Modelling the habitat suitability of cetaceans: Example of the sperm whale in the northwestern Mediterranean Sea

Emilie Praca [a,b,*], Alexandre Gannier [c], Krishna Das [b], Sophie Laran [a]

[a] Centre de Recherche sur les Cétacés—Marineland, 306 avenue Mozart, 06600 Antibes, France
[b] MARE Center—Laboratory for Oceanology, University of Liège, Sart Tilman, Bâtiment B6c, 4000 Liège, Belgium
[c] Groupe de Recherche sur les Cétacés, BP715, 06633 Antibes cedex, France

ARTICLE INFO

ABSTRACT

Cetaceans are mobile and spend long periods underwater. Because of this, modelling their habitat could be subject to a serious problem of false absence. Furthermore, extensive surveys at sea are time and money consuming, and presence–absence data are difficult to apply. This study compares the ability of two presence–absence and two presence-only habitat modelling methods and uses the example of the sperm whale (*Physeter macrocephalus*) in the northwestern Mediterranean Sea. The data consist of summer visual and acoustical detections of sperm whales, compiled between 1998 and 2005. Habitat maps were computed using topographical and hydrological eco-geographical variables. Four methods were compared: principal component analysis (PCA), ecological niche factor analysis (ENFA), generalized linear model (GLM) and multivariate adaptive regression splines (MARS). The evaluation of the models was achieved by calculating the receiver operating characteristic (ROC) of the models and their respective area under the curve (AUC). Presence–absence methods (GLM, AUC = 0.70, and MARS, AUC = 0.79) presented better AUC than presence-only methods (PCA, AUC = 0.58, and ENFA, AUC = 0.66), but this difference was not statistically significant, except between the MARS and the PCA models. The four models showed an influence of both topographical and hydrological factors, but the resulting habitat suitability maps differed. The core habitat on the continental slope was well highlighted by the four models, while GLM and MARS maps also showed a suitable habitat in the offshore waters. Presence–absence methods are therefore recommended for modelling the habitat suitability of cetaceans, as they seem more accurate to highlight complex habitat. However, the use of presence-only techniques, in particular ENFA, could be very useful for a first model of the habitat range or when important surveys at sea are not possible.

1. Introduction

Habitat modelling increases the knowledge about the spatial distribution of a species and its relationship with environmental variables. Such information is of great interest for theoretical studies on ecological niches or for practical purposes such as defining and managing protected areas. Habitat modelling can, moreover, be used to predict the impact of climate changes on species spatial distribution. In recent years, such conservation and management considerations have gained in ecological importance. At the same time, computational capabilities have considerably improved, leading to an increase in the number of habitat modelling techniques, using various

* Corresponding author at: Centre de Recherche sur les Cétacés—Marineland, 306 avenue Mozart, 06600 Antibes, France.
Tel.: +33 4 93 33 55 77; fax: +33 4 93 33 76 91.
E-mail address: emilie.praca@gmail.com (E. Praca).

statistical methods such as multiple regression or multi-factorial analyses.

The most used habitat modelling techniques (such as generalized linear models, GLMs) are based on presence–absence data (Guisan and Zimmermann, 2001; Redfern et al., 2006). 'True' absence data (when animals are actually absent) are not easy to collect for mobile or inconspicuous species. For example, Kelly (2000) cited in Hirzel et al. (2002) estimated that 34 visits to a site are needed to confirm the absence of a snake (*Coronella austriaca*). 'False' absence data, when animals are present but not detected, can significantly bias the analysis. As several cetacean species are able to spend long periods underwater and are very discreet at the surface, modelling their habitat with presence–absence methods may be subject to such biases, if absence data are not carefully considered. Moreover, collection of cetacean distribution data requires long and expensive surveys at sea.

This shortcoming can be avoided using presence-only methods such as principal component analysis (PCA) or ecological niche factor analysis (ENFA) (Hirzel et al., 2002). Because of the use of presence-only data, such methods tend to overestimate the area of suitable habitat. Indeed, presence-only methods seem to predict the potential distribution (fundamental niche), whereas presence–absence methods could reflect the present distribution (realized niche) of the species (Brotons et al., 2004; Zaniewski et al., 2002). Even though presence-only methods have limitations, they could be very useful for a first approach of habitat modelling for cetaceans.

The sperm whale (*Physeter macrocephalus*) is one of the eight common cetacean species inhabiting the northwestern Mediterranean Sea (NWMS, Fig. 1) (Duguy, 1991). In this area, sperm whales are exposed to anthropogenic disturbances such as noise and ship collisions (with ferries or high-speed boats), net entanglement and pollution (Aguilar et al., 2002; Di Natale and Notarbartolo di Sciara, 1994; Notarbartolo di Sciara and Gordon, 1997). With the creation of a marine protected area, the International Sanctuary for Marine Mammals, it was interesting to model the critical habitat of the sperm whale within the framework of management and conservation.

In the NWMS, the average patterns of the sperm whale typical deep dive are 45 min for the underwater feeding period and 9 min for the surface resting period (Aguilar et al., 2002; Drouot et al., 2004). Because the whales feed throughout the day/night cycle (Drouot et al., 2004; Watwood et al., 2006) and spend around 15% of the time at the surface, the use of only visual detection does not well represent their spatial distribution. We compensate for this lack by the use of passive acoustic detection along the survey track. Indeed, the sperm whale emits regular clicks during its feeding dives. These clicks are produced 80% of the time of the dive, allowing detection of individuals from several kilometres away when they are underwater (Watwood et al., 2006) and the use of both presence-only and presence–absence modelling methods. We therefore compared four methods to model the habitat suitability (HS) of the sperm whale: two well-established methods, PCA and GLM, and two more recent methods, ENFA and multivariate adaptive regression splines (MARS). We will discuss the statistical accuracy and the ecological meaning of the resulting models, in order to show the advantages and disadvantages of each technique for the habitat modelling of cetaceans.

