be statistically significant (Wilcoxon, $p < 0.05$).

In the random subsamples of the data used as testing sets, ANN were found to be more stable than DFA in terms of sensitivity. As shown in Table 1, the coefficient of variation for sensitivity ranged in ANN between 1.4 and 8.4 whereas in DFA between 8.6 and 23.7. It was further observed that ANNs' ability to predict presence accurately was always significantly higher than that of DFA (Wilcoxon, $p < 0.05$). A sensitivity value of 75% was attained for anglerfish using the ANN and 62.1% using the DFA model (Table 1). The sensitivity value obtained for red mullet was slightly lower, being 72.9% in ANN and 59.8% in DFA. A noteworthy difference between the sensitivity values of ANN (71%) and DFA (5.8%) for hake species was detected.

With regard to the predictability of true absences (specificity) in testing cases, the differences observed between the 2 methods were significant only for hake (Wilcoxon, $p < 0.05$).

In the testing sets, the prediction performance for red mullet was the highest, being 85.6% in ANN and 83.3% in DFA (Table 1). Hake had the worst prediction performance 83.3% for ANN and 49.5% for DFA.

Table 1. Testing results of the DFA and ANN models developed on species' presence/absence prediction. Sensitivity, specificity and accuracy values are expressed in percentages. SD: standard deviation; CV: coefficient of variation. Asterisk (*) in an ANN mean value indicates significant difference from the corresponding DFA value

Species	Sample	DFA			ANN		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Lophius budegassa	1	66.7	88.2	81.5	71.4	90.3	84.4
	2	78.9	83.5	82.2	84.2	85.6	85.2
	3	56.9	95.2	80.7	76.5	89.3	84.4
	4	61.2	89.5	79.3	75.5	89.5	84.4
	5	46.9	89.5	74.1	67.3	87.2	80.0
	Mean	62.1	89.2	79.6	75.0*	88.4	83.7*
	SD	11.9	4.2	3.3	6.3	2.0	2.1
	CV	19.1	4.7	4.1	8.4	2.2	2.5
Merluccius merluccius	1	7.8	83.3	54.8	76.5	92.9	86.7
	2	4.8	83.6	47.4	72.6	89.0	81.5
	3	6.3	74.7	50.4	68.8	89.7	82.2
	4	5.6	81.5	51.1	72.2	91.4	83.7
	5	4.3	64.0	43.7	65.2	91.0	82.2
	Mean	5.8	77.4	49.5	71.0*	90.8*	83.3*
	SD	1.4	8.3	4.2	4.3	1.5	2.1
	CV	23.7	10.7	8.4	6.0	1.7	2.5
Mullus barbatus	1	64.3	92.5	83.7	73.8	92.5	86.7
	2	51.4	93.0	82.2	74.3	89.0	85.2
	3	58.3	88.9	80.7	72.2	90.9	85.9
	4	62.5	93.2	85.9	71.9	90.3	85.9
	5	62.5	92.6	83.7	72.5	89.5	84.4
	Mean	59.8	92.0	83.3	72.9*	90.4	85.6
	SD	5.2	1.8	1.9	1.1	1.4	0.8
	CV	8.6	1.9	2.3	1.4	1.5	1.0

Anglerfish gave intermediate results, with 83.7% prediction performance for ANN and 79.6% for DFA. These higher accuracies of ANN models differed significantly from those of DFA for anglerfish and hake (Wilcoxon, $p < 0.05$) but not for red mullet (Wilcoxon, $p = 0.225$).

The discriminant grouping patterns of DFA for each species are illustrated in Fig. 3. The squared distances between presence and absence were higher for red mullet (2.56) than for anglerfish (1.69) and hake (0.09). Thus, discrimination was easy for red mullet, adequate for anglerfish and poor for hake. Graphical comparisons of DFA and ANN probability distributions outputs for each species, summarised how the presence/absence discrimination was better in the neural network method than in DFA (Fig. 4).

