Modelling the ecological niche of cetaceans: new perspectives and applications

Tese de Doutoramento

Marc Fernández Morrón

Doutoramento em

Biologia

Ponta Delgada
2017

Modelling the ecological niche of cetaceans: new perspectives and applications

Tese de doutoramento

Marc Fernández Morrón

Orientadores:

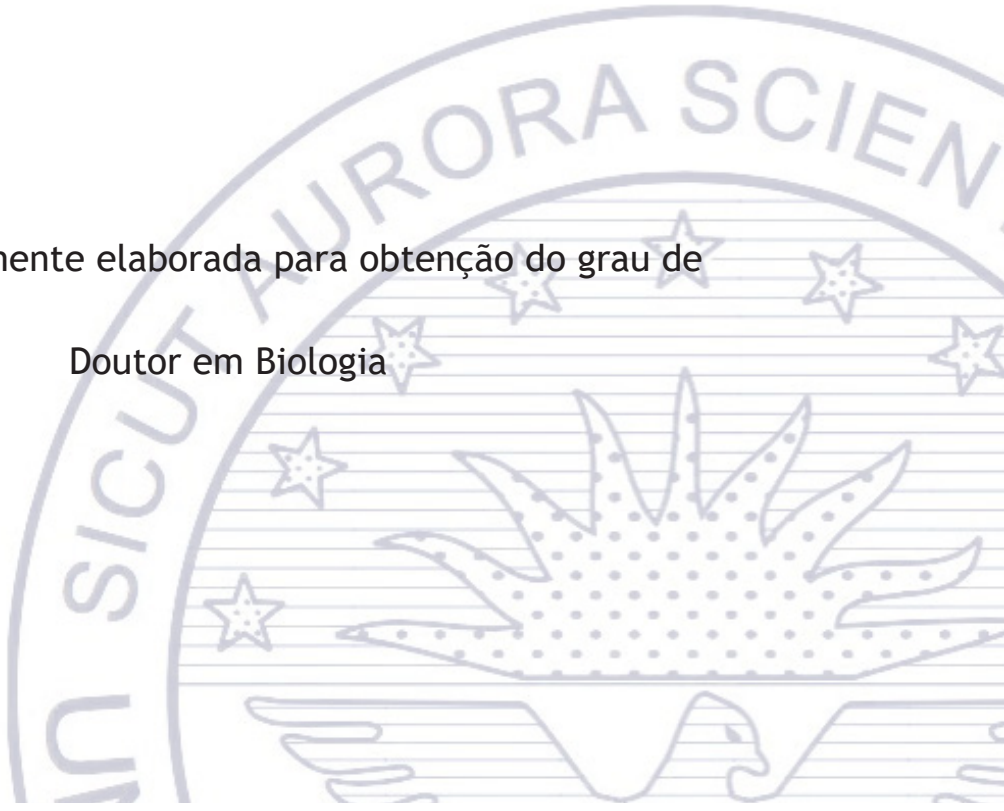
Professor José Manuel Viegas Oliveira Neto Azevedo

Doutor Chris Yesson

Doutor Alexandre Gannier

Tese especialmente elaborada para obtenção do grau de

Doutor em Biologia



"…essentially, all models are wrong, but some are useful"
Box & Draper (1987)

Marc Fernandez was supported by grant M3.1.2/F/028/2011 from the Fundo Regional para a Ciência e Tecnologia (Azores Government). This research was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT - Foundation for Science and Technology, under the project "PEst-C/MAR/LA0015/2013", by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT - Foundation for Science and Technology and European Regional Development Fund (ERDF), in the framework of the program PT2020 and by cE3c funding (Ref: UID/BIA/00329/2013). It was also partly supported by CIRN (Centro de Investigação de Recursos Naturais, University of the Azores), and CIIMAR (Interdisciplinary Centre of Marine and Environmental Research, Porto, Portugal.