2. Material and methods

2.1. Sampling surveys

From 1998 to 2005, summer surveys were conducted on a motor-sailing boat at a speed of 6 knots. In addition, during summer 2001, a motor boat was used for surveys at a speed of 11–12 knots (Fig. 1a) (Gannier, 2006). The survey track was designed as random zigzags from the upper slope to nearby pelagic waters, and crossings from France mainland to Corsica or to Balearic Islands
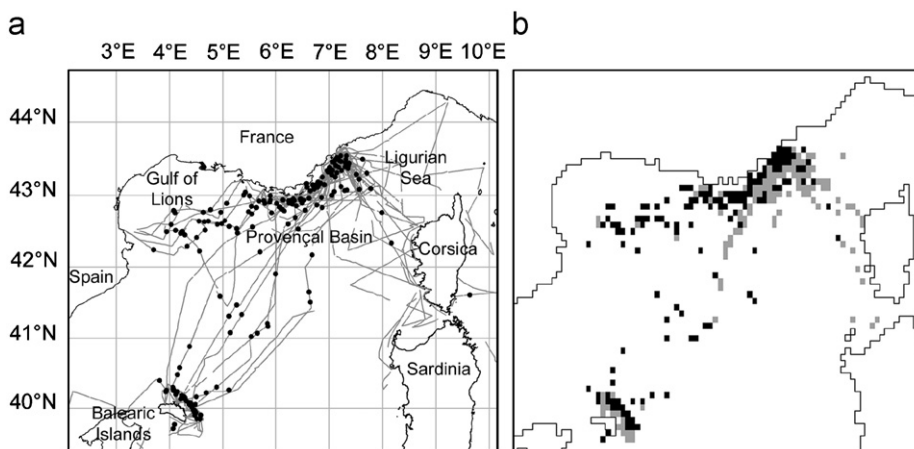

Fig. 1. Survey tracks (grey line) realized from 1998 to 2005 in the northwestern Mediterranean Sea and location of the observation sequences (black dots) of sperm whales (*Physeter macrocephalus*) (a); presence cells (black) and absence cells with a minimum of 5 km of survey effort (grey) of the sperm whale (b).

were performed when good meteorological conditions occurred during several days.

The protocol combined visual searching and systematic passive acoustic listening station (see Gannier et al., 2002). In brief, the visual survey was conducted by three experienced observers, scanning continuously with the naked eye the frontal sector ($-90°$ to $+90°$). The passive acoustic survey along the cruise track consisted of 1 min of listening station every 2 nm (3.7 km) and used a dual-channel towed hydrophone (Magrec Ltd., Lifton, UK). Sperm whales were recognized by their typical signal composed of regular clicks (Teloni, 2005). The following parameters were recorded at each listening station or sighting location: sea state, position of the boat and of the animals (if a sighting occurred), visual conditions (index V, varying between 0 and 6 and depending on the wind speed in Beaufort, the sky cover and the sea state) (Gannier, 1997; Gannier et al., 2002), background acoustic noise (index U, varying from 1 to 5) and the bio-acoustic signal level (index SL, varying between 0 and 5) (Gordon et al., 1998, 2000). V, U and SL were estimated by experienced observers. The significance of the data was improved by removing observations with bad visual or acoustical conditions, i.e. when $V < 4$, $U > 3$ or $SL < 2$.

The data were merged into observation sequences, in ArcGIS 8.3 (ESRI Inc., Redlands, USA), in order to minimize autocorrelation in the analysis. All successive acoustical or visual observations obtained with less than a 1-h time-lag (approximately 6 nm) were considered to be part of the same group (Gordon et al., 2000). One geographic position for each observation sequence was chosen either as the location of one central visual sighting or of the acoustic detection with the best SL.

2.2. Data treatment

A 9×9 km grid cell of the study area was created, in which both observation sequences and eco-geographical variables (EGVs) were implemented to construct presence–absence and EGVs grid cells. This cell size was chosen in order to use chlorophyll concentrations that were not available at higher resolution.

Presence cells were defined as cells where one or several observation sequences were located and had the value of 1. A homogenous searching effort was not feasible in our large study area, because the survey of offshore areas requires extended periods of good weather. A weight for presence cells was then used to balance this discrepancy. It was computed for each presence cell as the corresponding total number of observation sequences obtained divided by the total number of kilometres of searching effort obtained in this cell. This weight was used in the statistical softwares (see below) as an observation multiplier, for example, a cell with a weight of 3 will be considered three times. Absence cells were defined as cells on the survey track where no detections were obtained and had the value of 0. Absence data were maximized by selecting only absence cells with a minimum of 5 km of searching effort. The presence–absence data set was randomly split into calibration and validation data sets, representing 70% and 30% of the data set, respectively. GLM and MARS were performed with the total calibration data set, while only the presence cells of this data set were used for PCA and ENFA.

As information on the spatial distribution of sperm whale preys in the Mediterranean are scarce, hydrological and chlorophyll concentration data were used as proxies. However, sperm whales are not directly influenced by chlorophyll concentrations, as a gap occurs in trophic webs between primary production and cephalopods (Jaquet, 1996). The sperm whale summer distribution may then be influenced by the primary production situation during the phytoplankton bloom. Therefore, we modelled the summer distribution of sperm whale using data from both summer and phytoplankton bloom periods.

EGVs were variables used in previous cetacean habitat modelling studies (e.g. Gregr and Trites, 2001; Hamazaki, 2002), related to topography, temperature, salinity and primary production. Monthly resolution was used for the hydrological and biological EGVs, in order to compute seasonal situations for the two following periods: the summer (June–August) and the phytoplankton bloom period (February–April). These seasonal maps were then averaged over all survey years, resulting in two seasonal maps for each EGV.

This use of multi-year average situations, instead of daily or weekly data, was needed because we compiled sperm whale presence data interannually in order to have sufficient data. This was also required by the ENFA, as this method compares the mean available habitat in the study area and the species habitat (see below), preventing the use of close to real-time data. The same multi-year averages were then used in all models in order to perform the comparison of methods.

Depth, slope and the distance to the 200-m contour, which has been shown to be more relevant than the coastline for teuthophageous species (Mangion and Gannier, 2002), were obtained from the GEBCO Digital Atlas (IOC-IHO-BODC, 2003). Depth and slope were log-transformed in order to reduce their high variation range.

Sea surface temperature (SST) data were downloaded, depending on their availability, from the Pathfinder sensor (PO.DAAC) for 1998–2002 and from the Modis sensor (OceanColor) for 2002–2005. The front detection maps were computed in Idrisi Andes (Clark Labs, Clark University, Worcester, USA) applying a Sobel filter on the SST. This filter highlights horizontal and vertical gradients and replaces the value of a central cell, in a matrix of 3×3 cells, by the magnitude of the gradient (here in °C), using the following coefficients:

$$x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} c_1 & c_4 & c_7 \\ c_2 & c_5 & c_8 \\ c_3 & c_6 & c_9 \end{bmatrix} \tag{1}$$

$$y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \times \begin{bmatrix} c_1 & c_4 & c_7 \\ c_2 & c_5 & c_8 \\ c_3 & c_6 & c_9 \end{bmatrix} \tag{2}$$

The new value of the central cell c_5 is computed as

$$c_5 = \sqrt{(x^2 + y^2)} \tag{3}$$

Chlorophyll concentrations were obtained from the SeaWifs sensor website (OceanColor) for 1998–2005. Salinity data were obtained from the MEDAR/MEDATLAS II database (MODB). They were only available from 1998 until 2002, but were considered to be representative of average conditions of the study period.

