Hydrobiologia (2008) 612:5–20
DOI 10.1007/s10750-008-9493-y

FISH HABITAT MAPPING

Modelling of essential fish habitat based on remote sensing, spatial analysis and GIS

Vasilis D. Valavanis · Graham J. Pierce · Alain F. Zuur · Andreas Palialexis · Anatoly Saveliev · Isidora Katara · Jianjun Wang

Abstract We review the variety of existing modelling approaches applied to species habitat mapping and we discuss issues arising from the availability and nature of sampled biological data and corresponding ecological and environmental habitat descriptors, as well as the different spatial analysis approaches that are selected according to specific hypotheses. We focus on marine species habitat mapping, presenting an overview of work on modelling fish habitat carried out through a European Communities Policy-Support Action, EnviEFH 'Environmental Approach to Essential Fish Habitat (EFH) Designation' (2005–2008). The selection of the appropriate habitat model is dataset-specific and the resulting EFH maps are often similar in spite of using different models. Derived EFH maps are based on either environmental ranges (used as minimum and maximum environmental habitat descriptors) or probability of occurrence values. We apply model outputs to regions larger than sampled areas making use of the capacity of satellite data to cover wide areas.

Keywords Marine species · Statistical modelling · Fisheries · Environment · Ecology

Guest Editor: V. D. Valavanis
Essential Fish Habitat Mapping in the Mediterranean

V. D. Valavanis (✉) · A. Palialexis
Marine GIS Laboratory, Institute of Marine Biological Resources, Hellenic Centre for Marine Research, Thalassocosmos 71003, Heraklion Crete, Greece
e-mail: vasilis@her.hcmr.gr

G. J. Pierce
Centro Oceanográfico de Vigo, Instituto Español de Oceanografía, P.O. Box 1552, 36200 Vigo, Spain

A. F. Zuur
Highland Statistics Ltd, 6 Laverock road, AB41 6FN Newburgh, UK

A. Saveliev
Faculty of Geography and Ecology, Kazan State University, 8 Kremlyovskaya St., 420008 Kazan, Republic of Tatarstan, Russian Federation

G. J. Pierce · I. Katara · J. Wang
School of Biological Sciences (Zoology), University of Aberdeen, Tillydrone Avenue, AB24 2TX Aberdeen, UK

Introduction

The identification of Essential Fish Habitats (EFH), i.e. areas or volumes of water and bottom substrates that provide the most favourable habitats for fish populations to spawn, feed and mature throughout their full life cycle, is important for the conservation of biodiversity and sustainable fisheries management. The sustainability of fish populations and their associated fisheries could be conserved by limiting anthropogenic stressors in such habitats.

One of the foundational concepts underlying the ecosystem approach to fisheries management

(EAFM) is that different geographic areas have different biological production capacities and that it may be advantageous to focus applying science and management to protect overfished areas and areas of degraded habitats (Lutchman, 2003). EFH analysis should be able to identify those areas within the distribution of a species that contribute most to sustain the long-term viability of a population. Although it may be difficult to define the boundaries of EFH (for example, whether it should be the most important 10% or 15% or 20%, etc, of habitat), the definition of EFH areas, combined with management which recognizes the importance of such areas, represents a first step towards facilitating EAFM concepts and will thus contribute to the sustainability of marine ecosystems and their living marine resources as well as delivering the socioeconomic benefits with a healthy and sustainable fishery.

The extensive spatiotemporal variability, which characterizes dynamic marine ecosystems, presents inherent difficulties for the development of predictive species-habitat models. In order to identify relationships among ocean processes, environmental parameter distribution, biological responses and corresponding species distributions, scientific information and statistical analysis of habitat descriptors must accommodate the life cycle characteristics of the targeted species.

Satellite imagery provides an extensive (virtually worldwide) knowledge-base of sea-surface conditions, readily available in high or low resolution forms, allowing the mapping of important ocean processes that influence species distributions, albeit with the limitations that sub-surface phenomena cannot be described in this way and satellite data are available only since the early 1980s. In addition, extensive large-scale survey investigations often provide time-series of certain species distributions and sometimes also oceanographic data for the whole water column, allowing studies of relationships between environmental change and species environmental preferences. Finally, spatial statistical analysis and Geographic Information Systems (GIS) technology provide the tools to model species-habitat relationships and their variability and identify essential habitat areas (see, for example, Pierce et al., 2001, 2002; Valavanis et al., 2002, 2004).

Overviews of predictive species-habitat modelling approaches have been presented for various species groups in terrestrial (Guisan & Zimmermann, 2000; Elith & Burgman, 2002), freshwater (e.g. Olden & Jackson 2002; Behrouz et al., 2006) and marine ecosystems (e.g. Ferguson et al., 2006; Redfern et al., 2006). Related studies underline the fact that many marine species have wide distribution ranges and respond to environmental variation by changing their distribution patterns and habitat use (Perry et al., 2005; Laurel et al., 2007; Morrell & James, 2008). The marine environment is fundamentally dynamic: over a fixed background of bathymetry and seabed substrate, oceanographic conditions and prey availability vary in time (diurnally, seasonally, interannually) and space (vertically and horizontally) at various scales.

In the present article, we summarize a range of modelling approaches available to model species-habitat relations and map EFH for living marine resources, particularly fish, although some of these methods have been applied more often than others marine species datasets.

Objectives of essential fish habitat modelling

Fundamentally, EFH modelling is an applied science (in that it provides EFH maps based on analyzed scientific data), very often with the ultimate aim of providing tools to support the sustainable exploitation of living marine resources. Given the relatively low level of knowledge about external factors influencing the population dynamics of marine species, many published models are empirical, making few or no prior assumptions about underlying causal mechanisms, rather than mechanistic (process) or functional (e.g. optimization) models. Thus a relationship described by an empirical model may reflect a direct causal link, an indirect link or simply a coincidental (and most likely temporary) correlation. The whole process has been denigrated as "data mining" rather than hypothesis-driven science (Guisan et al., 2002). While various philosophers of science (e.g. Popper, 1963) (and indeed some national government funding bodies) have viewed falsification of hypotheses as the only legitimate form of scientific endeavor, in reality science is a much more complex process (e.g. Lakatos, 1970; Kuhn, 1996) and, indeed, European research funding under the Framework Programmes has mainly

targeted at science that offers economic and societal benefits rather science that aims solely to advance theory. We would argue that, in the context of applied sciences, such as fisheries science, data mining is a perfectly legitimate approach, which can lead to predictions and/or forecasts of fish distribution and abundance that are both testable and can be used to inform rational marine resource management. It is important to recognize that empirical models remain a form of hypothesis (regardless of whether an underlying causal mechanism can be identified) until, after an appropriate estimation of goodness-of-fit, they are tested on independent datasets. The majority of existing publications arguably fail in this respect, although most advance testable hypotheses.