The relative impact of the 5 predictor variables is given in Fig. 5. The B/A ratio and water depth had the highest contribution in predicting species' presence/absence in both analyses. All input variables had a contribution to the ANN models. On the contrary, DFA models have selected mainly 2 input variables. Specifically in ANNs, the B/A ratio was most strongly associated with all 3 species, whereas in DFA, water depth had the highest impact for hake. The other 3 input variables (i.e. longitude, month and latitude) in DFA

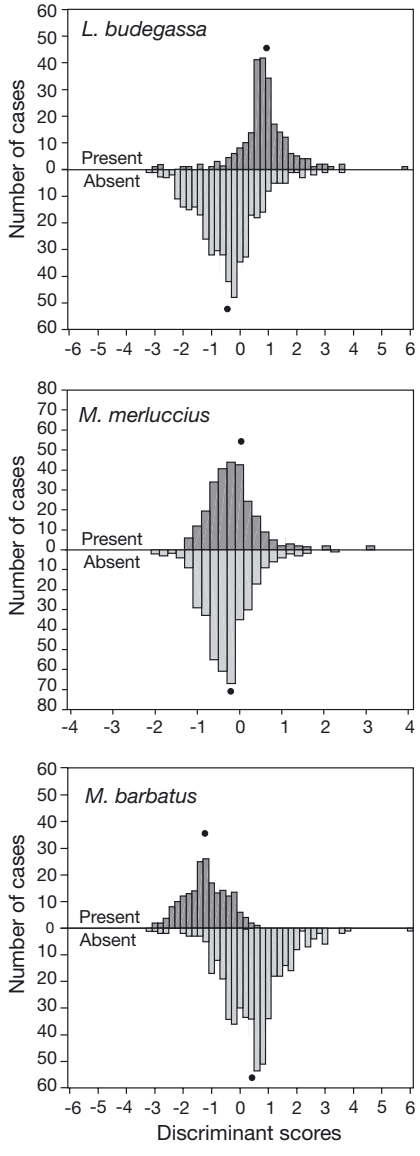

Fig. 3. Discriminant grouping patterns observed for each species. Values on the *x*-axis are discriminant scores obtained by the discriminant function analysis (DFA) and values on the *y*-axis are frequencies of cases. Dark grey bars indicate points where a species was present and light grey bars where it was absent. Group mean values (•: centroids)

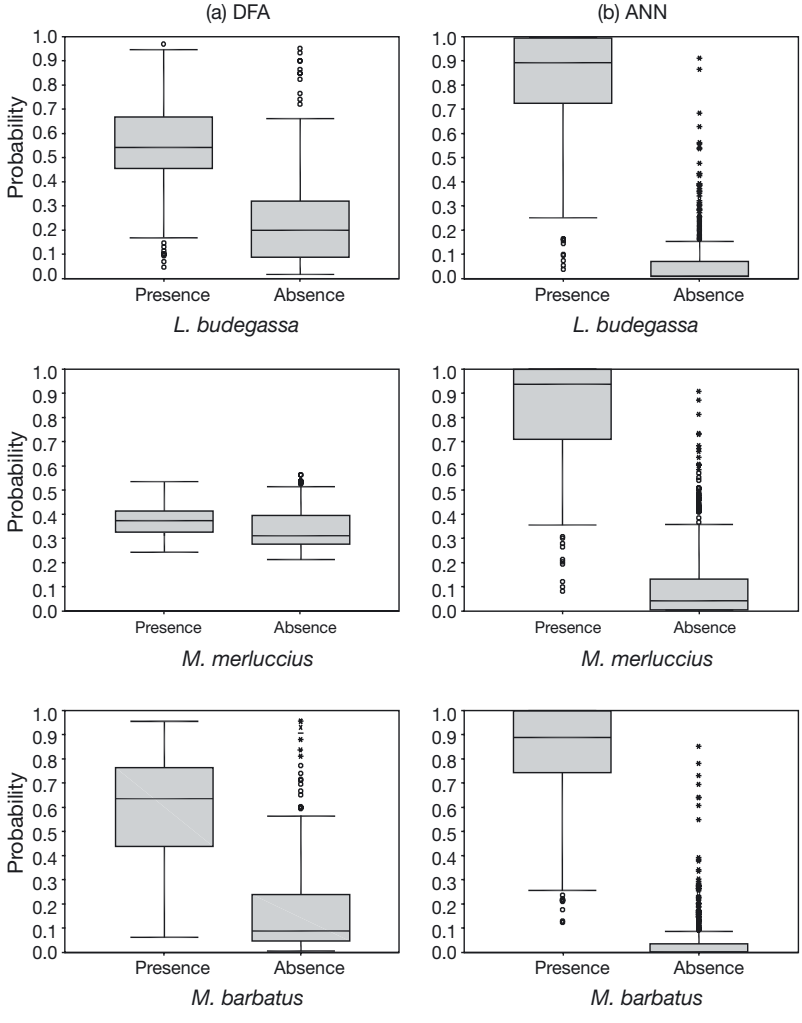

Fig. 4. Box plots of prediction probability by (a) discriminant function analysis (DFA) and (b) artificial neural network (ANN) for each species

contributed little to species' prediction. In ANNs, the relative impact of these 3 variables ranged between 12 and 17%.

