Comparison of habitat models for scarcely detected species

Auriane Virgili*, Mélanie Racine, Matthieu Authier, Pascal Monestiez, Vincent Ridoux

Centre d’Études Biologiques de Chizé—La Rochelle, UMR 7372 CNRS—Université de La Rochelle, Institut du Littoral et de l’Environnement, 17000 La Rochelle, France
Observatoire PELAGIS, UMS 3462 CNRS—Université de La Rochelle, Systèmes d’Observation pour la Conservation des Mammifères et des Oiseaux Marins, 17000 La Rochelle, France
BioSP, INRA, 84914 Avignon, France

Article info
Article history:
Received 11 July 2016
Received in revised form 19 December 2016
Accepted 21 December 2016
Available online 10 January 2017

Keywords:
Habitat modelling
Scarce species
GAM
GLM
MaxEnt
Poisson
Negative binomial
Tweedie
Zero-inflated Poisson

Abstract
When performing habitat models, modellers have to choose between presence-absence and presence-only models to estimate the habitat preferences of a species. Primarily, this choice depends on the data that are available and whether effort data are recorded in parallel to sighting data. For species that are rare or scarce, the models have to address a great number of zeros (i.e., no animal seen) that weakens the ability to make sound ecological inferences. We tested two types of habitat models (presence-absence vs. presence-only) to determine which type best dealt with datasets containing an excess of zeros, and we applied our models to a sighting dataset that included the common (Delphinus delphis) and striped (Stenella coerulea) dolphin (approximately 92% zeros). We used two types of presence-absence models (Generalised Additive models – GAMs, Generalised Linear Model – GLM) and one presence-only model, a MaxEnt model, and we used various criteria to compare these models (i.e., AIC, deviances, rootograms and distribution patterns predicted by the models). Overall, we observed that the presence-absence models made better predictions than the presence-only model. Among the presence-absence models, the GAM with a Negative Binomial distribution was better at predicting small delphinids habitats, even though the GAM with a Tweedie distribution exhibited similar results. However, the zero-inflated Poisson distributions exhibited less convincing results and was contrary to what was expected. Finally, despite 92% zeros, our dataset was not zero-inflated. Our study demonstrates the importance of selecting appropriate models to make reliable predictions of habitat use for species that are rare or scarce.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Identifying habitat needed and used by species is important for wildlife management and conservation (Cañadas et al., 2005; Bailey and Thompson, 2009). One means of identifying habitat is with statistical models that correlate the spatial distributions of animal sightings with environmental inputs (Austin, 2002; Guisan and Thuiller, 2005; Redfern et al., 2006). Such models allow the habitat of a species and presence to be estimated. They also allow for predictions in areas that have not been previously surveyed (Segurado et al., 2004).

Species distribution models have recently undergone rapid development and have been used for diverse applications (e.g., Elith et al., 2006; Elith and Leathwick, 2009; Mannocci et al., 2014a, 2014b, 2015). There are generally two categories of habitat models: presence-absence and presence-only models; the chosen model depends on the type of data used and, notably, whether effort data are recorded in parallel to sighting data (Guisan and Zimmermann, 2000).

The first group of models requires presence and effort data that are recorded during planned surveys, where each on-effort sighting represents a detection of the target species. Such presence-absence models include, among others, generalised linear models (GLM), generalised additive models (GAM), regression trees analyses such as boosted regression trees (BRT) (Guisan and Zimmermann, 2000; Brotons et al., 2004), or occupancy models (MacKenzie et al., 2002). Some of these models allow estimating detection probability, and consequently, prediction of habitat suitability of a species (Gormley et al., 2011). They also allow functional relationships to be fitted between species locations and local environmental conditions (Guisan and Zimmermann, 2000). The models of the second group...
only require detection data, such as opportunistic data, where the absence data are missing because effort data were not documented and non-detection data are not prospected and informed (Hirzel et al., 2002). These include Ecological Niche Factor Analysis (ENFA) or Maximum Entropy Modelling (MaxEnt), and allows for the identification of potentially suitable sites by evidencing the environmental conditions that are similar to the sites where animals were recorded (Elith et al., 2006). Nevertheless, the accuracy of presence-only model outputs is conditional on random or representative sampling of the habitat at the data collection stage (Yackulic et al., 2013). Presence-only data are the default option when data on absence (that is effort data) are not available (Zaniowski et al., 2002).