In the Gulf of Lions, the Rhône river exports high quantities of nutrients and particles (Conan et al., 1998), which increase the turbidity. This phenomenon leads to an overestimation of chlorophyll concentrations in satellite data ($> 0.8 \, \mathrm{mg\,m^{-3}}$ even in summer) and the Rhône panache can be classified as turbid case 2 water (Antoine et al., 1996). Consequently, the area influenced by the panache of the Rhône was removed from our analysis.

2.3. Modelling methods

PCA and ENFA are both presence-only multifactorial analyses, transforming the set of EGVs in the same number of non-correlated factorial axes (Hirzel et al., 2002; Legendre and Legendre, 1998). In PCA, the eigenvalues of the factorial axes are computed with the variance–covariance matrix of the EGV matrix (Hirzel et al., 2002; Legendre and Legendre, 1998), while ENFA introduces ecological significance in the computation of factorial axes (Hirzel et al., 2002). For this method, marginality (how much a species' habitat differs from the mean available conditions) is represented in the first factorial axis, and specialization (breadth of the habitat) is maximized in the subsequent axes. For both methods, the number of relevant axes was chosen using Mac Arthur's broken-stick method (Hirzel et al., 2006). Finally, HS maps were built with the median algorithm, which compares the position of each cell of the study area to the distribution of presence cells on the different factorial axes. A cell adjacent to the median of an axis would score 1, and a cell outside of the species distribution would score 0. All ENFA and PCA analyses were conducted using Biomapper 3.2 (Hirzel et al., 2006).

GLM and MARS are presence–absence methods (Friedman, 1991; McCullagh and Nelder, 1989), for which a logistic regression was used to relate the binary response variable (presence or absence of sperm whales) with the continuous EGVs:

$$\log it[P(Y=1)] = \frac{P(Y=1)}{1 - P(Y=1)} = b_0 + \sum b_i x_i \tag{4}$$

$$P(Y=1) = \frac{e^{b_0 + \sum b_i x_i}}{1 + e^{b_0 + \sum b_i x_i}} \tag{5}$$

where $P(Y=1)$ is the probability of presence varying between 0 and 1, x_i is an EGV and b_0 is the intercept. In GLMs, b_i is a scalar coefficient, leading to a linear relationship between the response variable and the EGVs (McCullagh and Nelder, 1989). An exhaustive estimation of the GLM models was computed, and the model with the lowest Akaike information criterion (AIC) was chosen as the more parsimonious (Tabachnick, 2000). In MARS, $b_i x_i$ is replaced by a piecewise basis function, composed of several linear segments with different slopes and breaking knots (Friedman, 1991). Basic functions are defined in pairs:

$$bf_i = \alpha_i \, max(0; t_i - x_i) \tag{6}$$

$$bf_{i+1} = \alpha_i \max(0; x_i - t_i) \tag{7}$$

where bf_i and bf_{i+1} are basic functions, α_i is the slope of the linear segment, t_i is the breaking knot and x_i is an EGV. The values of bf_i (Eq. (6)) will be $\alpha_i t_i$ when x_i is 0, declining to 0 as x_i approaches t_i and remaining at 0 when x_i is superior to t_i. In contrast, bf_{i+1} (Eq. (7)) takes the values of $\alpha_i(x_i - t_i)$ when x_i is greater than t_i and takes the value of 0 otherwise. More than 1 knot (i.e. bf pair) can be defined for each EGV, allowing the development of complex non-linear relationships. The whole of the basic functions initially over-fits the data. The model is then simplified using a backward/forward stepwise cross-validation in order to identify the significant functions. The probability of presence equations of GLM and MARS models were computed in Statistica 8.0 (Statsoft Inc., Tulsa, USA) and imported in Idrisi Andes to compute HS maps of GLM and MARS models with the relevant EGVs.

2.4. Model validation and comparison

The statistical accuracies of the model predictions were evaluated by comparing the probabilities of presence highlighted by HS maps and the validation data set. A method of validation commonly used is the confusion matrix, which cross-tabulates the observed and predicted presence and absence patterns (Fielding and Bell, 1997). It computes sensitivity as the fraction of presence cells well predicted as presence and specificity as the fraction of absence cells well predicted as absence. However, this method depends on a threshold between presence and absence, generally fixed to 0.5, which could introduce bias if this threshold is not optimal (Boyce et al., 2002). An alternative is the receiver operating characteristic (ROC) curves and their corresponding area under the curve (AUC) (Beck and Schultz, 1986). This method evaluates the proportion of correctly and incorrectly classified predictions over a continuous range of thresholds (Beck and Schultz, 1986; Boyce et al., 2002). ROC curves were obtained with Analyse-it (Analyse-it Software Ltd., Leeds, UK) for Microsoft Excel (Microsoft Corporation, Redmond, USA), plotting sensitivity vs. $1-$specificity pairs for each presence–absence threshold. A perfect model has an AUC of 1 and a random model an AUC of 0.5. The closer the AUC is to 1, the better is the fit of the model. AUCs of the different models were compared to the AUC of a random model with a Z-test (Boyce et al., 2002; DeLong et al., 1988). Finally, the point minimizing sensitivity–specificity, i.e. where the number of wrongly predicted cells as absence and presence is minimal, was chosen as the threshold between absence and presence.

3. Results

The data set was composed of 14,259 km of effort covered during the survey period (1998–2005), with 187 observation sequences, transformed in a grid cell with 135 cells of presence and 1025 cells of absence. The selection of the absence cells with a minimum of 5 km of effort resulted in 180 absence cells (Fig. 1b). Furthermore, considering the high correlation between salinity and SST of both seasons ($r^2 > 0.75$), models were tested only with the salinity or the SST of the summer period and the model with the best statistical validation was kept.

3.1. Principal component analysis

The Mac Arthur's broken-stick method retained the first three factorial axes as relevant to the PCA model. It had an AUC of 0.58, not significantly different from the random model (Z-test, $Z = 1.23$, $p = 0.11$, Fig. 2a). The threshold between predicted presence and absence was 0.47, with a sensitivity of 59% and a specificity of 54.5%.