Two-dimensional probability density maps of each species' presence were used to visualise the estimated effect of the predictor variables obtained using ANN models (Figs. 6, 7 & 8). On each of these figures, the response of individual species' presence to 2 input variables is presented. Each pixel represents the predicted median value which resulted from the 5 runs of ANN models (Sets 1 to 5). The *y*-axis was fixed to demonstrate the biotic information available, i.e. the B/A ratio. The *x*-axis represented 1 of the 4 predictive variables studied each time.

The probability of finding the anglerfish present was highest in waters up to 200 m, approximately carrying lower B/A values and in deeper waters (>250 m) having larger B/A values (Fig. 6a). With regard to the horizontal spatial dimension and for intermediate B/A values, anglerfish were likely to be found throughout

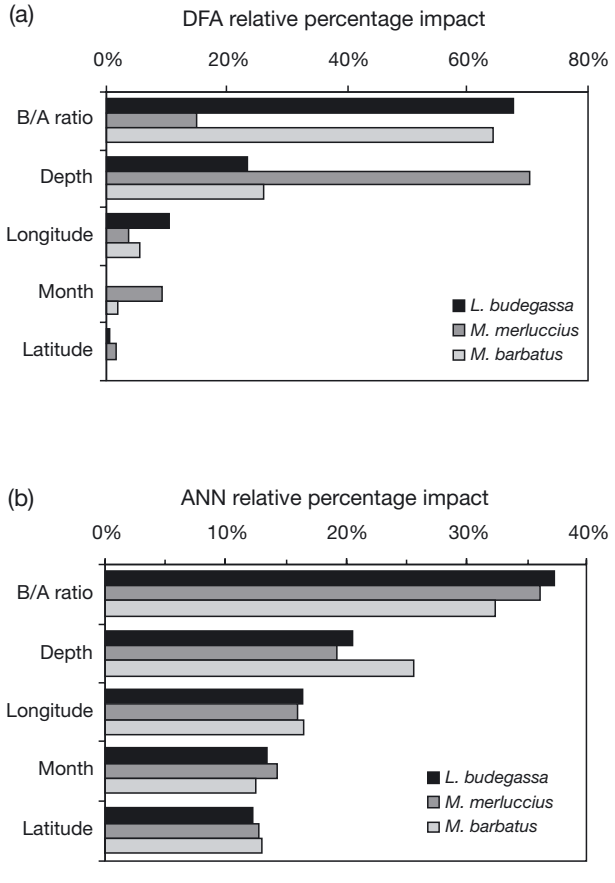

Fig. 5. Percentage contribution of each of the 5 variables to species presence/absence prediction. (a) Discriminant function analysis' (DFA) standardised canonical function coefficients expressed in relative form. (b) Artificial neural network's (ANN) impact factors obtained using Garson's (1991) algorithm