Except for presence-only models, which do not consider the zeros, choosing among presence-absence models might be difficult depending on the studied species, particularly when focusing on scarcely detected species, because of the inherent difficulty of models to accommodate a large number of absences. Due to restricted habitat range, low density and poor detection even in favourable habitats (Martin et al., 2005), the number of absences in some datasets (i.e., the zeros) can be large. True (or structural – the taxon is really absent from an area), and false (or sampling – the option when data on absence are not available) zero-inflated models are often over-dispersed, and second, depending on their data, they have different models: GAMs with a Poisson, a Negative Binomial, a Tweedie and a zero-inflated Poisson distribution; a GLM with a zero-inflated Poisson distribution and a presence-only model; the MaxEnt model. Due to their ability to model separately the absences and the presences (Lambert, 1992), we assumed a priori that a zero-inflated Poisson model would perform best. However, the Negative Binomial and Tweedie distributions can also provide good fits (Warton, 2005; Dunn and Smyth, 2005; Lindén and Mantyniemi, 2011). In addition, with its multiple applications (Yackulic et al., 2013), including those by managers, and its ability to take into account the complex interactions between response and predictor variables (Elith et al., 2006, 2011; Phillips et al., 2004; Phillips and Dudík, 2008), the MaxEnt model appears to be a relevant tool for modelling habitats of rare species (Wisz et al., 2008). Therefore, we also tested the model to assess its efficiency. This study aims to pragmatically answer some questions commonly asked by habitat modellers, such as those regarding the effective zero-inflation of their data and the relevance of the chosen model depending on their data.

2. Materials and methods

2.1. Datasets

2.1.1. Aerial surveys and data collection

The small delphinids sighting data were recorded during the SAMM survey (Suivi Aérien de la MégaFaune Marine; Aerial Census of Marine MegaFauna), which was dedicated to the observation of marine megafauna and conducted in the northeast Atlantic Ocean and the northwest Mediterranean Sea (Laran et al., in review; Lambert et al., in press). In the present work we focused on data collected in the summer of 2012 in the entire English Channel and the Bay of Biscay from the tip of Brittany to the Dover Strait in the north, and to the Spanish coast in the south (Fig. 1). The survey was carried out from mid-August to early August along 31,427 km of transect lines. A standard methodology for cetacean surveys was applied (Hammond et al., 2013) using twin-engine high-wing aircrafts equipped with bubble windows. The flights followed a zig-zag pattern, at a speed of 167 km/h and an altitude of 183 m. Observation conditions (Beaufort sea state, turbidity, cloud cover and glare severity) and sightings with group size were recorded following a line-transect methodology (Buckland et al., 2001). This implies that the angle between every sighting and the track line was recorded to estimate the Effective Strip Width (ESW; see the small delphinids detection function and estimated ESW in Laran et al., in review).

The common and striped dolphins were pooled because it was most often impossible to tell apart the two species from the plane. During the survey, 277 sightings of small delphinids were recorded in good observation conditions, corresponding to 14,477 individuals (Fig. 1).

2.1.2. Environmental predictors

To model the relationships between small delphinids and their environment, we used eight environmental predictors (Lambert et al., in press; Virgili et al., in review), of which there were two physiographic variables (depth and slope) and six oceanographic variables (mean, variance and gradient of Sea Surface Temperature—SST, mean and standard deviation of Sea Surface Height—SSH, and the maximum velocity of tidal currents (Table 1)). All oceanographic variables were computed at a seven day resolution, i.e., averaged over 6 days prior to the sampled day. Physiographic variables are static and relate to the bathymetry, and oceanographic variables are dynamic and describe water masses.
Fig. 1. Study area (A) showing small delphinids sightings (B) recorded during the survey. The study area covers the English Channel and the Bay of Biscay. Surveys were carried out along transects (dotted lines) following a zig-zag pattern, and sightings were classified by group size with each point representing a group of individuals (1, 2–20, 20 to 100, 20–100, and 100–700 individuals).

Table 1

<table>
<thead>
<tr>
<th>Environmental predictors</th>
<th>Sources*</th>
<th>Effects on pelagic ecosystems of potential interest to top predators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physiographic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depth (m)</td>
<td>A</td>
<td>Shallow waters could be associated with high primary production</td>
</tr>
<tr>
<td>Slope (°)</td>
<td>A</td>
<td>Associated with currents, high slope induce prey aggregation and/or primary production increasing</td>
</tr>
<tr>
<td>Oceanographic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean of SST (°C)</td>
<td>B</td>
<td>Variability over time and horizontal gradients</td>
</tr>
<tr>
<td>Variance of SST (°C)</td>
<td>B</td>
<td>of SST reveal front locations, potentially</td>
</tr>
<tr>
<td>Mean gradient of SST (°C)</td>
<td>B</td>
<td>associated to prey aggregations</td>
</tr>
<tr>
<td>Mean of SSH (m)</td>
<td>C</td>
<td>High SSH is associated with high mesoscale activity and</td>
</tr>
<tr>
<td>Standard deviation of SSH (m)</td>
<td>C</td>
<td>prey aggregation and/or primary production increase</td>
</tr>
<tr>
<td>Daily maximum intensity of the currents (m s⁻¹)</td>
<td>D</td>
<td>High currents induce water mixing and prey aggregation</td>
</tr>
</tbody>
</table>

These variables have been considered in the model selection procedure because they are all candidate drivers of small delphinids distribution via their effect on the functioning of pelagic ecosystems (Table 1, Appendix 1).