CONTENTS

CONTENTS

CONTENTS ..... **I**

LIST OF FIGURES: ..... **IV**

LIST OF TABLES: ..... **XIII**

ACKNOWLEDGEMENTS ..... **XV**

RESUMO **1**

ABSTRACT **3**

1 INTRODUCTION **7**

1.1 SPECIES DISTRIBUTION AND NICHE MODELING: THEORETICAL FRAMEWORK 7

What is the niche? *7*

Grillenian vs Eltonian niche *8*

Niches and geographic distributions *9*

1.2 NICHE MODELLING 11

Species occurrence data *11*

Enviromental data *13*

Modelling techniques *14*

1.3 NICHE MODELING AND DISTRIBUTION STUDIES FOR CETACEANS 16

The marine environment specificities *16*

Data sources for cetacean distributional studies *17*

Modelling cetaceans distributions *18*

1.4 RESEARCH OBJECTIVES 20

1.5 THESIS OUTLINE 20

Chapter 2: THE AZORES: A COMPLEX GEOGRAPHICAL AND OCEANOGRAPHIC SYSTEM *20*

Chapter 3: Enviromental and occurrences datasets *20*

Chapter 4: A THEORETICAL APPROACH TO MODELLING MOBILE SPECIES IN DYNAMIC ENVIRONMENTS *21*

Chapter 5: USING A HIGH TEMPORAL RESOLUTION OPPORTUNISTIC DATASET TO MODEL THE ECOLOGICAL NICHE OF CETACEANS *21*

Chapter 6: DISCUSSION *21*

2 THE AZORES: A COMPLEX GEOGRAPHICAL AND OCEANOGRAPHIC SYSTEM 25
2.1 GEOGRAPHY 25
2.2 OCEANOGRAPHY 26
The Azores confluence zone *27*
3 ENVIRONMENTAL AND OCCURRENCE DATASETS 31
3.1 ENVIROMENTAL LAYERS 31
Static layers *32*
Dynamic variables *36*
3.2 OCCURRENCE LAYERS 37
Dedicated survey dataset *37*
Opportunistic dataset *41*
4 A THEORETICAL APPROACH TO MODELLING MOBILE SPECIES IN DYNAMIC ENVIRONMENTS 55
4.1 INTRODUCTION 55
4.2 METHODS 56
Environmental data *56*
Virtual species: occurrence data *58*
Sampling survey design *63*
Temporal aggregations *65*
Modelling approaches *65*
Model evaluation *66*
4.4 RESULTS 67
Variable contributions *67*
Train and test AUC results *69*
Suitability projections *70*
2.5 DISCUSSION 72
Temporal resolution of dynamic variables *73*
Relationship between spatial and temporal scales *74*
5 USING A HIGH TEMPORAL RESOLUTION OPPORTUNISTIC DATASET TO MODEL THE ECOLOGICAL NICHE OF CETACEANS .. 79
5.1 INTRODUCTION 79
5.2 METHODS 80
ENVIRONMENTAL AND OCCURRENCE DATA *80*
Temporal and spatial resolutions *80*

Modelling techniques and evaluation procedures *81*
5.3 RESULTS 87
Model performance *87*
Species patterns *92*
5.4 DISCUSSION 143
Caveats and bias *143*
Spatial and temporal scale *144*
Sampling bias corrections *144*
Implications for species ecology *145*
6 GENERAL DISCUSSION 149
6.1 MODELLING CETACEAN DISTRIBUTIONS 150
6.2 OPPORTUNISTIC DATA FOR NICHE MODELLING OF CETACEANS 152
6.3 SUGGESTIONS FOR FUTURE STUDIES 153
How can niche models for cetaceans be improved? *153*
Dataset merging to produce cetacean niche models *155*
6.4 CONSERVATION IMPLICATIONS 155
7 REFERENCES 157

LIST OF FIGURES:

Figure 1.1 The BAM diagram (Soberon & Peterson 2005), depicting the interaction between abiotic (A), biotic (B) and movement (M) factors. G represents the entire geographic area under consideration, G_A the abiotically suitable area, G_0 the occupied distributional area, and G_I the invadable distributional area. Black circles indicate presences, white circles indicate absences..... 10

Figure 1.2 Probabilistic events leading to a presence or absence, modified from Peterson et al. (2011). Each bar represents a choice, with red filled circles representing "no" and green representing "yes". The first column represents the movement area, the Abiotic and Biotic columns represent the areas suitability for both cases, the Sampled columns represents if the area is sampled or not. Finally the Observation column reflects whether an observation is correctly performed or not..... 12

Figure 1.3 Distance to frontal thermal areas in the Azores archipelago area in two consecutive weeks (26/6/15 to 4/7/15 and 4/7/15 to 11/7/15). Distances are expressed in kilometers. Fronts were derived from MUR SST daily products..... 17

Figure 2.1 Bathymetry of the Azores. Black lines represent the 1000, 2000 and 3000m depth lines. Reddish colors represent shallower areas. Data extracted from the EMODnet portal..... 26

Figure 2.2 Large-scale circulation of the North-Atlantic depicted by OSCAR-derived mean (2004–2014) sea surface currents are represented as isolines over $\log^{10}$ of the current velocity in order to improve the representation of the extremes. The known ocean circulation currents are marked: GS, Gulf Stream; GSB, Gulf Stream Bifurcations; AC, Azores Current; PC, Portugal Current; CC, Canary Current; the dashed-line box is centered around the Azores and depicts the region where most of the historical analysis were carried out. Extracted from Caldeira & Reis (2017). 27

Figure 2.3 Overall mean (2004–2014) of Eddy Kinetic Energy ($cm^2\ s^{-2}$), emphasizing the main regional oceanographic processes: the Azores Current (AC), the Westward propagating Eddies (WE) and the Gulf Stream GS). Extracted from Caldeira & Reis (2017)..... 28

Figure 3.1 Canyons-like features (black shapes) from the Central and Eastern Group of the Azores area derived from the EMODnet Bathymetry using the Topographic Position Index. Grey areas represent the hill-shaded bottom relief. 34