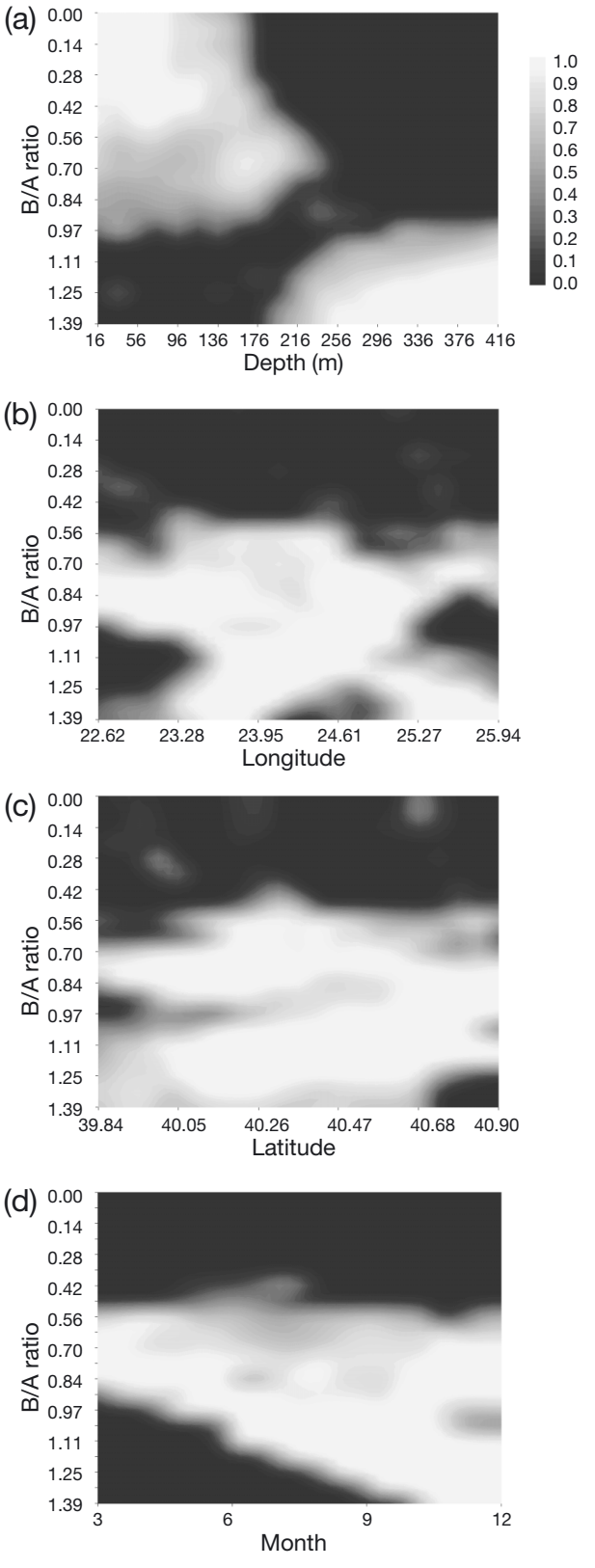

Fig. 6. *Lophius budegassa*. Artificial neural networks (ANN) estimated probability of *L. budegassa* presence as a function of predictor variables