2.2. Statistical models

2.2.1. Analytical strategy

In a first step, we arbitrarily chose a baseline model, a GAM with a Poisson distribution appropriate for equi-dispersed data, for comparison with the other models. For this baseline model, relationships between the abundances of small delphinids and environmental variables were investigated. Next, we fitted GAMS with a Negative Binomial and a Tweedie distribution, which are suitable for over-dispersed data, a GAM and a GLM with a zero-inflated Poisson distribution, which are suitable when over-dispersion is due to zero-inflation, and a MaxEnt model, which is specific to presence-only data. Using these models, we applied the variables previously selected in the baseline model. Even if a model performance is largely determined by its selected variables (Syphard and Franklin, 2009), in this study, we applied the same variables for each tested model to assess how the results were affected by the model alone when using the same dataset. To finish, we compared all models by using different criteria such as the Akaike Information Criterion (AIC), the deviances, the rootograms (Kleiber and Zeileis, 2016) and the predicted density maps to evaluate the predictive performance of each model.

2.2.2. Baseline model: GAM with a Poisson distribution

A Generalised Additive Model (GAM; Hastie and Tibshirani, 1986) with a Poisson distribution (variance equal to the mean), hereafter called PO-GAM, was retained as the baseline model (Table 2). The response variable was linked to the additive predictors using a log-link function. We included, as an offset, the effort per segment (Hastie and Tibshirani, 1986). This offset was calculated as the segment linear length multiplied by twice the ESW (Effective Strip Width estimated from Conventional Distance Sampling, see Laran et al., in review). The model was fitted using R-3.1.2 (R Core Team, 2016) with the mgcv package (Wood, 2006, 2013) by restricting polynomial smoothness to three degrees of freedom (Ferguson et al., 2006).

In the selection procedure, all models with a combination of one to four variables were tested and the combinations of variables with a correlation coefficient higher than |0.7| were excluded. A maximum of four covariates was implemented to avoid excessive complexity and difficulty of interpretation (Mannocci et al., 2014a, 2014b). The Akaike Information Criterion (AIC) was used to select the best models, where the lower the AIC the better the model (Akaike, 1974). Finally, we extracted the explained, null and resid-
Table 2
Details of the models used in the study. GAM: Generalised Additive Model; GLM: Generalised Linear Model; PO: Poisson; NB: Negative Binomial; TW: Tweedie; ZIP: Zero-Inflated Poisson; PA: Presence-Absence data; and AIC: Akaike Information Criterion. *R Core Team (2016).

<table>
<thead>
<tr>
<th>Generic models</th>
<th>Used names</th>
<th>Data</th>
<th>Settings and details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalised Additive Model with Poisson distribution</td>
<td>PO-GAM</td>
<td>Equi-dispersed PA</td>
<td>Used R-3.1.2*, package mgcv, function GAM, Poisson distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions, used AIC to select the best model</td>
</tr>
<tr>
<td>Generalised Additive Model with Negative Binomial distribution</td>
<td>NB-GAM</td>
<td>Over-dispersed PA</td>
<td>Used R-3.1.2*, package mgcv, function GAM, Negative Binomial distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions, no selection of the best model</td>
</tr>
<tr>
<td>Generalised Additive Model with Tweedie distribution</td>
<td>TW-GAM</td>
<td>Over-dispersed PA</td>
<td>Used R-3.1.2*, package mgcv, function GAM, Tweedie distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions, no selection of the best model</td>
</tr>
<tr>
<td>Generalised Additive Model with Zero-Inflated Poisson distribution</td>
<td>ZIP-GAM</td>
<td>Zero-inflated PA</td>
<td>Used R-3.1.2*, package mgcv, function GAM, ZIP distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions, no selection of the best model</td>
</tr>
<tr>
<td>Generalised Linear Model with Zero-Inflated Poisson distribution</td>
<td>ZIP-GLM</td>
<td>Zero-inflated PA</td>
<td>Used R-3.1.2*, package mgcv, function GAM, ZIP distribution log-link function, included an offset, no smooth, no selection of the best model</td>
</tr>
<tr>
<td>Maximum Entropy Modelling</td>
<td>MaxEnt</td>
<td>Presence-only</td>
<td>Used MaxEnt software version 3.3.3, no selection of the best model, hinge feature, default prevalence of 0.5, logistic output format</td>
</tr>
</tbody>
</table>

2.2.3. Challenger models

For all other models (Table 2), we used the variables associated with the best selected model following 2.3.1, and there was no covariate selection procedure. As for the baseline model, we extracted the explained, null and residual deviances and checked the distribution of residuals for each model, except for the MaxEnt model. In this part, the models are briefly described, but they are more developed in Appendix 2.

For the GAM with a Negative Binomial and a Tweedie distribution, we used the R package mgcv (Wood, 2006, 2013) with the gam, nb and tw functions to estimate the model parameters and the relationship between environmental variables and small delphinids densities. Hereafter, the fitted model will be called NB-GAM and TW-GAM.