Aside from the 'data mining' argument, there are at least two additional reasons for skepticism about the power and validity of empirical models based on environmental predictors. First, inclusion of a large number of putative explanatory variables in a model may led to overfitting, reducing predictive capacity and generality (Clark, 2005), assuming, that is, that a high ratio of explanatory variables to data points and/or collinearity between explanatory variables do not preclude fitting any model in the first place. A second issue is that the population dynamics of many (if not most) exploited species are driven by the amount, distribution and variability of fishing mortality, past and present. While this is true, the spatial distribution of EFH and abundance is still likely to be strongly dependent on the characteristics of the biotic and physical environment. As abundance increases, it may be expected that species expand from core 'preferred habitat' into increasingly marginal habitats. In this sense at least, knowledge of both the drivers of abundance and the habitat requirements (EFH) remains essential. In addition, environmental predictors may explain a high proportion of variance in recruitment strength and thus, in short-lived species (e.g. cephalopods, some small pelagic fish), a high proportion of variation in total abundance (Pierce et al. 2008, this volume).

The analysis of species distribution data has reached high statistical sophistication in recent years (Elith et al., 2006; Heikkinen et al., 2006). However, even the most complicated models cannot guarantee the improvement of our knowledge on the determinants of species distribution (Dormann et al., 2007).

Data acquisition and preparation

Species distribution modelling is only as good as the data used. The right sampling strategy can improve model results considerably and reduce the risk of making an inaccurate, biased or imprecise prediction. For that purpose, Hirzel & Guisan (2002) suggested some factors that could increase sampling efficiency. These are the increase of sample size, the use of regular sampling, and the use of environmental information to stratify sampling. Uncertainty, on which model inference and prediction depends, declines asymptotically with increasing sample size. The four strategies most frequently discussed are regular sampling (i.e. grid sampling), random sampling, equal random-stratified sampling, and proportional random-stratified sampling. A fifth approach, called gradsect (Austin & Heyligers, 1989, 1991), is close to a random-stratified sampling (either equal or proportional depending on its design) but sampling is concentrated within a few geographic transects designed across the main landscape gradients, mainly to reduce study costs (time- and cost-effective surveys) (Hirzel & Guisan, 2002; Hirzel & Arlettaz, 2003).

Examples of environmental and fish survey/fisheries datasets that may be used for EFH modelling are listed in Table 1. Fishery-independent survey datasets include a variety of surveyed parameters from fisheries acoustic data, experimental trawl data, ichthyoplankton and egg data. Usually at a coarser spatial resolution, commercial catch and fishing effort data can also provide distribution and abundance information for post-recruit fish. Environmental (ecogeographic) parameters that are likely to be relevant include interpreted satellite images for sea surface temperature (SST), chlorophyll-a (Chl-a), photosynthetically active radiation (PAR), euphotic depth (EUD), sea level anomaly (SLA), wind speed and direction, and modeled data for sea surface salinity (SAL) and surface currents (SSC). Hydrographic survey data can provide additional information on subsurface and sea bottom conditions. Spatial location variables and spatial patterns analyzed with statistical and geostatistical tools may add predictive power by acting as surrogates for one or more unknown environmental variables, or variables that cannot be measured easily. They can also potentially capture genuine geographic effects, such as proximity to favourable habitat features (e.g. spawning sites), or

Table 1 A list of datasets, their description and source that they were used in marine species habitat modelling

Parameter	Sensor/Model	Units	Resolutions	Source
Sea Surface Chlorophyll-a (CHLO)	SeaWiFS	mg/m^3	0.0833333°	http://oceancolor.gsfc.nasa.gov
Sea Surface Chlorophyll-a (CHLO)	MODISA	mg/m^3	0.0833333° and 0.0416667°	http://oceancolor.gsfc.nasa.gov
Sea Surface Temperature (SST)	AVHRR	°C	0.0128748°	http://eoweb.dlr.de:8080
Sea Surface Temperature (SST)	MODISA	°C	0.0833333° and 0.0416667°	http://oceancolor.gsfc.nasa.gov
Photosynthetically Active Radiation (PAR)	SeaWiFS	einstein/m^2/day	0.0833333°	http://oceancolor.gsfc.nasa.gov
Sea Surface Wind Speed and Direction (WIND)	QSCAT	m/sec and ° from N	0.25°	www.ssmi.com
Sea Surface Current Speed and Direction (SSC)	Merged T/P, Jason–1, ERS-2, Envisat	cm/sec and ° from N	0.125°	www.jason.oceanobs.com
Mean Sea Level Anomaly (MSLA)	Merged Jason-1, Envisat, ERS-2, GFO, T/P	cm	0.2942888°	www.jason.oceanobs.com
Sea Surface Salinity (SAL)	CARTON-GIESE SODA, CMA BCC GODAS, and NOAA NCEP EMC CMB GODAS models	psu	0.3333309°	http://iridl.ldeo.columbia.edu
Euphotic Depth (ZEU)	SeaWiFS (Lee and/or Morel)	m	0.0833333° and 0.0416667°	http://oceancolor.gsfc.nasa.gov
Bathymetry (BATH)	GEBCO	m	0.0166666°	www.ngdc.noaa.gov
Bathymetry (BATH)	Geosat and ERS-1	m	0.0280322°	http://ibis.grdl.noaa.gov/SAT/

The listed datasets are commonly georeferenced in a GIS database under the EnviEFH Project requirement (world-wide and/or Mediterranean coverage, weekly and/or monthly resolutions for the general period 1997-current, some earlier to 1997)

where juvenile dispersal is aided by particular habitat combinations (Francis et al., 2005). Similarly, inclusion of time of day, month/season and year in models may capture temporal patterns without explaining them.

For ocean processes, SST and Chl-a data can be used to locate thermal and productivity-enhancing fronts (Ullman & Cornillon, 2000; Valavanis et al., 2005), and marine productivity hotspots (Valavanis et al., 2004) and thus determine the distance of each sampling point from such features. Fixed physical features include bathymetry and derived variables, such as seabed slope, depth and slope variability, aspect, distances from coast and specific bathymetry zones, and sea bottom substrate types (where available). The final selection of candidate explanatory variables for EFH modelling is, as far as possible, based on knowledge of the biology and ecology of the species. Ideally, explanatory variables should describe characteristics of the ecology of the species and indicate the presence/strength of relevant ocean processes (e.g. upwelling or fronts) by using, for example, distances of surveyed points from such processes (Table 2). It may be important to include temporally and/or spatially displaced (e.g. time-lagged and teleconnected) environmental conditions (e.g. because the distribution of adults reflects processes affecting earlier life stages).