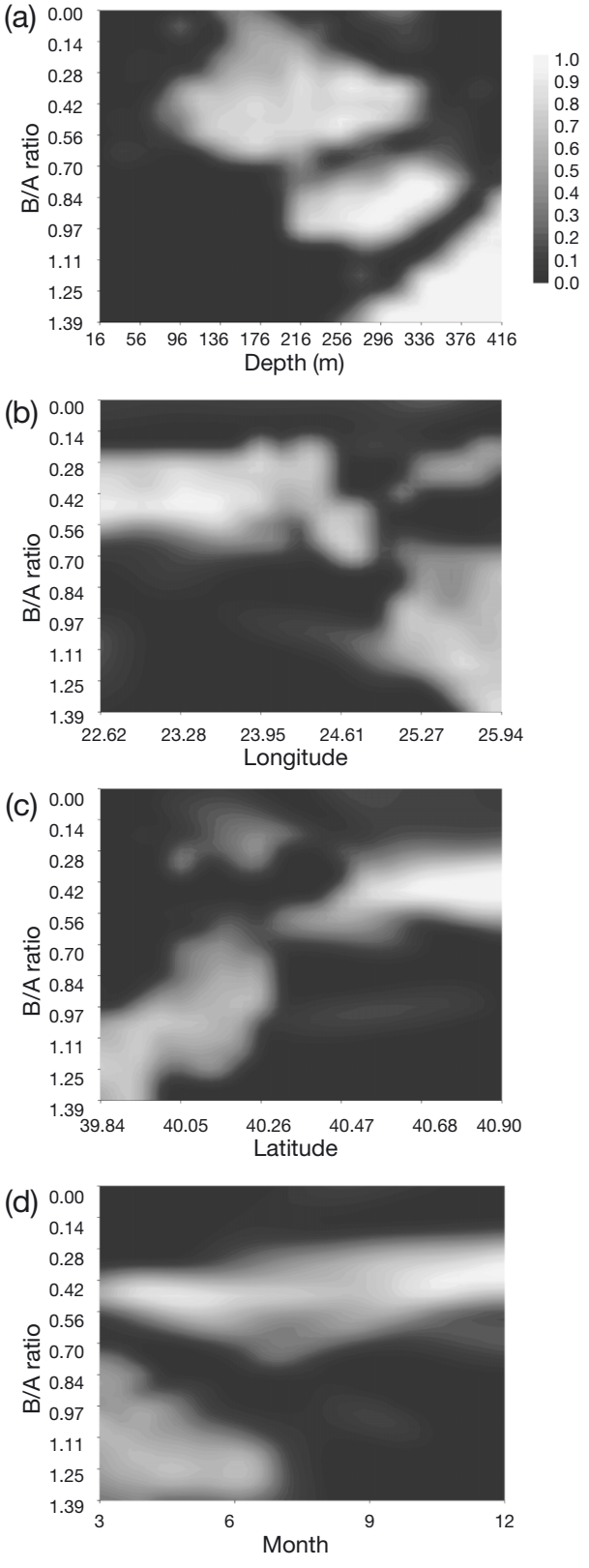

Fig. 7. *Merluccius merluccius.* Artificial neural networks (ANN) estimated probability of *M. merluccius* presence as a function of predictor variables

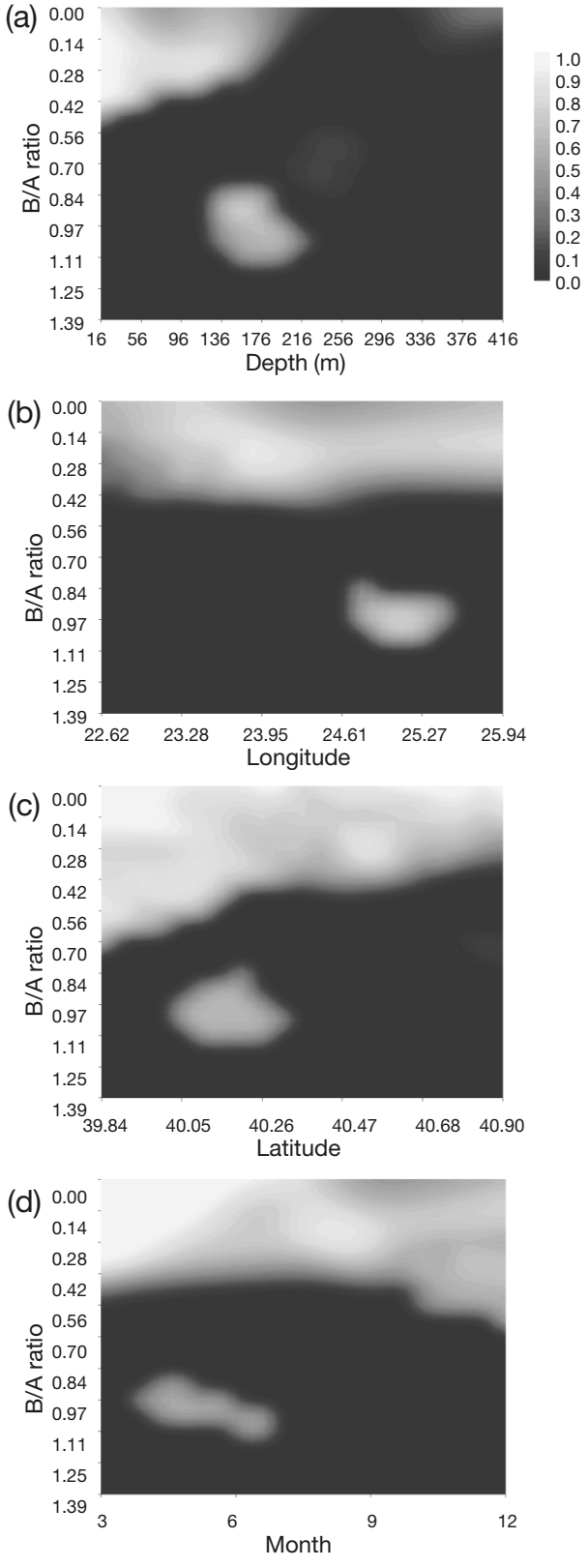

Fig. 8. *Mullus barbatus.* Artificial neural networks (ANN) estimated probability of *M. barbatus* presence as a function of predictor variables

the studied area (Fig. 6b). Areas with longitude values between 23.28 to 24.13° and 25.15 to 25.94° E with the largest B/A values also had a high probability of anglerfish encounter (Fig. 6b). Anglerfish distribution was not confined in a north-south geographical dimension, at least for intermediate and high B/A values (Fig. 6c). There was a high probability of anglerfish' presence in the first half of the year when B/A values were intermediate as well as in the latter half of the year, when B/A values were intermediate or the highest (Fig. 6d).