Next, we tested two types of zero-inflated models, with linear (GLM) and nonlinear (GAM) relationships between the response variable and the predictors and considering a zero-inflated Poisson distribution. We used the mgcv package (Wood, 2006, 2013) with the gam and zip functions to estimate the parameter of the models. Hereafter, the fitted models will be called ZIP-GAM and ZIP-GLM.

Finally, we fitted a presence-only model with Maximum Entropy (MaxEnt), in which relationships with the environment are estimated using background samples of the environment instead of absence locations (Elith et al., 2011). We used the MaxEnt version 3.3.3 (http://www.cs.princeton.edu/~schapire/maxent/; Phillips et al., 2006). The input file was the same as the baseline model, but we removed all absences; hence, each line corresponded to one observation of small delphinids and for the environmental predictors, we used the four covariates selected by the baseline model. Regarding model parameters, we used the “hinge” feature to generate models with smooth functions similar to GAM ones, with a default prevalence of 0.5 and a logistic output format to compare it to the probability of presence (Phillips and Dudík, 2008; Elith et al., 2011; Merow et al., 2013).

2.2.4. Predictions

For each fitted model, except for MaxEnt, which directly provides a prediction map, we computed the predictions and their associated uncertainty maps for each day of the survey (85 days) on a 0.05° × 0.05° resolution grid. Next, daily predictions were averaged over the entire period to produce maps of averaged density of small delphinids. Finally, we provided uncertainty maps that corresponded to the standard errors of the predictions. To limit extrapolation, all predictions were constrained within the envelope of sampled values of covariates used to fit the model.

In addition, we assessed whether a prediction was an extrapolation or an interpolation using the non-parametric Gower’s distance (King and Zeng, 2007). An extrapolation is a prediction for a combination of covariate values that falls outside the convex hull which is defined by the covariate data used to calibrate the model (King and Zeng, 2007; Authier et al., 2016). However, even if a prediction falls outside this convex hull, this extrapolation can nevertheless be informed by calibration data lying in its neighbourhood. The neighbourhood of a prediction was defined as the calibration covariate data within a radius of one geometric mean Gower’s distance of the prediction (King and Zeng, 2007). The geometric mean was computed from all pairs of calibration data point. The results from this extrapolation analysis were mapped to visually assess how trustworthy the predictions were.

2.3. Model comparison

Evaluating the predictive performance of a model requires demonstrating its consistency with raw observation data and comparing the outputs of several models (Pearce and Ferrier, 2000). Each assessment criterion quantifies a particular aspect of a model performance and several criteria must be used in combination (Elith and Graham, 2009). We calculated different selection measures to improve the relevance of model comparison.

First, an Akaike Information Criterion was computed for each model to assess model relative fit: the lower the AIC, the better the model (Akaiche, 1974). Second, we examined several deviance-based quantities (null, residual and explained) as a proxy of the model reliability to predict the frequencies of species occurrence (Elith and Graham, 2009). A high explained deviance can indicate a good fit, whereas a high null deviance and a high residual deviance can indicate a bad one. Finally, to evaluate the absolute goodness-of-fit of the models and how they handled the excess of zeros, we plotted rootograms that compared, with histograms, the raw data frequencies to the frequencies fitted with the models (Kleiber and Zeileis, 2016).

The methods cannot be readily applied with presence-only models, which leads to some complexity in the methods of model comparison. To evaluate the predictive performance of MaxEnt, we used the Area Under the receiver operating characteristic Curve (AUC; Elith et al., 2006). This method works only on binary data (not on count data) and measures how a model can differentiate the sites where the species is present and the sites where it is absent. A perfect discrimination of the sites is revealed by a score of 1,
a discrimination equivalent to a random distribution is indicated by a score of 0.5 and for a score lower than 0.5, the model performance is worse than a random guess (Elith et al., 2006). This AUC is directly provided by the MaxEnt software. However, with this method, we cannot compare the model performance to the fitted baseline model. We thus transformed the PO-GAM prediction maps (only this one) to probability of presence with the formula: presence probability = 1 – e^−predicted density

3. Results

3.1. Model selection

Among the eight environmental predictors, the variables selected by the best PO-GAM, defined as the baseline model, were the mean and variance of SST and the mean and standard deviation of SSH (Fig. 2). The highest densities of delphinids were predicted for stable temperatures at approximately 16 °C (variance around 0 °C), and a rather stable low average altimetry (SSH, around −0.5 m, standard error around 0.5 m).

The NB-GAM (k = 0.028), TW-GAM (p = 1.573) and ZIP-GAM (θ = −4.861, 0.262) showed fairly similar smooth functions compared to the PO-GAM, except for sd SSH (Fig. 2). However, confidence intervals around the functional relationships were significantly smaller and the predicted densities were higher in the case of the ZIP-GAM. The smooth functions of the ZIP-GLM showed increasing small delphinids densities with increasing SST mean and SSH mean and decreasing densities with decreasing SST variance and sd SSH, which was an opposite trend compared to the other models for SSH mean.