Table 2 Typical GIS output table including surveyed fisheries data (spatiotemporal and biological measurements) as well as derived habitat environmental and other descriptors*

SPATIAL					FISHERIES		TEMPORAL	HABITAT DESCRIPTORS										
No	LAT	LON	X	Y	ANCH	SARD	month	SSTm	CHLm	PARm	ALTm	SALm	DEP	Dcoast	Chl-AN	SST-AN	Dist-MPH	SED	Dfronts
1	39.67	25.07	849326.6	4399545.5	0	1398	may04	17.25	0.124	53.557	-0.354	38.388	-162	5677	-0.011	1.25	0	4	147902
2	39.69	25.07	849219.8	4401389.5	0	137	may04	17.38	0.124	53.665	-0.345	38.384	-179	4104	-0.011	0.25	0	4	149861
3	39.71	25.07	849098.8	4403231.5	0	0	may04	17.38	0.124	53.776	-0.339	38.382	-193	2153	NA	NA	NA	4	120048
4	39.72	25.07	848951.6	4405074.0	0	0	may04	17.5	0.124	54.005	-0.333	38.38	-192	3675	NA	-0.13	NA	3	115489
5	39.74	25.07	848771.1	4406910.0	2526	0	may04	17.5	0.124	54.128	-0.326	38.377	-177	1917	-0.013	0.88	0	3	100341
6	39.75	25.07	848659.3	4407951.0	0	0	may04	17.5	0.124	54.253	-0.32	38.375	-159	7370	-0.014	0.63	0	3	90599
7	39.75	25.06	847451.7	4408240.0	0	0	may04	17.5	0.136	54.256	-0.246	38.374	-167	6231	0.01	0.13	0	3	63389
8	39.77	25.06	847696.4	4410078.0	0	0	may04	17.38	0.143	54.393	-0.269	38.37	-148	1205	-0.013	0.13	0	3	82044
9	39.79	25.06	847959.0	4411899.0	0	0	may04	17.13	0.15	54.493	-0.262	38.367	-115	1283	-0.014	0	0	2	89459
10	39.80	25.07	848075.3	4413745.0	0	0	may04	17	0.157	54.721	-0.256	38.366	-95	3687	-0.01	0.13	0	4	73837
11	39.82	25.07	848005.6	4415589.5	2736	0	may04	16.75	0.173	54.821	-0.25	38.364	-64	5520	0.011	-1	1	4	64944
12	39.83	25.06	847642.7	4416778.5	0	0	may04	-9999	0.171	54.852	-0.244	38.361	-16	2399	-0.011	0.25	0	4	46951
13	39.47	25.28	868503.5	4379981.5	0	0	jun04	22.5	0.104	59.806	-0.847	38.425	-166	3986	0.01	-0.5	1	4	48395
14	39.49	25.28	868477.4	4381825.0	0	0	jun04	22.5	0.104	59.723	-0.888	38.422	-149	3417	NA	NA	NA	3	57233
15	39.51	25.29	868475.4	4383674.5	0	0	jun04	22.38	0.103	59.661	-0.985	38.416	-139	3012	0.01	0.13	0	3	42807
16	39.52	25.29	868476.3	4385493.5	0	0	jun04	22.38	0.102	59.591	-1.034	38.412	-130	2746	0.012	-0.13	1	3	51328
17	39.54	25.29	868096.5	4387366.5	0	0	jun04	22.38	0.102	59.481	-1.08	38.41	-123	7010	0.01	0.13	0	3	44246
18	39.56	25.28	868196.9	4389120.0	0	0	jun04	22.5	0.104	59.5	-1.126	38.407	-117	7736	NA	NA	NA	3	58220
19	39.57	25.29	868134.4	4390970.0	0	0	jun04	22.5	0.106	59.548	-1.22	38.402	-112	8215	0.011	-1	1	2	57385
20	39.59	25.29	868007.2	4392811.5	0	0	jun04	22.5	0.104	59.604	-1.269	38.4	-108	15861	0.011	-0.88	1	2	71647
21	39.61	25.29	867889.1	4394660.0	0	0	jun04	22.5	0.103	59.707	-1.32	38.397	-107	1722	0.012	-0.63	1	2	33055
22	39.62	25.29	867760.6	4396633.0	0	0	jun04	22.5	0.102	59.739	-1.427	38.392	-106	6688	0.011	-0.63	1	2	46440
23	39.64	25.29	867541.8	4398329.0	110	38	jun04	22.5	0.103	59.746	-1.483	38.39	-106	1937	0.01	-0.88	1	3	81875
24	39.66	25.28	867441.1	4400182.0	0	97	jun04	22.5	0.105	59.655	-1.542	38.387	-105	3920	NA	NA	NA	4	121545
25	39.67	25.28	867359.1	4402029.5	0	0	jun04	22.63	0.109	59.565	-1.602	38.385	-103	5307	0	0.38	0	4	123635
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

*LAT-LON (decimal degrees) and X-Y (meters): coordinates

ANCH, SARD, species (anchovy and sardine abundance acoustic index)

SST-DEP, monthly-averaged environmental variables and bathymetry; Dcoast, distance to coast (m); Chl-AN-SST-AN, environmental anomalies (indication of upwelling); MPH, presence/absence of marine productivity hotspots; SED, sediment types; Dfronts, distance to thermal fronts (m)

Data processing

Environmental and other habitat descriptor datasets can be inserted into GIS and projected under a common georeference system in order to extract a suite of environmental parameters for each sampling station. There are many tools available for this purpose (e.g. ArcGIS, MapInfo, etc) as well as a variety of data format conversion utilities for data communication among tools. For example, in the EnviEFH project, ESRI's ArcGIS software and Arc Macro Language (AML) programming language were used to create vector and raster layers of information (ESRI, 1994). Interpreted satellite images are processed as regular grids (ArcGIS GRID module) while fisheries data are placed in coverages of point topology (ArcGIS ARC module). Environmental data can be assigned to each sampling point of fisheries data by means of controlled cursors (pointers that move one-by-one through a selected set of geographic features) between vector and raster datasets (e.g. among selected sets of spatial point features and associated attribute tables and related remotely sensed parameter grids) using ArcGIS INFO/TABLES and ARCPLOT modules. An important consideration is the extent of the buffer zone (area around each sampling point), which is used to calculate an average for environmental parameters, e.g. the SST associated with a given sample could be a weekly or monthly average within a range of anything from 1 km to several kms of latitude and longitude. Derived descriptors, such as closest distances from thermal and productivity-enhancing fronts, marine productivity hotspots, temperature and chlorophyll-a anomalies, sediment types (when available) as well as distances from coast and bathymetry zones, are quantified using ArcGIS embedded distance functions under the ARC module.

Approaches to essential fish habitat modelling

Nature is too complex and heterogeneous to be predicted accurately in every aspect of time and space from a single, although complex, model (Guisan & Zimmermann, 2000). Levins (1966) formulated the principle that only any two out of three desirable model properties (generality, reality, precision) can be improved simultaneously while the third property has to be sacrificed. According to this principle, models can be classified to analytical (Pickett et al., 1994), mechanistic (e.g. Prentice, 1986) and empirical (Decoursey, 1992; Korzukhin et al., 1996). According to Lehmann et al. (2002), mechanistic models may provide more robust predictions than statistical models but the former are much more difficult to develop. Methods used to make spatial predictions should meet several criteria; they should be general enough to deal with the wide variety of attributes that need to be predicted; they should be rigorous and data-defined to make predictions in an objective and defensible manner; they should be standardized to produce uniform results and streamlined to facilitate the required analyses.

Modelling methods depend on data accuracy as well as the type of data. We can discriminate among various types of data, for example presence only data and presence/absence data (i.e. binary presence/absence 1:0). In the former case, we have available samples from locations that a species is found (1:presence) and in the latter we have samples from locations where a species is found as well as from those where it is not found (1:presence–0:absence). Abundance data (e.g. counts, catch rates) may sometimes be converted to presence only or presence/absence data depending on the biological and logistical constraints governing a conservation monitoring situation (Pollock, 2006).