The probability density maps of hake were relatively hazier compared with those of the other 2 species examined (Fig. 7a–d). During these surveys, the probability of finding hake present was higher for the largest B/A values found in deeper waters (Fig. 7a). Throughout the studied area, hake were also likely to be encountered when B/A values were intermediate, with the exception of shallower waters (<90 m). The probability of hake being present was highest in the western part of the area and for intermediate B/A values; however, it was also high in the eastern part, carrying the largest B/A values (Fig. 7b). Concerning the north-south geographic dimension, hake encounter was less probable moving from north to south and from intermediate to high B/A values (Fig. 7c). There was no obvious seasonal variation in the probability of hake being present, al least for intermediate B/A values (Fig. 7d). The first half of the year had a notable probability of encountering high B/A values of hake.

Deeper waters had the lowest probabilities of finding red mullet present (Fig. 8a). There was a higher probability of red mullet being present in shallower water (up to 170 m) areas, carrying low B/A values in the southern than in the northern part of the study area (Fig. 8a,c). Red mullet also had a high probability of presence in lower B/A values throughout the year and the longitude range of these surveys (Fig. 8b,d).

DISCUSSION

The use of statistical models to predict the likely occurrence or distribution of fish species is becoming an increasingly important tool in conservation planning and fisheries management. Results from the present work suggested that the main processes modulating the occurrence of anglerfish, hake and red mullet in the NE Mediterranean Sea can be approximated by linear functions only to a limited extent. Due to their ability to mimic non-linear systems, ANNs proved far more effective in modelling the distribution of these species in the marine ecosystem. According to the current findings, ANN-based models of demersal fish distribution attained better recognition and prediction performance, and thus outperformed conventional models (i.e. DFA). Although the same order of performance was observed with DFA and ANN for the 3 species, ANN gave consistently the highest scores.

An important property of successful presence/absence ecological models when applied to independent datasets is their ability to predict presence accurately. Hence, sensitivity is considered of primary importance compared to specificity (and overall accuracy) since the latter might suffer from the species' prevalence effect, i.e. frequency of occurrence. This effect will have profound implications whenever the distribution of overexploited or endangered species needs to be predicted for conservation and management purposes; for example, in introducing marine protected areas (Pearce & Ferrier 2000). However, such an effect on predictive power is still to be included in fisheries distribution models. In the current study, true absence outnumbered true presence. The prevalence effect was reflected in both methods' results predicting true absence better than true presence. Furthermore and most importantly, the present results confirmed the ability of neural networks to predict presence more accurately than DFA when tested against independent field data, thus revealing ANN's superiority.

The B/A ratio used in this study may be utilised as a relative index of each species' size distribution in the survey area. A low B/A ratio at a specific location (i.e. station) suggested relatively higher abundance and lower biomass values of the examined species in that sampling station. Likewise, a high B/A ratio indicated relatively higher biomass and lower abundance values. Therefore, it seems reasonable to suggest that in the latter case larger and fewer specimens were present compared with numerous smaller specimens observed in the former case.

Two-dimensional probability density maps of each species' presence enabled the visualisation of the ANN estimated effect of the predictor variables and generated a number of interesting observations. The ANN model displayed a relatively high demersal fish classification accuracy which was consistent with present understanding of the aggregating effects of the examined parameters on these species' distribution. Although a number of studies have been carried out in the NE Mediterranean Sea (Stergiou et al. 1992, Vassilopoulou & Papaconstantinou 1992, Papaconstantinou & Stergiou 1995, Tserpes et al. 1999), information on the ecology of the studied species is sparse due to the analyses concentrating mainly in 1 survey and/or 1 yr and specific regions. To date, the current work represents the most comprehensive study of the relationship between the water depth, spatial location, sampling season and the distributional abundance of 3 economi-