To complete the comparison of the models, we analysed the residuals of each fitted model (Appendix 3). In all cases, there was an accumulation of residuals at zero and an over-dispersion of positive values, but it was less important for the TW-GAM. In addition, we calculated Cook’s distances (Appendix 4) to determine if some values highly influenced the fitted models (Cook’s distance > 1). It appeared that some values greatly influence the PO-GAM and ZIP-GLM. However, for NB-GAM, TW-GAM and ZIP-GAM, no value appeared to affect the models (Cook’s distance < 1) and the only values that could influence them correspond to non-extreme values of covariates. Consequently, that strengthened the results provided by the fitted models, especially for NB-GAM, TW-GAM and ZIP-GAM.

Finally, with an AUC of 0.822, the MaxEnt model predicted delphinids presence probabilities much better than a random prediction would do (AUC of 0.5).

3.2. Predictions

Prediction maps of the PO-GAM showed a concentration of delphinids in offshore waters, from the continental shelf to the oceanic waters, with the highest densities over the slope (Fig. 3). The highest densities, which reached 30 individuals km\(^{-2}\), were predicted in the north of Galicia, which is outside the survey area. In addition, we noticed a good match between observations and predictions of the model (Fig. 1). Within the survey area, predictions were associated with low uncertainties (Appendix 6), which strengthened the results. In contrast, outside the survey area, patches of high densities were associated with higher uncertainties and needed to be considered with caution.

The TW-GAM predicted exactly the same distribution as the previous model but with slightly higher densities (maximum at 35 individuals km\(^{-2}\); Fig. 3, Appendix 5). The NB-GAM also predicted the same distribution as PO-GAM but with higher densities (maximum at 73 individuals km\(^{-2}\), Fig. 3, Appendix 5). The ZIP-GAM showed the same distribution patterns in the Bay of Biscay but with lower densities (maximum at 11 individuals km\(^{-2}\), Fig. 3, Appendix 5) and more individuals predicted near the coasts. However, contrary to the PO-GAM, this model predicted delphinids in the western English Channel, with a concentration of individuals around the Channel Islands (Appendix 5). Regarding the ZIP-GLM, densities were also predicted in offshore waters, approximately 5 and 10 individuals km\(^{-2}\) and similar to the other models, but a larger patch was identified and located west of the Isle of Wight with more than 2000 individuals km\(^{-2}\) (Fig. 3, Appendix 5).

Similarly to the PO-GAM, high predicted densities of the NB-GAM and TW-GAM were associated with high uncertainties outside the survey area but low uncertainties in the survey area. For ZIP-GAM and ZIP-GLM, patches of high densities in the survey area were associated with uncertainties, making the predictions less reliable (Appendix 6).

Extrapolation and neighbourhood maps (Fig. 4) allowed us to assess the reliability of the predictions obtained with the fitted models. Overall, a high percentage of extrapolation and a low percentage of calibration data used to inform neighbouring cells (neighbourhood) indicated unreliable predictions. Hence, a model that predicted densities in the English Channel was inconsistent, which was particularly the case with ZIP-GLM and to a lesser extent, the case with ZIP-GAM. PO-GAM, NB-GAM and TW-GAM all predicted high densities outside the survey area, but according to Fig. 4, these predictions were reliable because they were informed by approximately 20% of the data used to calibrate the model. However, NB-GAM made more extreme extrapolations than the other models in the Bay of Biscay.

The MaxEnt model predicted higher probabilities of occurrence in the Bay of Biscay, particularly over the slope but also fairly evenly spread along the coasts of the Bay of Biscay and southwest England (Fig. 5). Compared to the PO-GAM (see prediction map in probability of presence, Fig. 5), MaxEnt hardly extrapolated beyond the sampled area and the predicted probabilities of presence were lower.

3.3. Evaluation and comparison of models

The NB-GAM showed the lowest AIC and was followed by TW-GAM, whereas the PO-GAM showed the highest (Table 3). The explained deviances varied between 7.3% for the ZIP-GLM and 39.1% for the TW-GAM (Table 3). In addition, the lowest null and residual deviances were computed for the NB-GAM, which indicated a better fit of the model compared to TW-GAM that, despite a high explained deviance, showed very high null and residual deviances (Table 3). The ZIP models performed worse than NB-GAM or TW-GAM but better than PO-GAM. Overall, the NB-GAM showed a better predictive performance than the other models.

Visible with rootograms (Fig. 6), all fitted models were not adequate for counts between 0 and 5 individuals. In all cases, the observed frequencies (blue bars) were not included in the confidence intervals (red dots). PO-GAM, NB-GAM and TW-GAM did not predict enough zeros and predicted too many sightings from 1 to 5 individuals. Likewise, ZIP-GAM and ZIP-GLM did not predict enough zeros and predicted too many sightings beyond 5 individuals. Albiet all fitted models tended to over-predict the frequencies between 1 and 5 individuals, the highest number of observed frequencies included in the confidence interval was observed for the NB-GAM thus making it the best fitted model.
Fig. 2. Forms of smooth functions for the selected covariates for each presence-absence model. The solid line in each plot is the smooth function estimate and shaded regions represent approximate 95% confidence intervals. The y-axis indicates the logarithm of the abundance in individual/km². The x-axis indicates the values of the covariates and zero on the x-axis indicates no effect of the covariate. Best model fits are between the vertical lines indicating the 10th and 90th quantiles of the data.