The first factorial axis seemed to highlight the influence of topography on the habitat of the sperm whale (Table 1). It indicated the importance of depth and distance to the 200-m contour (coefficients of 0.82 and 0.91, respectively) and low slope (−0.73). In contrast, the second axis highlighted the influence of biological EGVs with important chlorophyll concentrations in summer (0.91) and during the phytoplankton bloom period (0.79). The third axis highlighted the influence of thermal fronts in both periods (0.80 for the summer period and 0.52 for the phytoplankton bloom period) (Table 1). The HS map of the PCA model (Fig. 4a) showed a core habitat on the continental slope in the Ligurian Sea and near the Balearic Islands, but also on the continental shelf close to Sardinia, and offshore between Corsica and the Spanish coast.

3.2. Ecological niche factor analysis

For the ENFA, Mac Arthur's broken-stick method retained the first five factorial axes. This model had an AUC of 0.66, significantly different from the random model (Z-test, $Z = 2.33$, $p = 0.01$, Fig. 2a). The presence–absence threshold was of 0.45, with a sensitivity of 64.1% and a specificity of 63.6%.

The marginality factorial axis indicated a strong relationship for cells with steep slope (coefficient of 0.61) (Table 2). The first specialization axis highlighted the restriction of the species to the lower SST, lower distance to the 200-m contour (0.54) and lower chlorophyll concentrations for the phytoplankton bloom period (0.57). The last three axes showed the restriction of the species to waters with the lower frequencies of SST fronts (0.57), the higher chlorophyll concentrations in summer (0.56) and the steeper slopes (0.56). This method did not highlight the water depth as a significant variable (Table 2). The HS map of this method revealed a core habitat on the continental slope in almost the whole study area: near the Sardinian coast, near the French coast in the

Table 1
Relevant axes (with their eigenvalues) and the EGV coefficients of the principal component analysis model.

EGVs	Axis 1 (0.31)	Axis 2 (0.23)	Axis 3 (0.15)
Depth (log)	0.82	−0.16	0.08
Chlorophyll concentrations$_{summer}$	0.08	0.91	0.14
Chlorophyll concentration$_{bloom}$	0.50	0.79	0.09
Distance to the 200-m contour	0.91	−0.23	0.12
Salinity$_{summer}$	0.37	0.29	−0.51
Slope (log)	−0.73	0.42	0.02
Thermal front detection$_{bloom}$	−0.15	−0.17	0.80
Thermal front detection$_{summer}$	0.18	0.24	0.52

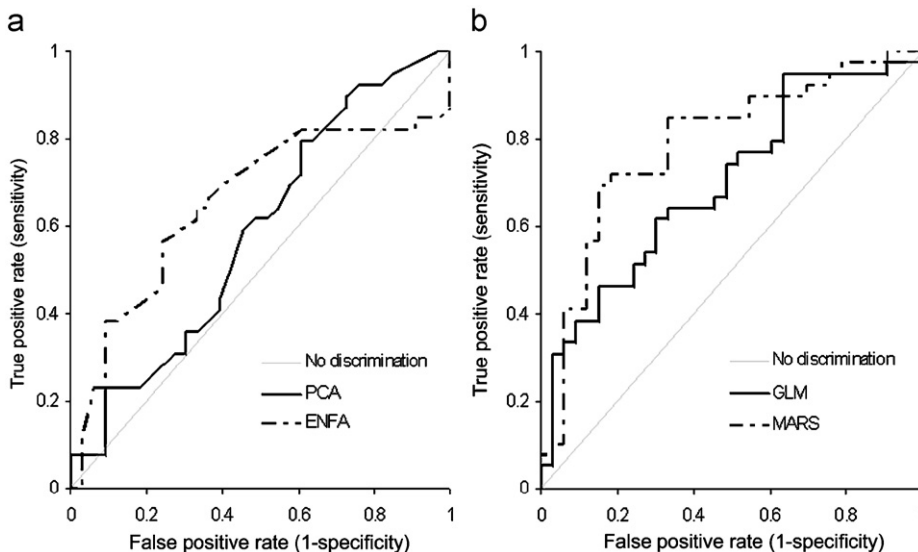Summer: summer period, bloom: phytoplankton bloom period.

Fig. 2. Receiver operating characteristic curves of the four modelling methods used: PCA and ENFA (a) GLM and MARS (b).

Table 2
Relevant axes (with their eigenvalues) and the EGV coefficients of the ecological niche factor analysis model.

EGVs	Axis 1 (0.16)	Axis 2 (0.23)	Axis 3 (0.18)	Axis 4 (0.16)	Axis 5 (0.10)
Depth (log)	−0.14	0.40	0.29	0.16	0.28
Chlorophyll concentrations$_{summer}$	0.40	0.11	0.03	0.56	0.73
Chlorophyll concentration$_{bloom}$	−0.20	0.57	0.18	0.56	0.03
Distance to the 200-m contour	−0.48	0.54	0.46	0.01	0.22
Salinity$_{summer}$	−0.40	0.38	0.06	0.09	0.13
Slope (log)	0.61	0.06	0.59	0.41	0.56
Thermal front detection$_{bloom}$	−0.18	0.25	0.57	0.40	0.12
Thermal front detection$_{summer}$	−0.03	0.09	0.08	0.13	0.04

The positive or negative sign is relevant for the first axis coefficients (marginality), but in the following axis (specialization) only the absolute value of coefficients is considered (summer: summer period, bloom: phytoplankton bloom period).

Table 3
Eigenvalues, intercept and *p*-values of the relevant EGVs retained for the generalized linear model.

EGVs	Eigenvalues	*p*-Values
Intercept	−2.55	0.06
Depth (log)	0.43	0.02
Chlorophyll concentrations$_{summer}$	17.89	0.0001
Chlorophyll concentrations$_{bloom}$	−7.59	>0.0001
Distance to the 200-m contour	0.04	>0.0001
Slope (log)	0.58	0.0005

Summer: summer period, bloom: phytoplankton bloom period.

Table 4
Slopes and breaking knots of the relevant basic functions (bf) of the multivariate adaptive regression splines model.

	EGVs	Slopes	Breaking knots
bf 1	Distance to the 200-m contour	0.003	28.93
bf 2	Distance to the 200-m contour	0.02	28.93
bf 3	Chlorophyll concentrations$_{bloom}$	−1.44	0.59
bf 4	Depth (log)	0.26	5.30
bf 5	Salinity$_{summer}$	−2.23	37.80

Summer: summer period, bloom: phytoplankton bloom period.