With a short description provided later in the text, common methods for modelling presence data include ENFA, BIOCLIM, DOMAIN, GARP and MAXENT, while ANN, GLM, GAM and CART require accurate presence/absence data in order to generate statistical functions or discriminative rules that allow habitat suitability to be ranked according to distributions of presence and absence of species (Manel et al., 1999; Guisan & Zimmerman, 2000). Although a data-specific characteristic some of the latter methods could be used with presence only data (Brotons et al., 2004). Other categories of spatial modelling, also using binary presence/absence data are implemented via Kriging and Simulation techniques. All together could be integrated with a hierarchical Bayesian approach thinking about the scientific method as an iterative process: First, a hypothesis is formulated, rooted in current knowledge. Then, data are collected against which the hypothesis can be tested. Finally, current knowledge is updated in light of the data, repeating the process as appropriate. The Bayesian paradigm is similar, replacing current knowledge with prior probability distributions, data with a likelihood, and updated knowledge with posterior probability distributions, which may now serve as prior probability distributions for future studies. Zaniewski et al. (2002) argue that pure presence-only methods (such as ENFA) are more likely to predict potential distributions that more closely resemble the fundamental niche of the species, whereas presence/absence modelling is more likely to reflect the present natural distribution derived from realized niche. However, both methods aim at predicting distributions by sampling real distributions, and therefore, providing different estimations of the realized niche of the species (Loehle & LeBlanc, 1996). MacLeod et al. (2008, this volume) compared the performance of several presence-only models with that of GLM and showed that, although the latter had the highest predictive power, presence-only models could perform almost as well. Where survey effort is very uneven, both presence only and presence–absence models can give biased results, but only in the latter case is it possible to correct for the bias, e.g. by using effort as a weighting variable. In both cases, inadequate coverage of the full range of habitat types could lead to biased models.

Applying a statistical model consists of various main steps: parameter estimation, model selection, uncertainty estimation, model validation, and generating and testing predictions. Several techniques can be used in each step. For example, hypothesis testing or information criteria like the Akaike Information Criterion (Akaike, 1974) or Bayesian Information Criterion (Schwarz, 1978) can be used for model selection, uncertainty estimation can be done with classical methods based on the Fisher information matrix or using bootstrapping and jackknife techniques (Efron & Tibshirani, 1991). Also for model

validation, many methods are available to verify the underlying assumptions (e.g. auto-correlation and semi-variograms for testing the independence assumptions).

A common problem with ordinary least squares regression approaches for modelling species responses to environmental variables is a bias introduced due to unmeasured variables. Typically, since only some of the factors that affect a species distribution are measured and included in statistical models, the influence of an unmeasured factor could mask the predictive relationship between response and explanatory variables. Even by taking into account all factors, some systems include an unexplained stochastic component or show chaotic behaviour.

An alternative approach is to view explanatory variables as constraints rather than as correlates. This approach stems from a fundamental ecological principal, namely Liebig's law of the minimum. Conventional correlation and regression analyses are not based on the concept of limiting factors (Thompson et al., 1996; Cade et al., 1999). Quantile regression is based on this principle and quantifies the effects of limiting factors by fitting regression curves in quantiles near the maximum response (e.g. 0.90 regression quantile) (Eastwood et al., 2001; Cade & Noon, 2003; Hiddink, 2005). That way the effect of other measured or unmeasured factors is disregarded and only the cases when the tested factor has a limiting effect are taken into account. At the same time quantile regression reveals hidden bias and the existence of important processes that are not adequately represented by the measured variables (Cade et al., 2005). Another technique based on the fact that the upper boundary of the distribution of abundance reveals the limiting effect of a factor was developed by Blackburn et al. (1992). This method estimates the regression slope of the upper boundary by dividing the data into size classes and using the highest abundance for the calculations. Other techniques based on the law of the minimum have been proposed by Maller (1990), Kaiser et al. (1994) and Thompson et al. (1996). However, Liebig's law is open to criticism since resources might exhibit interactive effects, i.e. a factor can have a limiting effect only in the presence or absence of an other factor (Huisman & Weissing, 2002) or all the resources can be limiting simultaneously ('multiple limitation hypothesis', Rubio et al., 2003).

Austin (2007) in a critical review of current modelling approaches introduces structural equation models as an alternative offering the possibility to incorporate latent variables in the model. Structural equation models are a descendant of 'path analysis' developed by Wright (1921) to provide a mathematical description of a hypothetical causal scheme between traits of a species and abiotic/biotic environmental variables. Structural equation models are a contemporary fusion of factor analysis with path analysis, which is able of testing causal claims. The method is based on the fact that, although correlation does not imply causation, causation does necessarily imply particular types of statistical independencies and this constraint is what is tested. Shipley (1999) characterizes structural equation models as the most sophisticated method of performing statistical control on causal relationships. Its major disadvantages are the inevitable linearity of the relationships, the multivariate nature of the data, and the necessity for large sample sizes.

These methods have been developed in an attempt to solve technical problems associated with ordinary least squares regression, such as heteroscedasticity, extreme values, overdispersion and bias due to unmeasured explanatory variables, but most importantly to introduce biological thinking into statistical modelling. The lack of a biological basis for the most broadly used statistical methods leads to lack of interpretable results and, therefore, biological processes and causal relationships are being overlooked. Moreover results of null hypothesis testing on observational studies are arbitrary because of the lack of control over the response and predictor variables. The information-theoretic approach, namely the coupling of statistical tools with ecological theory to develop robust and interpretable models, has already started to dominate species distribution modelling (Rushton et al., 2004).

Presence-only models

ENFA (Ecological Niche Factor Analysis) compares the statistical distributions of the ecogeographical variables for a presence dataset consisting of locations where the species has been detected with the predictors' statistical distributions over a wider geographic area. Like principal component analysis

(PCA), ENFA summarizes all predictors into a few uncorrelated factors retaining most of the information. However, in this case, the factors have specific ecological meanings: the first factor is the 'marginality', reflecting the direction and distance in which the species niche differs most from the available conditions in the wider area. Subsequent factors represent the 'specialisation'; they are extracted successively by computing the direction that maximizes the ratio of the variance of the global distribution to that of the species distribution (Hirzel et al., 2001; Dettki et al., 2003; Brotons et al., 2004). Although developed for modelling habitats of terrestrial species, it has been recently applied to harbour porpoises on the west coast of Scotland (UK) (MacLeod et al., 2008, this volume).

BIOCLIM is an 'envelope' method that implements a bioclimatic envelope algorithm (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. For each environmental variable the algorithm finds the mean and standard deviation (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.

DOMAIN (Carpenter et al., 1993) 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. The Genetic Algorithm for Rule-Set Prediction (GARP) use 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). MAXENT estimates a target probability distribution by finding the probability distribution of maximum entropy (i.e. that is 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).