Table 3
Indices used for the comparison of the presence-absence models. AIC: Akaike Information Criterion.

<table>
<thead>
<tr>
<th></th>
<th>PO-GAM</th>
<th>NB-GAM</th>
<th>TW-GAM</th>
<th>ZIP-GAM</th>
<th>ZIP-GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>70,082</td>
<td>4284</td>
<td>5869</td>
<td>25,438</td>
<td>28,001</td>
</tr>
<tr>
<td>Explained deviance</td>
<td>28.6%</td>
<td>38.4%</td>
<td>39.1%</td>
<td>17.1%</td>
<td>7.3%</td>
</tr>
<tr>
<td>Null deviance</td>
<td>96,266</td>
<td>1001</td>
<td>41,341</td>
<td>27,305</td>
<td>27,146</td>
</tr>
<tr>
<td>Residual deviance</td>
<td>68,742</td>
<td>616</td>
<td>25,168</td>
<td>25,164</td>
<td></td>
</tr>
</tbody>
</table>

4. Discussion

4.1. General

We compared different types of habitat models, particularly presence-absence and presence-only models, to choose one that would be the most suitable for a scarce species. We found that a GAM with a Negative Binomial distribution was the most appropriate model to fit the data, even though the GAM with a Tweedie distribution also showed good predictions. In contrast, the zero-inflated Poisson distributions showed less convincing results, which was contrary to what was expected. We also found that MaxEnt provided quite good results compared to PO-GAM. These findings suggest that habitat for species that are rare or seldom seen are best described using presence-absence models such as GAM with a Negative Binomial distribution (Warton, 2005). However, it is important to also recognise that we used a particular biological system and we did not test all possible models.

4.1.1. Biological system

Small delphinids, defined here as common and striped dolphins taken together, were selected as study material for several reasons. First, they provided a large enough dataset to allow proper statistical analyses. The present dataset included 277 sightings corresponding to approximately 14,477 individuals; consequently, the models tested here could be fitted without generating error messages during optimisation. Second, this dataset included more than 90% zero observations, which could be suggestive of zero-inflation. Third, in a previous investigation, Lambert et al. (in press) showed that small delphinids have well-defined patterns of distribution over the Bay of Biscay, which helped in evaluating predictive accuracy.
4.1.2. Baseline model

In this study, we chose a panel of eight environmental predictors, both static and dynamic, as they are considered proxies for primary production and consequently, prey distribution (Austin, 2002). To compare the habitat models of small delphinids, we used a GAM with a Poisson distribution (PO-GAM) as a baseline model. The effectiveness of GAMs has been repeatedly demonstrated (Ferguson et al., 2006; Vilchis et al., 2006; Becker et al., 2010; Mannocci et al., 2014a, 2014b) and we chose a Poisson distribution as the baseline because it characterises equi-dispersed data. Equi-dispersion is expected in the idealised situation where each detection event is independent of the others. In addition, GAMs are commonly used because they are more flexible than GLMs thanks to their semi-parametric functions that can accommodate non-linear relationships between animal densities and environmental predictors (Becker et al., 2010). We then established a variable selection procedure to define the best fitted model to be used as the baseline. This fitted model showed a relatively high explained deviance (28.6%) and interpolated rather than extrapolated (Fig. 4). The selected model appeared ecologically consistent, which strengthened our model choice. A preference of eastern North Atlantic small delphinids for waters warmer than 15 °C and depths between 400
Fig. 4. Extrapolation analysis using Gower’s distance (King and Zeng, 2007). The extrapolation map (A) assesses whether a prediction was an extrapolation (100%) or an interpolation (0%) and the neighbourhood map (B) represents the percentage of calibration covariate data which informed each cell.

Fig. 5. Distributions predicted by PO-GAM (A) versus MaxEnt (B) in the Bay of Biscay and the English Channel. Dotted lines represent the survey area. The same scale was applied for the two model to facilitate the comparison.

and 1000 m and with concentrations along the shelf edge and lower densities in the western Channel and in coastal waters of the Bay of Biscay has been shown (Figs. 2 and 3) and already mentioned by Cañadas et al. (2009), MacLeod et al. (2009), Murphy et al. (2013) and Lambert et al. (in press).

4.1.3. Challenger models choice

The choice of the challenger models was an important step in the comparison process. For the presence-absence models, the aim was not to test all existing models but to answer the pragmatic question: “What type of model should we use if the dataset contains more than 90% zeros?” To answer this question, we built realistic models that included linear (GLM) or non-linear (GAM) relationships between the response variable and predictors and tested different structural choices for the data likelihood (Tweedie, Negative Binomial and ZIP). All these models handle differently the datasets with extra zeros (Appendix 2). A ZIP model links two sub-models: a binomial model for the zero count that distinguishes between true and false absences; and a Poisson count model for non-zero observations. Conditional on an observation not being a true absence, equi-dispersion is assumed. True absences are in this case the only source of over-dispersion. Tweedie and Negative Binomial models directly include an assumption on the relationship between the mean and the variance, in that they address over-dispersion in a more phenomenological way because the micro-level process generating over-dispersion is not explicit (Dunn and Smyth, 2005; Ridout et al., 1998; Zeileis et al., 2007; Wenger and Freeman, 2014).