cally important demersal species, covering a wide geographic area, i.e. the NE Mediterranean and is the culmination of a 3 yr study programme. Perhaps more importantly, and unusually for this type of work in the study area, the analysis covers not 1, but 8 consecutive surveys. Predicting the presence or absence was found to be easier for red mullet and anglerfish than for hake. Smaller red mullet were found to be preferentially distributed in shallower waters in the south part of the area throughout the year and avoided waters deeper than 200 m. Larger red mullet were observed in the period between March and June. These results are consistent with prior sporadic information reported in other local studies in the area (Vassilopoulou & Papaconstantinou 1992, Tserpes et al. 1999). Large anglerfish were significantly associated with deeper waters during the autumn-winter period. The present work also provided direct evidence for small anglerfish exhibiting a wide bathymetric distribution from shallow to deeper waters of approximately 300 m depth. The current findings were consonant with anecdotal evidence suggesting that the area contains substantial abundances of young *Lophius*, although this has not been studied quantitatively (Anonymous 1993, C. Papaconstantinou unpubl. data). In common with previous investigations (Anonymous 1993, Papaconstantinou & Stergiou 1995), the present study indicated that large hake were associated with deeper waters in the SE region of the study area during spring and summer. Small and intermediate hake displayed rather diffused distribution patterns with specimens found throughout the survey area. The idea behind the 3 yr study (1996 to 1998) was that a thorough knowledge of the fish distribution in space in relation to environment would help to understand the systems that maintain them. Additionally, it would allow proper modelling of their population dynamics as well as their aggregation at fishing and prespawning grounds. Several impediments, during the surveys however, prevented the collection of additional habitat variables that would have enhanced the ecological reality of the present models. Clearly, there is still a considerable need for additional research that will further clarify the distribution patterns of the studied species.

Although it is possible to predict the likely occurrence of a species at a particular surveyed site using the ANN model, it may not be feasible to do so at an unsurveyed site. This is due to the absence of biomass and abundance data that have been incorporated as predictors in initial model development. The current research surveys have been conducted covering a broad geographical area, i.e. the NE Mediterranean Sea. When the survey's coverage is extensive and sampling is considered to be representative, the biomass and abundance data collected usually include a wide range of values extending from 0 to the highest values. Apart from providing information on the amplitude distribution of each species' data, a further possibility might be to utilise this information to separate the data into categories, e.g. into low, middle and high ranges of values. Thus, it is still possible to predict the probabilities of presence (or absence) of a given species providing the model with an *a priori* threshold value of B/A ratio as input.

One significant advantage of the ANN methodology, as applied in the present study, is that it was trained adaptively for the given input data; thus, the network could be specifically trained to the particular ecosystem of interest. This would help to obtain answers to specific problems to the target ecosystem, as well as to compare them with other ecosystems. Another advantage of the ANN may be that the network is generally more feasible in solving non-linear problems which are essential features of real-life systems. On the other hand, it is not possible to predict an unknown event that has not occurred in training data; therefore, the values of the training data should cover as wide a range as possible.

Results from this study imply that field data may be used to correctly categorise the presence/absence of a species at a site. Although this might seem intuitively realistic to suggest, it is also highly unlikely to be true due to sampling error and its implications. Ecological investigations attempting to correctly categorise presence/ absence of a species at a particular site usually suffer from such sampling error. Several studies have attempted to specifically address this issue. Among them, of particular interest, is that of Scardi (2001). Using constrained training and metamodeling techniques, Scardi (2001) tested the effect of low, medium and high primary production values in neural network development and obtained improved models in cases of limited available data. However, it was beyond the scope of the present study to quantify the implications of sampling error in detail. Notwithstanding this, a key aspect in model development that is related to sampling error and required particular attention in the present work, was that of representative sampling. This was because if the training data are representative, the training data therefore contain the correct proportion of incorrect and correct evaluations, which are used to provide the training set. When training is successful, one would expect that the incorrectly presented data would end up as difficult to separate. Thus, the resulting distributions in the parameter space would overlap. This might be seen in the results as failure to fully train. This would then be carried through to the classification stage with similar errors. When the training set is not representative or is insufficient in number and

the occurrence of training values that are in error do not overlap or overlap enough in the parameter space, this would give rise to the possibility of incorrect training (J. Simmonds, FRS Marine Laboratory, Aberdeen, pers. comm.). Thus, taking representative samples in our analysis was considered crucial to partly address the implications of sampling errors in the results. Obtaining a sufficiently good training set was also regarded as an important issue since it informed us about error not perpetuating it.

In conclusion, ANNs were proved to be more effective in capturing non-linear interactions and attained a better predictive performance than conventional models; thus, they may prove promising in other situations where the association between dependent and independent variables is complex and non-linear. It is more than likely that a further improvement in models' performance may be achieved through incorporation of additional habitat factors not currently included in the analysis, e.g. prey availability, substrate type, temperature profiles. Predicting the occurrence of economically important demersal fish in a multispecies marine environment, such as the Mediterranean Sea, can be of considerable value to the long-term sustainable development of the fishing industry and to the protection of biodiversity. Reliable habitat models can delineate areas of likely occurrence of target species, and that might potentially reduce the by-catch of non-target species. Evidently, such habitat models could be useful tools for decision making and conservation planning.