Presence/absence models

Classification and regression trees

Classification and regression trees (CART) function by way of recursive binary partitioning of data into increasingly homogenous groups with respect to the dependent variable. The two most homogenous groups of data with respect to the response variable are chosen (using the explanatory variables) and the resulting model is a tree-like structure consisting of a series of nodes (Lawler et al., 2004; Bourg et al., 2005).

In boosted regression trees (BRT), each of the individual models consists of a simple CART while the boosting algorithm uses an iterative method for developing a final model in a forward stage-wise fashion, progressively adding trees to the model by re-weighting the data to emphasize cases poorly predicted by the previous trees. Advantages offered by a BRT model include its ability to accommodate both 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). Leathwick et al. (2006a, b) used BRT to analyze fish species richness, environmental parameters and trawl characteristics.

Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) provide an alternative regression-based method for fitting non-linear responses, using piecewise linear fits rather than smooth functions (Friedman, 1991). MARS offer similar level of performance to other non-linear modelling techniques but may be extended by generalized boosted models (rarely used in ecological studies) where the estimation of the classifier's prediction is based on learning algorithms while systematically varying the training sample. For example, MARS were used to predict the distribution of freshwater diadromous freshwater fish in New Zealand (Leathwick et al., 2005).

Generalized linear, additive and mixed models

Generalized linear models (GLM) are extensions of linear regression in the sense that they use different

distributions (e.g. the Poisson distribution for count data, the binomial distribution for binary and proportional data, the negative binomial distribution for overdispersed count data). Furthermore, they use a link function between the expected values of the response variable and explanatory variables that ensures that the fitted values make sense (e.g. larger than 0 for count data, or between 0 and 1 for binary data) (McCullagh & Nelder, 1989). In the context of EnviEFH, GLM were used with a predictive rather than inductive goal. In such circumstances accuracy of model predictions is more important than significance of particular ecogeographic variables (Legendre & Lengendre, 1998). Nishida and Chen (2004) applied GLM on yellowfin tuna CPUE data of the Japanese longline fisheries in the Indian Ocean.

Generalized additive models (GAM, Hastie & Tibshirani, 1990; Wood, 2006) are straightforward extensions of GLM, which allow linear and other parametric terms to be replaced by smoothing functions; they are now widely applied in fisheries science (e.g. Zuur et al., 2007). In a comparison of modelling techniques, Moisen & Frescino (2002) found that GAM built on real (as opposed to simulated) data, performed marginally better than other techniques (CART, ANN and GLM). In addition, GAM are well-suited to model continuous relationships, provided the samples are spread out over the entire measured gradient (GAM gives wide confidence bands when most observations have the same values for the explanatory variables, or if the explanatory variables have extreme observations). GAM is perhaps the most common and well developed method for modelling fish habitats. For example, Maravelias et al. (2007) used GAMs to identify the distribution of Morocco dentex (*Dentex maroccanus*) in the NE Mediterranean and the environmental factors that are related with species distribution. Giannoulaki et al. (2006) used GAMs to identify the relationship between anchovy presence and environmental variables. Francis et al. (2005) predicted small fish presence and abundance in northern New Zealand harbours. Many authors suggested several approaches to improve model fitting and prediction capacity in GAM. Ridge regression and lasso (Tibshirani, 1996; Harrell, 2001; Hastie et al., 2001) and model averaging (Burnham & Anderson, 2004; Johnson & Omland, 2004) are promising alternatives in a model stepwise procedure. The use of CART techniques in a complementary way to GLM and GAM enables identification of ecologically meaningful interactions (Guisan et al., 2002).

Species distributional or trait data based on range map (extent-of-occurrence) or atlas survey data often display spatial autocorrelation, i.e. locations close to each other exhibit more similar values than those further apart. If this pattern remains present in the residuals of a statistical model based on such data, one of the key assumptions of standard statistical analyses, that residuals are independent and identically distributed, is violated. The violation of this assumption may bias parameter estimates and can increase type I error rates (falsely rejecting the null hypothesis of no effect) (Dorman et al., 2007). Often, spatial correlation is due to a missing covariate or interaction term in the model, the use of the wrong link function, or modelling a non-linear effect as linear. If refitting the model still gives residuals with spatial correlation, then a spatial correlation structure has to be incorporated in the model. It is also possible that due to the nature of the data, there is spatial correlation, in which case, a correlation structure has to be included anyway. Zuur et al. (2007) showed that falsely ignoring residual spatial autocorrelation, can give *P*-values that result in incorrect ecological interpretation. Including a spatial correlation structure provides more accurate predictions. In this case, modellers can quantify and integrate the spatial correlation structure (but should not ignore it).

Wagner & Fortin (2005) describe three different approaches to deal with spatial autocorrelation in models: (1) regression models incorporating a spatial term (autoregressive models: Keitt et al., 2002), (2) partialling-out of the spatial component in the species-environment relationship (variance partitioning: Legendre, 1993), and (3) residuals analysis (multiscale ordination, Maggini et al., 2006). Redfern et al. (2006) separated methods for addressing spatial autocorrelation into two general categories: (1) removing autocorrelation from the data and (2) explicitly accounting for autocorrelation in statistical tests and models. Fotheringham et al. (2002) discussed the approach of geographically weighted regression (similar to kernel regression) where spatial autocorrelation in parameter estimation is treated through the assignment of weights, such that those observations near the point in space where the

parameter estimates are desired have more influence on the result than observations further away.

In Generalized Regression Analysis and Spatial Prediction (GRASP), which uses GAM for spatial predictions, improvement is achieved by using either cross-validation as a model selection method, or weighted absences, or limited absences, or predictors accounting for spatial autocorrelation, or a factor variable accounting for interactions between all predictors (Maggini et al., 2006). With regard to spatial autocorrelation, model performance and stability can be improved by incorporate large spatial trends, although better models are obtained by accounting for local spatial autocorrelation. Interaction factors built from a regression tree on residuals of a first environmental model proved to be an efficient way to account for interactions between all predictors but this can lead to some overfitting (Maggini et al., 2006). BRUTO provides a rapid method to identify both variables to be included and the degree of smoothing to be applied in a GAM (Leathwick et al., 2006a, b). The final choice of model strategy should always depend on the nature of the available data and the specific study aims (Maggini et al., 2006).

Artificial neural networks

Artificial neural networks (ANN) are non-linear mapping structures based on hundreds or thousands of simulated neurons connected together in much the same way as the brain's neurons. ANN learn from experience (not from programming) and their behaviour is defined by the way its individual computing elements are connected and by the strength of those connections (weights). ANN can be trained to recognize patterns, classify data, and forecast future events (Kohonen, 1996; Ripley, 1996; Bishop, 1997). They have been shown to be universal and highly flexible function approximations for any data and any data dependencies. These make powerful tools for models, especially when the underlying data relationships are unknown (Lek & Guegan, 1999). However, ANN are not very common for fish habitat modelling. Brosse et al. (1999) used ANN to assess fish abundance and occupancy in the littoral zone of the lake Pareloup (SW France). Lek & Guegan (1999) presented an introduction of ANN as a tool in ecological modelling.