Ligurian Sea, in the Gulf of Lions and near the Balearic Islands (Fig. 4b).

3.3. Generalized linear model

The best GLM model (AIC of 254.28) had an AUC of 0.70 significantly different from the random model (Z-test, $Z = 3.20$, $p = 0.0007$, Fig. 2b). It presented a threshold between presence and absence of 0.65, with a sensitivity of 64.1% and a specificity of 63.6%.

The eigenvalues of the relevant EGVs showed a strong relationship with chlorophyll concentration in summer (17.89) and medium influence of the bathymetry (0.43), the slope (0.58) and the distance to the 200-m contour (0.04). A relatively strong negative relationship with chlorophyll concentrations for the phytoplankton bloom period (7.59) was also shown (Table 3). The HS map of this method showed a core habitat on the whole continental slope and the western offshore waters of the study area (Fig. 4c).

3.4. Multivariate adaptive regression splines

The breaking knots and slopes of the significant basic functions are given in Table 4. The resulting model had an AUC of 0.79, significantly different from the random model (Z-test, $Z = 5.12$, $p < 0.0001$, Fig. 2b). The threshold between predicted presence and absence of this model was of 0.79, with a sensitivity of 71.8% and a specificity of 72.7%.

The basic functions highlighted a decreasing affinity of the sperm whales for areas with distance to the 200-m contour inferior to 29 km and an increasing affinity after this breaking knot (Fig. 3a). They also showed an influence of waters with low chlorophyll concentrations for the phytoplankton bloom period (Fig. 3b), important depth (Fig. 3c) and low salinity in summer (Fig. 3d). The HS map of this model revealed a principal habitat on the continental slope near the Balearic and Sardinia Islands, the Spanish coast, the Gulf of Lions and the Provençal coast. Furthermore, the offshore waters in the western part of the study area were also highlighted as important for the species (Fig. 4d).

3.5. Statistical comparison between the modelling methods

The different AUC and *p*-values of the Z-tests are summarized in Table 5 to compare the different modelling methods. The only significant difference between the models was found between the MARS and the PCA models (difference of 0.20, $p = 0.007$). The other differences of AUC varied between 0.04 and 0.13 and were not significant ($0.070 < p < 0.545$).

4. Discussion

4.1. Data used for the habitat modelling

The interannual compilation of presence data and the ENFA required the use of average data. In order to perform the method comparison, we preferred to use the same data for all the four methods tested. This use of average data could confound the effect of interannual variation in sperm whale occurrence and environmental conditions.
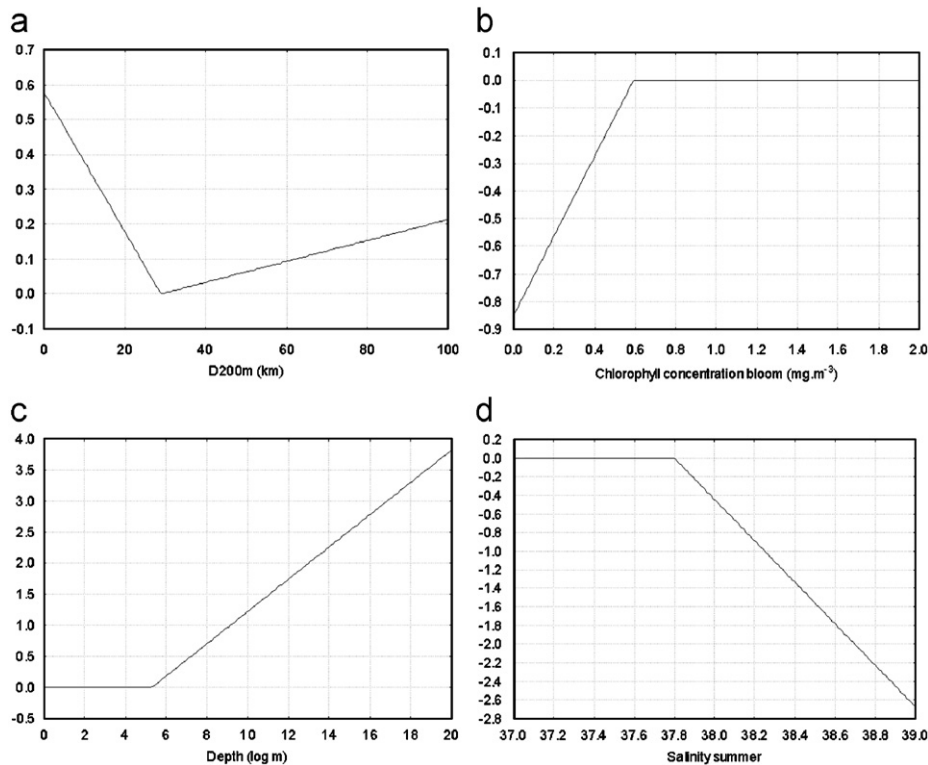
Fig. 3. Relevant basic functions (bf) relating the presence and absence of sperm whales and the EGVs highlighted by the MARS model: (a) distance to the 200-m contour, bfs 1 and 2, (b) chlorophyll concentrations in summer, bf 3 (c) depth, bf 4 and (d) salinity in summer, bf 5 (summer: summer period, bloom: phytoplankton bloom period).

However, our objectives were here to attempt a global description of the sperm whale habitat in a temporally and spatially heterogeneous area, the entire NWMS, and to compare the ability of different modelling methods in such heterogeneous area.

As a gap occurs between the primary production and cephalopods, the main prey of sperm whale, we modelled the summer habitat of this species using hydrological and biological EGVs from the summer and the phytoplankton bloom periods. In our results the chlorophyll concentration of the phytoplankton bloom period was highlighted as significant EGVs. Furthermore, for the PCA, ENFA and MARS method, the eigenvalues of this EGVs were higher than those of the summer chlorophyll concentration. The chlorophyll concentration of the phytoplankton bloom period seems then relevant to reflect the summer sperm whale habitat.

Our work modelled the summer sperm whale habitat using data from June to August. The habitat modelling of the sperm whale distribution in autumn, winter and spring would require extensive surveys in order to increase data on this species during these seasons. However, sperm whale has been observed, throughout the year, in both continental slope and offshore waters in the Ligurian Sea (Laran and Drouot-Dulau, 2007; Laran and Gannier, 2006) and in the Liguro-Provençal Basin (Praca, 2008).The seasonal distribution of the sperm whale seems therefore close to what we modelled for the summer (see below).