Acknowledgements. We thank John Simmonds (FRS Marine Laboratory, Aberdeen) and 3 anonymous reviewers for useful comments that improved the manuscript.

LITERATURE CITED

Anonymous (1993) Investigation of the abundance and distribution of the demersal stocks of primary importance to the Greek fishery in the North Aegean Sea. Final report, Contract No DG XIV Fisheries MA-1–90, European Commission, Brussels

Baran P, Lek S, Delacoste M, Belaud A (1996) Stochastic models that predict trout population density or biomass on a mesohabitat scale. Hydrobiologia 337:1–9

Brosse S, Lek S, Dauba F (1999a) Predicting fish distribution in a mesotrophic lake by hydroacoustic survey and artificial neural networks. Limnol Oceanogr 44:1293–1303

Brosse S, Tourenq JN, Lek S (1999b) The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecol Model 120:299–312

Garson GD (1991) Interpreting neural network connection weights. Artif Intellig Expert 6:47–51

Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. Neural Comput 4:1–58

Girosi F, Jones M, Poggio T (1995) Regularization theory and neural networks architectures. Neural Comput 7:219–269

Goh ATC (1995) Back-propagation neural networks for modelling complex systems. Artif Intellig Engin 9:143–151

Haralabous J, Georgakarakos S (1996) Artificial neural networks as a tool for species identification of fish schools. ICES J Mar Sci 53:173–180

James FC, McCulloch CE (1990) Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annu Rev Ecol Syst 21:129–166

Kohavi R (1995) A study of cross-validation and bootstrap for estimation and model selection. In: Mellish CS (ed) Proc 14th Int Joint Conf Artif Intelligence. Morgan Kaufmann Publishers, Berlin, p 1137–1143

Lek S, Belaud A, Baran P, Dimopoulos I, Delacoste M (1996) Role of some environmental variables in trout abundance models using neural networks. Aquat Living Resour 9:23–29

Mann KH (1993) Physical oceanography, food chains, and fish stocks: a review. ICES J Mar Sci 50:105–119

Maravelias CD, Reid DG (1997) Identifying the effects of oceanographic features and zooplankton on prespawning herring abundance using generalized additive models. Mar Ecol Prog Ser 147:1–9

Mastrorillo S, Lek S, Dauba F, Belaud A (1997) The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshw Biol 38:237–246

Papaconstantinou C, Stergiou KI (1995) Biology and fisheries of the eastern Mediterranean hake (*M. merluccius*). In: Alheit J, Pitcher TJ (eds) Hake: biology, fisheries and markets. Chapman & Hall, London, p 149–180

Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecol Model 133:225–245

Ripley BD (1994) Neural networks and related methods for classification. J R Stat Soc B 56:409–456

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. Nature 323: 533–536

Scardi M (1996) Artificial neural networks as empirical models for estimating phytoplankton production. Mar Ecol Prog Ser 139:289–299

Scardi M (2001) Advances in neural network modeling of phytoplankton primary production. Ecol Model 146:33–45

Shepherd JG, Pope JG, Cousens RD (1984) Variations in fish stocks and hypotheses concerning their links with climate. Rapp P-V Reun Cons Int Explor Mer 185:255–267

Stergiou KI, Petrakis G, Papaconstantinou C (1992) The Mullidae (*Mullus barbatus*, *Mullus surmuletus*) fishery in Greek waters, 1964–1986. FAO Fish Rep 477:97–113

Stern HS (1996) Neural networks in applied statistics. Technometrics 38:205–214

Tserpes G, Peristeraki P, Potamias G, Tsimenides N (1999) Species distribution in the southern Aegean Sea based on bottom-trawl surveys. Aquat Living Resour 12:167–175

Vassilopoulou V, Papaconstantinou C (1992) Aspects of the biology and dynamics of red mullet (*Mullus barbatus*) in the Aegean Sea. FAO Fish Rep 477:115–126

Wahl PW, Kronmal RA (1977) Discriminant functions when covariances are unequal and sample sizes are moderate. Biometrics 33:479–484

Editorial responsibility: Otto Kinne (Editor), Oldendorf/Luhe, Germany

Submitted: October 1, 2001; Accepted: January 15, 2003
Proofs received from author(s): April 15, 2003