Model comparisons

The development of validation tools for prediction methods provides comparison methods. Although validation and comparison of models depends on specific datasets, predictions based on presence/absence data perform generally better than presence only data (Brotons et al., 2004) while presence/absence models perform generally better than abundance models (Francis et al., 2005). Presence only models can perform as well if survey coverage is evenly and widely distributed (see MacLeod et al., 2008, this volume) but they contain no mechanism to control for biased sampling. Given effort data, effort can be used as a weighting factor to compensate for unevenly distributed survey effort, although compensation for inadequate survey coverage of the full range of habitats is not possible. In principle, abundance models should be more informative, however, their poor performance in practice relates 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.

Brotons et al. (2004) suggested that GLM with presence/absence data predict more accurately than ENFA (presence data), although MacLeod et al. (2008, this volume) show that the performance of the two classes of model is similar given good survey coverage. Elith et al. (2006) evaluated the prediction of eleven distinct models and sixteen approaches that use presence-only data. They classified the models in 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 data only (BIOCLIM, DOMAIN and LIVES, Li & Hilbert, in press). 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 are often superior to methods such as traditional single decision trees (Ferrier & Watson, 1997; Elith & Burgman, 2002, Moisen & Frescino, 2002, Munoz & Felicisimo, 2004; Segurado & Araujo, 2004).

Among some available methods for predictions (e.g. GLM, GAM, ANN) and based on specific datasets, Lehmann et al. (2002) selected GAM because of the ecological interpretability of its non-parametric response curves and of the advantage of being statistically well defined, allowing good inference, but also flexible enough to fit the data closely. Leathwick et al. (2006a, b) fitted GAM and MARS models between the distributions of 15 freshwater fish species and their environment, and based on ROC values, results indicated little difference between the performances of both models. According to Olivier & Wotherspoon (2005), GLM classification accuracy on both test and training data was higher than that of CART and these authors finally suggested that the application of CART in a complementary way to GLM and GAM proved very useful in the model building phase as a guide to identify meaningful interactions using tree nodes.

Essential fish habitat mapping

Prediction versus range

There are at least two ways to produce an EFH map from a model. The first one is based on the model's graphical and numerical output. Estimated regression parameters, their signs, and significance levels indicate the strength and (partial) effect on the response variable. Some methods also provide graphical output for the explanatory variable (e.g. GAMs), which can be used for assessing its (partial) effect. Application of these ranges within GIS grids generates maps and indicates areas where variable ranges simultaneously meet, as potential EFH. Environmental ranges extracted from a specific surveyed area (e.g. North Aegean Sea in Eastern Mediterranean) can be applied to satellite data that cover the whole region (e.g. the whole Mediterranean basin), thus providing potential EFH maps for the region of interest.

The other approach includes the use of either a new set of values (or the original ones) for the explanatory variables of a fitted GAM model, in order to produce predictions. In this case, the predicted values can directly be mapped. The interpretation of prediction maps depends on the model's response variable (e.g. predicted values from a GAM with a response variable in presence/absence format gives probabilities of presence). The extent and the resolution of the predicted area depend on the set of the values used as the explanatory variables.

Use of satellite data

There are many advantages in using satellite datasets in prediction models. Due to the great spatial and temporal coverage of satellite data, it is easy to extract EFH maps that expand the sampling area. The outcome can sometimes lead to underestimated or biased predictions for the areas outside the sampling area. Since the model is based on environmental parameters within specific ranges at the sampling area, predictions out of those ranges might be unrealistic. Applying the models to a new area or different time period (prospective sampling) provides different results on habitat availability and this will usually result in a change in the model coefficients and apparent selection. However, one measure of the robustness of a habitat model is its capacity to be applied in other areas (Boyce et al., 2002; Olivier & Wotherspoon, 2005).

Seasonal or pooled

Given seasonal data, GAM can be applied in each season or in the whole dataset. In each case, there are advantages and disadvantages. The final choice depends on the study's objectives (prediction or description). If the whole dataset is used, the model will have a wider applicability because of the wider range of values of the explanatory variables. If only one season is used, the model will describe accurately the specific season (better than the previous model) but would be weak to predict other seasons. As a general rule, although always depending on the nature of specific datasets, the use of whole year datasets is recommended for predictive models whereas seasonal datasets might be used for descriptive models. The same type of argument can be applied to data collected across several years or several areas. Boyce et al. (2002) concluded that an overall model (all-year pooled data) is generally not a good indicator for individual-year models.

Model validation

There are several techniques to validate a model or to compare the accuracy of prediction among different models. Kappa statistics, Receiver Operating Characteristic (ROC), k-Fold cross validation, confusion matrices and classification tables are well described by Boyce et al. (2002) for presence/absence data. ROC statistics are preferable to Kappa, because, unlike Kappa, the ROC method avoids the problem of choosing a threshold value (Lehmann et al., 2002). For presence data, better model evaluation is achieved by withholding data (k-Fold partitioning) for testing model predictions or by comparing Resource Selection Function (RSF) predictions using models developed at other times and places (prospective sampling) (Boyce et al., 2002). One step towards improving evaluation of model performance in predicting distributions of species is to use independent, well structured presence–absence datasets for validation (Elith et al., 2006). On the other hand, Lehmann et al. (2002) indicated that cross-validation is generally more practical because it creates relatively independent random subsets and allows the use of all available data in the modelling process. By using entirely independent datasets, there is a risk of comparing different sampling strategies instead of evaluating a model (Lehmann et al., 2002). In addition, adequate data for independent validation may be difficult to collect and modelers usually prefer to use all available data to fit their model (Araujo et al., 2005; Maggini et al., 2006). The jackknife is also used (Jaberg & Guisan, 2001) for model validation.

False positives and false negatives are the types of prediction errors in modelling based on presence/absence data. Some data partitioning methods for the allocation of cases to training and testing datasets are resubstitution (Stockwell, 1992), which tends to provide optimistic measures of prediction success, bootstrapping (Buckland & Elston, 1993) in which accuracy is usually reported as a mean and confidence limits, randomization and prospective sampling (Capen et al., 1986) that could be from a different region or time, k-fold partitioning and jackknife sampling (Fielding & Bell, 1997).

Conclusions

We conclude this overview with the statement that the best model selection depends on the specific dataset and on the aims of the modelling process. In the EnviEFH project, the final objective was to produce essential fish habitat maps for small, large pelagic and demersal fish resources in the Mediterranean and adjacent seas, thus we focused more on the fact that different methods produce similar maps rather than examining in great detail the pros and cons of the statistics behind each model.

Published studies have moved from using simple correlation and regression (assuming normal distributions, homogeneity of variance and linear/non-linear relationships) to techniques that can account for the effects and interactions of multiple explanatory variables, response variables with a range of different distributions, non-linear relationships, heteroscedasticity, time-lagged effects and temporal autocorrelation and, one of the less tractable of these issues, spatial autocorrelation. Particular attention should be given to ensuring adequate and representative sampling and robust model validation methods.

Acknowledgement This review was supported by the EU-FP6-SSP Action EnviEFH 'Environmental Approach to Essential Fish Habitat Designation' (Project No: 022466).