4.2. Accuracy of the modelling methods

The HS of the sperm whale was modelled using four methods, and the accuracy of each method predicting whale distribution was compared. ENFA, GLM and MARS provided models significantly different from the random model, while the PCA model was not significantly different from it. Considering the statistical accuracy, the best model was provided by MARS with an AUC of 0.79, followed by GLM (0.70), ENFA (0.66) and PCA (0.58).

The four models highlighted the influence of both topographical and hydrological EGVs to characterize sperm whale distribution. The chlorophyll concentrations of the phytoplankton bloom period and the distance to the 200-m contour were the only EGVs highlighted as important by all four models, but the selected EGVs showed either positive or negative influence according to the different models. However, the differences in the EGV selection could be due to the computation techniques used by the modelling methods, which are intrinsically different. Comparing the relevant EGVs of the different models could be meaningless, and the ecological significance of the models will thereafter be compared using the HS maps that they produced.

Disparities were observed between the different HS maps. Only the PCA model highlighted a preferred area on the continental shelf between Corsica and Sardinia and on the continental slope north of Corsica. Only the MARS model did not show an important habitat on the
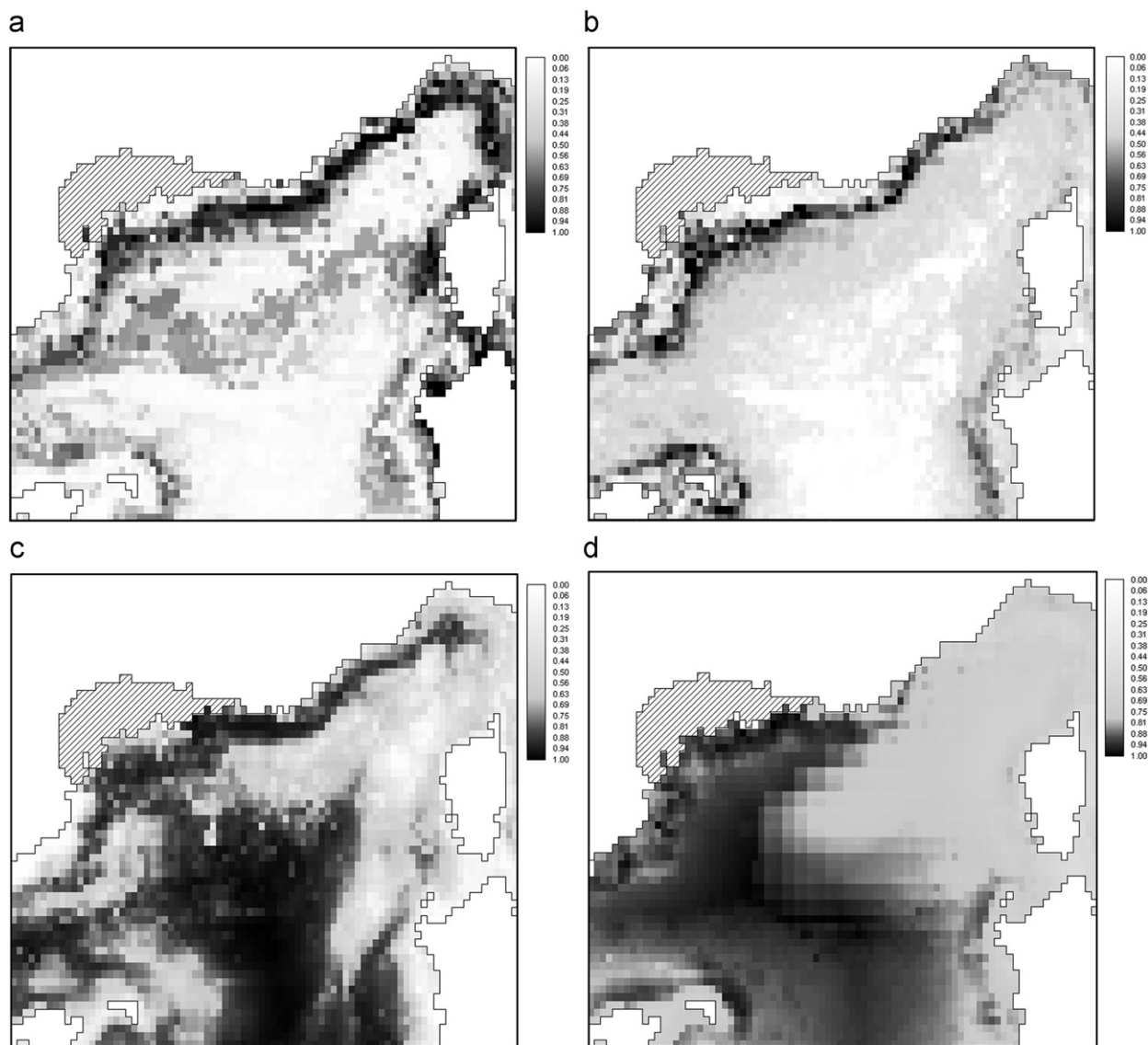
Fig. 4. PCA (a), ENFA (b), GLM (c) and MARS (d) habitat suitability maps of the sperm whale (the hatched area was removed from the analysis). Colours from dark grey to black indicated the predicted presence area.

Table 5
Statistical differences between the AUCs and *p*-values of the *Z*-tests comparing the modelling methods: PCA, ENFA, GLM and MARS.

	Differences	*p*-Values
PCA vs. ENFA	0.07	0.315
PCA vs. GLM	0.11	0.070
PCA vs. MARS	0.20	0.007
ENFA vs. GLM	0.04	0.545
ENFA vs. MARS	0.13	0.097
GLM vs. MARS	0.09	0.118

continental slope near the French coast, in the Ligurian Sea. Important differences also appeared for the offshore waters. The PCA model showed a small area, with moderate probability of presence, between Corsica and Spain, and the ENFA model highlighted a decreasing importance of habitat from continental slope to offshore waters, whereas both the GLM and the MARS models showed a wide area, with significant presence probabilities, in the western offshore part of the study area.

However, the continental slope was shown as a suitable area for the sperm whale by the four HS maps. The influence of this topographic feature on the sperm whale distribution has been widely documented in other parts of the world such as the Gulf of Mexico (Baumgartner et al., 2001; Davis et al., 1998), the North Atlantic Ocean (Hamazaki, 2002; Waring et al., 2001) and the Alboran Sea (Cañadas et al., 2002).