Variable selection was only done on the baseline model. This could have led to sub-optimal models for the other likelihood choices (Tweedie, Negative Binomial, etc.). The performance of a model is largely determined by its selected variables (Syphard and Franklin, 2009). We decided to hold the set of covariates constant over models to assess how the results were affected by the structural choice in the model only. This corresponds to the idealised situation where the habitat of a species is known from previous investigations, but there is uncertainty in the exact model structure to predict its habitat. In practice, variable selection depends also on the model structure.

To assess the robustness of our results, we also ran a variable selection procedure for each model (not shown). For ZIP-GAM, the same variables were selected, but for NB-GAM and TW-GAM, the gradient of SST was selected over the variance of SST (two variables for which the potential effect on the pelagic ecosystems is quite similar; Table 1). All other variables were identical. The model with the four variables selected by the baseline model (PO-GAM) was the second model in the two cases. All GAMs were almost identical with respect to the set of selected covariates. The biggest difference was observed in ZIP-GLM where the mean and variance of SST were replaced by the slope and gradient of SST; the model with the four variables selected in the baseline model was the 13th model. A complete comparison (Appendix 7) revealed that the predictions of all the “best” GAMs were similar, but the predictions of the ZIP-GLM were greatly different from the rest and they were more likely to be extrapolations.
4.1.4. MaxEnt

We also wanted to test a presence-only model because it is not expected to be bad for scarce species as it does not see the zeros and it is easier to fit since corresponding data are more readily available (Tsoar et al., 2007). Indeed, most available data of species distribution are presence-only records because they are easier to collect, and contrary to presence-absence data, do not require recording effort data (Tsoar et al., 2007). In the case of rare and elusive species, opportunistic data, a common example of presence-only data, often represent the largest set of available data (Pearce and Boyce, 2006). Hence, it appeared necessary for habitat modellers, who have to choose between several models, to know if presence-only models could provide good predictions of species distribution relative to other more comprehensive methods based on presence-absence data (which requires effort data to be properly recorded). Among all presence-only models (Elith et al., 2006; Tsoar et al., 2007; Monk et al., 2010), we chose the MaxEnt model because it appeared more suitable to model the predictions of species distribution with complex interactions between the response and the predictor variables (Elith et al., 2006, 2011; Phillips et al., 2004; Phillips and Dudík, 2008) and seemed to manage datasets characterised by scarce data well (Wisz et al., 2008).

4.2. Pragmatic habitat modelling of scarce species

Although environmental variables used in the models were identical, each fitted model showed a different predictive performance based on its own characteristics. Overall, NB-GAM and TW-GAM were very similar in the improvement they provided over PO-GAM (Appendix 3) and estimated similar non-linear relationships with environmental covariates. NB-GAM exhibited the best predictive performances with the smallest AIC and a moderate explained deviance. Habitat predictions from models PO-GAM, TW-GAM and NB-GAM were qualitatively similar, suggesting robustness with respect to extrapolation (Fig. 4) and consistency in the results but predicted densities were larger in magnitude with NB-GAM. The overall bad performance of the only GLM among the candidate set of models stressed the importance of non-linear relationships in habitat modelling of small delphinids in the Bay of Biscay. Thanks to their flexibility, GAMs are appropriate for modelling the distribution of sparsely distributed megafauna either marine or terrestrial (Wood, 2006; Becker et al., 2010; Hegel et al., 2010). Thus, NB-GAM and TW-GAM were able to fit the data well despite the huge number of zeros, as seen on the rootograms (Fig. 6).

Fitted ZIP models showed lower explained deviances (17.1% for ZIP-GAM and 7.3% for ZIP-GLM), lower predictive performances (higher AIC) and less ecologically consistent predictions with extrapolation of the predicted densities. Indeed, ZIP-GAM and ZIP-GLM predicted large densities of small delphinids in the English Channel where no sightings were recorded. Moreover, previous studies evidenced that these species generally avoid this area (Cañadas et al., 2009; MacLeod et al., 2009; Murphy et al., 2013; Lambert et al., in press). The disappointing performance of ZIP-GAM was somewhat surprising. We expected, following Barry and Welsh (2002), a better performance of this model because it mixes a zero-inflated model with the non-parametric functions of a GAM. In fact, the results were less convincing than NB-GAM or TW-GAM results because of the lower explained deviance, higher AIC and unrealistic densities predicted in the English Channel. This is likely due to the current parametrisation of the ZIP family in mgcv. In fact, the current parametrisation uses the linear predictors and linearly scales them on a logit scale to generate extra-zero observations (see the help pages in mgcv v1.8–9; Wood, 2013). This parametrisation implicitly assumes that areas with lower densities have a higher probability of non-detection, which is a priori reasonable. However, it does not allow for incorporating detection-specific covariates which may better explain non-detection patterns. Despite 92% zeros in the data, ZIP models showed worse results than NB-GAM or TW-GAM; over-dispersion was not mainly due to zero-inflation. Even though the best model we selected did not completely accom-
moderate all the zero observations, suggesting some zero-inflation (Fig. 6), the latter was arguably less prevalent than initially thought.