References

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

Araujo, M. B., R. J. Whittaker, R. J. Ladle & M. Erhard, 2005. Reducing uncertainty in projections of extinction risk from climate change. Global Ecology and Biogeography 14: 529–528.

Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological Modelling 200: 1–19.

Austin, M. P. & P. C. Heyligers, 1989. Vegetation survey design for conservation: gradsect sampling of forests in north-east New South Wales. Biological Conservation 50: 13–32.

Austin, M. P. & P. C. Heyligers, 1991. New approach to vegetation survey design: gradsect sampling. In Margules, C. R. & M. P. Austin (eds), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Australia: 31–36.

Behrouz, A.-N., A. St-Hilaire, M. Berube, E. Robichaud, N. Thiemonge & B. Bobee, 2006. A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. River Research and Applications 22: 503–523.

Bishop, C. M., 1997. Neural Networks for Pattern Recognition. Oxford University Press, New York: 484.

Blackburn, T. M., J. H. Lawton & J. Perry, 1992. A method for estimating the slope of upper bounds in plots of body size and abundance in natural animal assemblages. Oikos 65: 107–112.

Bourg, N. A., W. J. McShea & D. E. Gill, 2005. Putting a CART before the search: successful habitat prediction for a rare forest herb. Ecology 86: 2793–2804.

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

Brosse, S., J. Guegan, J. Tourenq & S. Lek, 1999. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecological Modelling 120: 299–311.

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.

Buckland, S. T. & D. A. Elston, 1993. Empirical models for the spatial distribution of wildlife. Journal of Applied Ecology 30: 478–495.

Burnham, K. P. & D. R. Anderson, 2004. Multimodel inference – understanding AIC and BIC in model selection. Sociological Methods & Research 33: 261–304.

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.

Cade, B. S. & B. R. Noon, 2003. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment 1: 412–420.

Cade, B. S., B. R. Noon & C. H. Flather, 2005. Quantile regression reveals hidden bias and uncertaintly in habitat models. Ecology 86: 786–800.

Cade, B. S., J. W. Terrell & R. L. Schroeder, 1999. Estimating effects of limiting factors with regression quantiles. Ecology 80: 311–323.

Capen, D. E., J. W. Fenwick, D. B. Inkley & A. C. Boynton, 1986. On the measurement of error: multivariate models of songbird habitat in New England forests. In Verner, J. A., M. L. Morrison & C. J. Ralph (eds), Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates. University of Wisconsin Press, Madison: 171–175.

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

Clark, S. J., 2005. Why environmental scientists are becoming Bayesians. Ecology Letters 8: 2–14.

Decoursey, D. G., 1992. Developing models with more detail: do more algorithms give more truth? Weed Technology 6: 709–715.

Dettki, H., R. Lofstrand & L. Edenius, 2003. Modelling habitat suitability for moose in coastal Northern Sweden: empirical vs. process-oriented approaches. Ambio 32: 549–556.

Dormann, C. F., J. M. McPherson, M. B. Araujo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. D. Kissling, I. Kuhn, R. Ohlemuller, P. R. Peres-Neto, B. Reineking, B. Schroder, F. M. Schurr & R. Wilson, 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30: 609–628.

Eastwood, P. D., G. J. Meaden & A. Grioche, 2001. Modelling spatial variations in spawning habitat suitability for the sole *Solea solea* using regression quantiles and GIS procedures. Marine Ecology Progress Series 224: 251–266.

Efron, B. & R. Tibshirani, 1991. Statistical data analysis in the computer age. Science 253: 390–395.

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, Covelo CA: 303–314.

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.

ESRI, 1994. ARC Macro Language. Environmental Systems Research Institute Inc. Redlands, CA, USA: 3–37.

Ferguson, M. C., J. Barlowa, P. Fiedler, S. B. Reilly & T. Gerrodette, 2006. Spatial models of delphinid (family *Delphinidae*) encounter rate and group size in the eastern tropical Pacific Ocean. Ecological Modelling 193: 645–662.

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.

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.

Fotheringham, A. S., C. Brunsdon & M. E. Charlton, 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester, Wiley.

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.

Giannoulaki, M., A. Machias, V. D. Valavanis, S. Somarakis, A. Palialexis & C. Papaconstantinou, 2006. Spatial modelling of the European anchovy habitat in the Eastern Mediterranean basin using GAMs and GIS technology. General Fisheries Commission for the Mediterranean Scientific Advisory Committee, Sub-Committee for Stock

Assessment Working Group on Small Pelagic Species FAO, Rome, 11–14 September 2006.
Guisan, A., J. Edwards, C. Thomas & T. Hastie, 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157: 89–100.
Guisan, A. & N. E. Zimmermann, 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135: 147–186.
Harrell, F. E., 2001. Regression Modelling Strategies With Applications to Linear Models Logistic Regression and Survival Analysis. Springer-Verlag, New York.
Hastie, T. & R. Tibshirani, 1990. Generalized additive models. Chapman and Hall, London.
Hastie, T., R. Tibshirani & J. H. Friedman, 2001. The elements of statistical learning: data mining, inference and prediction. Springer-Verlag, New York.
Heikkinen, R. K., M. Luoto, M. B. Araujo, R. Virkkala, W. Thuiller & M. T. Sykes, 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography 30: 1–27.
Hiddink, J. G., 2005. Implications of Liebig's law of the minimum for the use of ecological indicators based on abundance. Ecography 28: 264–271.
Hirzel, A. H. & R. Arlettaz, 2003. Modelling habitat suitability for complex species distributions by the environmental-distance geometric mean. Environmental Management 32: 614–623.
Hirzel, A. H. & A. Guisan, 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling 157: 331–341.
Hirzel, A. H., V. Helfer & F. Metral, 2001. Assessing habitat suitability models with a virtual species. Ecological Modelling 145: 111–121.
Huisman, J. & F. J. Weissing, 2002. Oscillations and chaos generated by competition for interactively essential resources. Ecological Research 17: 175–181.
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.
Johnson, J. B. & K. S. Omland, 2004. Model selection in ecology and evolution. Trends in Ecology and Evolution 19: 101–108.
Kaiser, M. S., P. L. Speckman & J. R. Jones, 1994. Statistical models for limiting nutrient relations in inland waters. Journal of the American Statistical Association 89: 410–423.
Keitt, T. H., O. N. Bjornstad, P. M. Dixon & S. Citron-Pousty, 2002. Accounting for spatial pattern when modelling organism–environment interactions. Ecography 25: 616–625.
Kohonen, T., 1996. Self-organizing Maps. Springer-Verlag, New York: 428.
Korzukhin, M. D., M. T. Ter-Mikaelian & R. G. Wagner, 1996. Process versus empirical models: which approach for forest ecosystem management? Canadian Journal of Forest Research 26: 879–887.
Kuhn, T. S., 1996. The Structure of Scientific Revolutions. University of Chicago Press, USA.
Lakatos, I., 1970. Falsification and the methodology of scientific research programmes. In Lakatos, I. & A. Musgrave (eds), Criticism and the Growth of Knowledge. Cambridge University Press, London: 91–195.
Laurel, B. J., A. W. Stoner & T. P. Hurst, 2007. Density-dependent habitat selection in marine flatfish: the dynamic role of ontogeny and temperature. Marine Ecology Progress Series 338: 183–192.
Lawler, J. J., R. J. O'Connor, C. T. Hunsaker, K. B. Jones, T. R. Loveland & D. White, 2004. The effects of habitat resolution on models of avian diversity and distributions: a comparison of two land-cover classifications. Landscape Ecology 19: 515–530.
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.
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.
Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74: 1659–1673.
Legendre, P. & L. Legendre, 1998. Numerical Ecology. Elsevier.
Lehmann, A., J. Mc, C. Overton & J. R. Leathwick, 2002. GRASP: generalized regression analysis and spatial prediction. Ecological Modelling 157: 189–207.
Lek, S. & J. F. Guegan, 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling 120: 65–73.
Levins, R., 1966. The strategy of model building in population ecology. American Scientist 54: 421–431.
Li, J. & D. Hilbert, in press. A new predictive model, LIVES, for the potential distributions and habitats of species using presence-only data. Ecological Modelling (in press).
Loehle, C. & D. LeBlanc, 1996. Model-based assessments of climate change effects on forests: a critical review. Ecological Modelling 90: 1–31.
Lutchman, I., 2003. New technical approaches in fisheries management: the precautionary approach and the ecosystem approach. In De Fontaubert, C. & I. Lutchman (eds), Achieving Sustainable Fisheries: Implementing the New International Legal Regime. IUCN, The World Conservation Union: 31–45.
MacLeod, C. D., L. Mandleberg, C. Schweder, S. M. Bannon & G. J. Pierce, 2008. A comparison of approaches to modelling the occurrence of marine animals. Hydrobiologia. doi:10.1007/s10750-008-9491-0.
Maggini, R., A. Lehmann, N. E. Zimmermann & A. Guisan, 2006. Improving generalized regression analysis for the spatial prediction of forest communities. Journal of Biogeography 33: 1729–1749.
Maller, R. A., 1990. Some aspects of a mixture model for estimating the boundary of a set of data. Journal du Conseil International pour l'exploration de la Mer 46: 140–147.