On the other hand, only GLM and MARS models presented high probabilities of presence in the western offshore waters. The influence of hydrological features,

such as fronts or eddies, have been shown (e.g. Biggs et al., 2000; Gannier and Praca, 2007; Waring et al., 2001), but at smaller spatial and temporal scales than in our study. For example, in the NWMS, Gannier and Praca (2007) used only data at depths greater than 2000 m and a weekly temporal scale. Our results are the first, to our knowledge, to model the offshore habitat of the sperm whale at such a scale in the Mediterranean Sea. First, this habitat is certainly related to the presence of a spatially fluctuating but permanent front: the North Balearic Front. It is the convergence zone of the southern modified warm Atlantic waters and the northern deep and cold upwelled waters (Le Vourch et al., 1992). Furthermore, the anti-clockwise circulation in the NWMS brings nutrient-rich waters of the Ligurian Sea to the west. Upwelling is also frequent in the Gulf of Lions in relation to the occurrence of Mistral and Tramontane winds (Millot, 1999; Millot and Wald, 1990). These hydrological features enrich the waters of the Provençal Basin and favour the development of the food web (Le Vourch et al., 1992), probably attracting cephalopods, the main prey of sperm whales in the Mediterranean Sea (Astruc and Beaubrun, 2005; Roberts, 2003). However, as the effort in this part of the study area was less than the effort on the continental slope, the offshore habitat of the sperm whale should be confirmed by additional surveys in offshore waters.

In the eastern part of the study area, PCA, ENFA and GLM highlighted an important habitat only on the continental slope, while the MARS model showed a more diffuse habitat in the whole Ligurian Sea. Several studies in this area have shown that there is no significant distinction of the sperm whale distribution between the continental slope and the offshore waters (Gannier et al., 2002; Gordon et al., 2000; Laran, 2005). The Ligurian current and fronts, combined with a nearby steep slope (Le Vourch et al., 1992; Millot, 1999), probably result in a disperse distribution of the cephalopods and then of the sperm whales in the whole area. This phenomenon is then better reflected in the MARS model.

This study has highlighted the difficulty of modelling the habitat of cetaceans in a wide and heterogeneous region, such as sperm whales in the NWMS. Such opportunistic species are very mobile and inhabit different areas with variable characteristics. The results showed a clear disparity between presence-only models and presence–absence models. Indeed, the models provided by the presence–absence methods had better statistical accuracies, highlighted the habitat on the continental slope and better reflected the offshore habitat of the sperm whale. They seem to reflect better the complex distribution of cetaceans, and as for other organisms, such as plants (Zaniewski et al., 2002), birds (Brotons et al., 2004) or insects (Gallego et al., 2004), their use is recommended. In particular, MARS allows the development of complex non-linear relationships between the presence–absence data and the EGVs, but uses simple piecewise basis functions. Furthermore, MARS models can be produced rapidly and are easily comprehended (Leathwick et al., 2006). Presence–absence data collection requires significant effort of observation. As cetaceans are difficult to observe, the achievement of such effort at sea involves long surveys and significant expenditure of resources. In particular, the use of acoustic is an important tool to enlarge data set. Presence-only methods could be a good alternative in regions where conservation and management decisions have to be taken, but where extensive surveys are not feasible. Indeed, presence-only methods allow the use of opportunistic data sets (e.g. collected by recreational sailors, fishermen or whale-watching industry) or the merging of coherent presence data from different surveys (collected by different institutions, with different protocols). Although the ENFA is limited by the use of average data, this method yielded a better statistical accuracy than PCA and introduces an ecological meaning in the factorial axes computation. ENFA should therefore be recommended in situations where data set are insufficient to provide the basis of presence–absence methods.

References

Aguilar, A., Borrell, A., Reijnders, P.J.H., 2002. Geographical and temporal variation in levels of organochlorines contaminants in marine mammals. Mar. Environ. Res. 53, 425–452.

Antoine, D., André, J.-M., Morel, A., 1996. Oceanic primary production: 2. Estimation at global scale from satellite (coastal zone color scanner) chlorophyll. Global Biogeochem. Cycles 10 (1), 57–69.

Astruc, G., Beaubrun, P., 2005. Do Mediterranean cetacean diets overlap for the same resources? Eur. Res. Cet. 19, 81.

Baumgartner, M.F., Mullin, K.D., May, L.N., Leming, T.D., 2001. Cetacean habitats in the northern Gulf of Mexico. Fish. Bull. 99, 219–239.

Beck, J.R., Schultz, E.K., 1986. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch. Pathol. Lab. Med. 110, 13–20.

Biggs, D.C., Leben, R.R., Ortega-Ortiz, J.G., 2000. Ship and satellite studies of mesoscale circulation and Sperm Whale habitats in the Northeast Gulf of Mexico during GulfCet II. Gulf Mexico Sci. 18 (1), 15–22.

Boyce, M., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K., 2002. Evaluating resource selection functions. Ecol. Model. 157, 281–300.

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

Cañadas, A., Sagarminaga, R., Garcia-Tiscar, S., 2002. Cetacean distribution related with depth and slope in the Mediterranean waters off southern Spain. Deep-Sea Res. I 49, 2053–2073.

Conan, P., Pujo-Pay, M., Raimbault, P., Leveau, M., 1998. Variabilité hydrologique et biologique du Golfe du Lion. II. Productivité sur le bord interne du courant. Oceanol. Acta 21, 767–782.

Davis, R.W., Fargion, G.S., May, N., Leming, T.D., Baumgartner, M., Evans, W.E., Hansen, L.J., Mullin, K., 1998. Physical habitat of cetaceans along the continental slope in the north-central and western Gulf of Mexico. Mar. Mamm. Sci. 14, 490–507.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Di Natale, A., Notarbartolo di Sciara, G., 1994. A review of the passive fishing nets and trap fisheries in the Mediterranean Sea and the cetacean bycatch. Rep. Int. Whaling Comm. 44 (Special Issue 15), 189–202.

Drouot, V., Gannier, A., Goold, J.C., 2004. Diving and feeding behavior of sperm whales (*Physeter macrocephalus*) in the northwestern Mediterranean Sea. Aquat. Mamm. 30 (3), 419–426.

Duguy, R., 1991. Les mammifères marins de la Méditerranée occidentale. Bull. Soc. Zool. Fr. 114 (3), 89–96.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24 (1), 38–49.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19, 1–141.

Gallego, D., Canovas, F., Esteve, M.A., Galian, J., 2004. Descriptive biogeography of *Tomicus* (Coleoptera: Scolytidae) species in Spain. J. Biogeogr. 31, 2011–2024.