MaxEnt showed a fairly high predictive performance (AUC = 0.82) and distribution patterns relatively similar to those of PO-GAM, albeit more spread out in the whole study area. However, this model underestimated the probabilities of presence compared to PO-GAM and did not extrapolate beyond the study area. This presence-only model appeared relatively efficient to establish distribution patterns in a given survey area (Tsoar et al., 2007) and to identify areas of high probabilities of presence when only presence data were available (Zaniewski et al., 2002; Cromeley et al., 2011). However, this may result more from the sampling design than from MaxEnt modelling per se. Data were collected with a standard protocol that ensured almost uniform coverage over the Bay of Biscay; no area was over- or under-sampled. The main issue with presence-only models is that they cannot account for uneven effort and must assume that the sampling of the habitat was random in order to interpret MaxEnt predictions correctly (Yackulic et al., 2013). Thus, our results might give too much of an optimistic outlook of the performance of MaxEnt. To moderate that optimism, cross-validation with portions of the study area removed “in block” would have been useful but was beyond the scope of this study.

Finally, as Warton (2005) warned, “many zeros does not mean zero-inflation” of the data, and even 92% zeros does not necessarily mean zero-inflation. We would recommend to habitat modellers, even if they study scarce species, to first test over-dispersed models such as GAMs with Tweedie or Negative Binomial distributions before testing zero-inflated models. Obviously, the predictive performance of the model has to be assessed for the tested model. A useful visual method to assess whether a model adequately addresses many zeros is the rootogram (Minami et al., 2007; Kleibe and Zeileis, 2016).

4.3. Management applications

Habitat models can be useful to delineate marine protected areas (Cañadas et al., 2005; Lambert et al., in review). These models allow investigating species habitat preferences (Austin, 2002) and revealing contiguous areas of high predicted densities, thereby highlighting potential areas of conservation (Cañadas et al., 2005). They can accommodate low sampling effort and remain useful in identifying suitable habitat despite few recorded sightings. For example, due to the zig-zag pattern of the SAMM survey transects, areas on the continental slope were not entirely prospected but habitat models predicted relatively high densities throughout the stratum without extrapolating. Habitat models can also help with sampling gaps that might necessitate further extensive effort to validate predictions (Bailey and Thompson, 2009).

Scarce detected species present additional challenges for habitat modelling. Due to the small number of sightings compared to the deployed effort, it can be difficult to obtain reliable predictions and establish conservation plans for these species. However, these sparsely distributed species may face many threats and require conservation action plans. We outlined in this study a pragmatic approach to build habitats models when focusing on scarce or rare species.

5. Conclusions

Modelling the habitats of cetaceans or large predators in general is challenging because these organisms are by nature sparsely distributed compared to lower trophic levels, and their detection is often imperfect. This situation results in scarce datasets when survey effort is low to modest, and heavy zero-rich datasets when the amount of survey effort is large. However, statistical models generally require large presence-absence datasets to fit count data to environmental predictors. It is arguably easier for managers to use presence-only data rather than presence-absence data because effort data are not required.

Thanks to a homogeneous subjacent effort, the MaxEnt model, a presence-only model, provided relatively good results in this study. Despite its lower accuracy, it would provide good enough predictions for small delphinids presence in the Bay of Biscay, although its ability to predict outside the survey area seemed limited. Among the presence-absence models, non-linear models predicted best small delphinids habitat. Contrary to what was expected, zero-inflated models were not the best predictive models; we thought that with 92% zeros, the data would be zero-inflated, but a thorough analysis revealed that they were mostly over-dispersed and not zero-inflated. Our study shows the importance of selecting appropriate models (beyond variable selection) to make reliable predictions of habitat use for species that are rare or scarce and that an abundance of zeros does not necessarily mean zero-inflation.

Acknowledgments

We are grateful to the French Ministry in charge of the environment (Ministère de l’Environnement, de l’Energie et de la Mer) and the Agency for marine protected areas (Agence des aires marines protégées) for funding the project. We thank Hélène Falchetto for processing the survey data that were used in the study. We are grateful to PREVIMER for providing us outputs from MARS-2D and MARS-3D models. We thank the Direction Générale de l’Armement (DGA) for a financial support to this study and funding A. Virgili’s PhD. We thank Charlotte Lambert and Andrew Trites for their very helpful comments on this manuscript. We are grateful to reviewers for their comments on the article.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2016.12.013.

References