Manel, S., J. M. Dias, S. T. Buckton & S. J. Ormerod, 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology 36: 734–747.

Maravelias, C. D., E. V. Tsitsika & C. Papaconstantinou, 2007. Evidence of Morocco dentex (*Dentex maroccanus*) distribution in the NE Mediterranean and relationships with environmental factors determined by Generalized Additive Modelling. Fisheries Oceanography 16: 294–302.

McCullagh, P. & J. A. Nelder, 1989. Generalized Linear Models. Chapman & Hall, London.

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

Morrell, L. J. & R. James, 2008. Mechanisms for aggregation in animals: rule success depends on ecological variables. Behavioral Ecology 19: 193–201.

Munoz, J. & A. M. Felicisimo, 2004. A comparison between some statistical methods commonly used in predictive modeling. Journal of Vegetation Science 15: 285–292.

Nishida, T. & D. G. Chen, 2004. Incorporating spatial autocorrelation into the general linear model with an application to the yellowfin tuna (*Thunnus albacares*) longline CPUE data. Fisheries Research 70: 265–274.

Olden, J. D. & D. A. Jackson, 2002. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biology 47: 1976–1995.

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.

Perry, A. L., P. J. Low, J. R. Ellis & J. D. Reynolds, 2005. Climate change and distribution shifts in marine fishes. Science 308: 1912–1915.

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

Pickett, S. T. A., G. Kolasa & C. G. Jones, 1994. Ecological Understanding: the Nature of Theory and the Theory of Nature. Academic Press, New York.

Pierce, G. J., V. D. Valavanis, A. Guerra, P. Jereb, L. Orsi-Relini, J. M. Bellido, I. Katara, U. Piatkowski, J. Pereira, E. Balguerias, I. Sobrino, E. Lefkaditou, J. Wang, M. Santurtun, P. R. Boyle, L. C. Hastie, C. D. MacLeod, J. M. Smith, M. Viana, A. F. González & A. F. Zuur, 2008. A review of cephalopod-environment interactions in European Seas and other world areas. Hydrobiologia.

Pierce, G. J., J. Wang & V. D. Valavanis, 2002. Application of GIS to cephalopod fisheries: workshop report. Bulletin of Marine Science 71: 35–46.

Pierce, G. J., J. Wang, X. Zheng, J. M. Bellido, P. R. Boyle, V. Denis & J.-P. Robin, 2001. A cephalopod fishery GIS for the Northeast Atlantic: development and application. International Journal of Geographical Information Science 15: 763–784.

Pollock, J. F., 2006. Detecting population declines over large areas with presence–absence, time-to-encounter, and count survey methods. Conservation Biology 20: 882–892.

Popper, K. R., 1963. The Growth of Scientific Knowledge. Routledge, London.

Prentice, I. C., 1986. Some concepts and objectives of forest dynamics research. In Fanta, J. (ed.), Forest Dynamics Research in Western and Central Europe. PUDOC, Wageningen: 32–41.

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.

Ripley, B. D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, London: 416.

Rubio, G., J. Zhu & J. P. Lynch, 2003. A critical test of the two prevailing theories of plant response to nutrient availability. American Journal of Botany 90: 143–152.

Rushton, S. P., S. J. Ormerod & G. Kerby, 2004. New paradigms for modelling species distributions? Journal of Applied Ecology 41: 193–200.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6: 461–464.

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

Shipley, B., 1999. Testing causal explanations in organismal biology: causation, correlation and structural equation modelling. Oikos 86: 374–382.

Stockwell, D. R. B., 1992. Machine learning and the problem of prediction and explanation in ecological modelling. Ph.D. Thesis, Australian National University.

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.

Thompson, J. D., G. Weiblen, B. A. Thompson, S. Alfaro & P. Legendre, 1996. Untangling multiple factors in spatial distributions: lilies, gophers and rocks. Ecology 77: 1698–1715.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society 58: 267–288.

Ullman, D. S. & P. C. Cornillon, 2000. Evaluation of front detection methods for satellite-derived SST data using in situ observations. Journal of Atmopheric and Oceanic Techology 17: 1667–1675.

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., S. Georgakarakos, D. Koutsoubas, C. Arvanitidis & J. Haralabous, 2002. Development of a marine information system for cephalopod fisheries in the Greek seas (eastern Mediterranean). Bulletin of Marine Science 71: 867–882.

Valavanis, V. D., I. Katara & A. Palialexis, 2005. Marine GIS: identification of mesoscale oceanic thermal fronts. International Journal of Geographical Information Science 19: 1131–1147.

Wagner, H. H. & M. J. Fortin, 2005. Spatial analysis of landscapes: concepts and statistics. Ecology 86: 1975–1987.

Wood, S. N., 2006. Generalized Additive Models: An Introduction with R. CRC Press, London.

Wright, S., 1921. Correlation and causation. Journal of Agricultural Research 20: 557–585.

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.

Zuur, A. F., E. N. Ieno & G. M. Smith, 2007. Analysing Ecological Data. Springer Series: Statistics for Biology and Health.