Gannier, A., 1997. Estimation de l'abondance estivale du Rorqual commun *Balaenoptera physalus* (Linné, 1758) dans le bassin Liguro-provençal (Méditerranée Nord-Occidentale). Rev. Ecol. (Terre & Vie) 52, 69–86.

Gannier, A., 2006. Le peuplement estival de cétacés dans le Sanctuaire Marin Pelagos (Méditerranée nord-occidentale): distribution et abondance. Mammalia 70 (1), 17–27.

Gannier, A., Praca, E., 2007. SST fronts and the summer sperm whale distribution in the north-west Mediterranean Sea. J. Mar. Biol. Assoc. UK 8 (1), 187–193.

Gannier, A., Drouot, V., Goold, J.C., 2002. Distribution and relative abundance of sperm whales in Mediterranean Sea. Mar. Ecol. Prog. Ser. 243, 281–293.

Gordon, J.C.D., Moscrop, A., Carlson, C., Ingram, S., Leaper, R., Matthews, J., Young, K., 1998. Distribution, movements and residency of sperm whales off the Commonwealth of Dominica, Eastern Caribbean: implication for the development and regulation of the local whale-watching industry. Rep. Int. Whaling Commn. 48, 551–557.

Gordon, J.C.D., Matthews, J.N., Panigada, S., Gannier, A., Borsani, F.J., Notarbartolo di Sciara, G., 2000. Distribution and relative abundance of striped dolphins, and distribution of sperm whales in the Ligurian Sea cetacean sanctuary. J. Cetacean Res. Manage. 2 (1), 27–36.

Gregr, E.J., Trites, A.W., 2001. Predictions of critical habitat for five whale species in the waters of coastal British Columbia. Can. J. Fish. Aquat. Sci. 58, 1265–1285.

Guisan, A., Zimmermann, N.E., 2001. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Hamazaki, T., 2002. Spatiotemporal prediction models of cetacean habitats in the mid-western North Atlantic Ocean (from Cape Hatteras, North Carolina, USA, to Nova Scotia, Canada). Mar. Mamm. Sci. 18 (4), 920–939.

Hirzel, A., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83 (7), 2027–2036.

Hirzel, A., Hausser, J., Perrin, N., 2006. Biomapper 3.2. Laboratory for Conservation Biology. University of Lausanne URL: ⟨http://www.u-nil.ch/biomapper⟩.

IOC-IHO-BODC, A., 2003. Centenary Edition of the GEBCO Digital Atlas. The Intergovernmental Oceanographic Commission, The International Hydrographic Organization and the British Oceanographic Data Centre, Liverpool, UK.

Jaquet, N., 1996. How spatial and temporal scales influence understanding of sperm whale distribution: a review. Mamm. Rev. 26, 51–65.

Laran, S., 2005. Variations spatio-temporelles du peuplement de cétacés en Mer Ligure (Méditerranée Nord-Occidentale) et relations avec les conditions environnementales. Ph.D. Thesis, Free University of Brussels, Belgium.

Laran, S., Drouot-Dulau, V., 2007. Seasonal variation of striped dolphins, fin- and sperm whales' abundance in the Ligurian Sea (Mediterranean Sea). J. Mar. Biol. Assoc. UK 87, 345–352.

Laran, S., Gannier, A., 2006. Variation saisonnière de la présence du cachalot (*Physeter macrocephalus*) dans le Sanctuaire Pelagos (Mer Ligure). Mésogée 61, 71–77.

Le Vourch, J., Millot, C., De Stephanis, R., Castagné, N., Le Borgne, P., Olry, J.P., 1992. Atlas des fronts thermiques en Méditerranée d'après l'imagerie satellitaire. Mémoires de l'Institut Océanographique.

Leathwick, J.R., Elith, J., Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptative regression splines for statistical modelling of species distributions. Ecol. Model. 199, 188–196.

Legendre, P., Legendre, L., 1998. Numerical Ecology. Elsevier, Amsterdam.

Mangion, P., Gannier, A., 2002. Improving the comparative distribution picture for Risso's dolphin and long-finned pilot whale in the Mediterranean Sea. Eur. Res. Cet. 16, 68.

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

Millot, C., 1999. Circulation in the Western Mediterranean Sea. J. Mar. Syst. 20, 423–442.

Millot, C., Wald, L., 1990. Upwelling in the Gulf of Lions. Coast. Upw., 160–166.

MODB, MEDAR/MEDATLAS II Database, data available at ⟨http://modb.oce.ulg.ac.be/medar/medar.html⟩.

Notarbartolo di Sciara, G., Gordon, J.C.D., 1997. Bioacoustics: a tool for the conservation of cetaceans in the Mediterranean Sea. Mar. Freshwater Behav. Physiol. 30, 125–146.

OceanColor. Data available at ⟨http://oceancolor.gsfc.nasa.gov⟩.

PO.DAAC. Physical Oceanography Distributed Active Archive Center. Data available from ⟨http://poet.jpl.nasa.gov/⟩.

Praca, E., 2008. Ecological niches of the teuthophageous odontocetes in the northwestern Mediterranean Sea. Ph.D. Thesis, University of Liège, Belgium.

Redfern, J.V., Ferguson, M.C., Becker, E.A., Hyrenbach, K.D., Good, G., Barlow, J., Kaschner, K., Baumgartner, M., Forney, K.A., Ballance, L.T., Fauchald, P., Halpin, P., Hamazaki, T., Pershing, A.J., Qian, S.S., Read, A.J., Reilly, S.B., Torres, L., Werner, F., 2006. Techniques for cetacean-habitat modeling. Mar. Ecol. Prog. Ser. 310, 271–295.

Roberts, S.M., 2003. Examination of the stomach contents from a Mediterranean sperm whale found south of Crete, Greece. J. Mar. Biol. Assoc. UK 83, 667–670.

Tabachnick, B.G., 2000. Using Multivariated Statistics. Allyn and Bacon.

Teloni, V., 2005. Patterns of sound production in diving sperm whales in the northwestern Mediterranean. Mar. Mamm. Sci. 21 (3), 446–457.

Waring, G.T., Hamazaki, T., Sheehan, D., Wood, G., Baker, S., 2001. Characterization of Beaked Whale (Ziphiidae) and Sperm Whale (*Physeter macrocephalus*) summer habitat in shelf-edge and deeper waters off the northeast U.S. Mar. Mamm. Sci. 17 (4), 703–717.

Watwood, S.L., Miller, P.J.O., Johnson, M., Madsen, P.T., Tyack, P.L., 2006. Deep-diving foraging behaviour of sperm whales (*Physeter macrocephalus*). J. Anim. Ecol. 75, 814–825.

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