Figure 3.2 Static of environmental variables used for the analysis (depth, slope, bottom curvature, distance to the 200m bathymetric line and

distance to the 1,000m bathymetric line) for the study area. The curvature variable is calculated as an index and therefore has no units. Distance to canyon-like features are depicted in Fig. 3.1. 35

Figure 3.3 Median of the dynamic environmental variables used in the present study (Sea Surface Temperature, distance to major frontal areas and surface Chlorophyll-a). Data for the second week of July 2014 is shown as an example. 37

Figure 3.4 Tracks and sightings from the 2013 survey 40

Figure 3.5 Tracks and sightings from the 2014 survey. 41

Figure 3.6 Distribution of sightings on the Azores archipelago (down left) and mainland South-West Portugal (top right). Grey polygons comprise all the observations; lines represent the 1,000 and 2,000 m bathymetric lines in the Azores and 200 and 1,000 m in mainland Portugal. 43

Figure 3.7 Annual observation effort of the MONICET database, expressed as monthly mean number of trips (± standard deviation) across all the years. 43

Figure 3.8 Classification tool on the photo-identification interface. The screen shows the two main photos being compared, together with a carrousel of photos of individuals filtered using the Characterization tool. 46

Figure 3.9 Sightings of short-beaked common dolphin (Delphinus delphis) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 48

Figure 3.10 Sightings of sperm whales (Physeter macrocephalus) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 48

Figure 3.11 Sightings of bottlenose dolphins (Tursiops truncatus) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 49

Figure 3.12 Sightings of Atlantic spotted dolphins (Stenella frontalis) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 49

Figure 3.13 Sightings of fin whales (Balaenoptera physalus) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 50

Figure 3.14 Sightings of Risso's dolphins (Grampus griseus) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 50

Figure 3.15 Sightings of short-finned pilot whales (Globicephala macrorhynchus) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 51

Figure 3.16 Sightings of sei whales (Balaenoptera borealis) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 51

Figure 3.17 Sightings of striped dolphins (Stenella coeruleoalba) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 52

Figure 3.18 Sightings of blue whales (Balaenoptera musculus) from MONICET dataset from 2009 to 2015. Black lines represent the 1000, 2000 and 3000 bathymetric lines. 52

Figure 4.1 Variable contribution for dynamic species: Sea Surface Temperature (left) and distance to thermal fronts (right). Figure extracted from Fernandez et al. 2017. 60

Figure 4.2 Theoretical suitability maps for the Dynamic species on 55 consecutive days of 2013 summer (July-August). Meaningful differences between suitability areas found thought the study period. Figure extracted from Fernandez et al. 2017. 60

Figure 4.3 Variable contribution for the Static species: depth (left), distance to the 200m bathymetric line (centre) and sea floor slope (right). Figure extracted from Fernandez et al. 2017. 61

Figure 4.4 Theoretical suitability maps for the Static species on 55 consecutives days of 2013 summer (July-August). No differences are present between suitability areas thought the study period. Figure extracted from Fernandez et al. 2017. 62

Figure 4.5 Variable contributions for pseudoreal species: Sea Surface Temperature (left), slope (center), and distance to thermal fronts (right). Figure extracted from Fernandez et al. 2017. 62

Figure 4.6 Theoretical suitability maps for the Pseudoreal species on 55 consecutive days of 2013 summer (July-August). Meaningful differences are found between suitability areas thought the study period. Figure extracted from Fernandez et al. 2017. 63

Figure 4.7 Study area map (Eastern Group and Central Group, Azores Archipelago) with the virtual transects used for the niche modelling calculations. Figure extracted from Fernandez et al. 2017. 64

Figure 4.8 Study area map (São Miguel Island, Azores Archipelago) with the tracks used as effort measure for the non-linear niche modelling calculations. Figure extracted from Fernandez et al. 2017. 65

Figure 4.9 Results of variable selection for the three temporal aggregations (daily, weekly and monthly – in rows), two models algorithms (GLM and GBM), and 3 virtual species (dynamic, static and pseudoreal – in columns). Results of the GBM models are expressed as mean variable contribution over the 1000 iterations according to variable relative importance. Results of the GLM are expressed as the number of times a specific variable was selected for the model after the AIC stepwise selection procedure. Figure extracted from Fernandez et al. 2017. 68

Figure 4.10 Results for the training and testing AUC for a linear survey using sampling data for the GBM and GLM model algorithms (rows), and the three temporal grain selections (daily, weekly and monthly) and 3 virtual species (dynamic, static and pseudoreal), (columns). AUC ranges from 0 to 1. Figure extracted from Fernandez et al. 2017. 69

Figure 4.11 Results for the test AUC for a non-linear survey of the two models algorithms (GLM and GBM), for the three temporal grain selections (daily, weekly and total) and for the 3 virtual species (dynamic, static and pseudoreal). AUC ranges between 0 to 1. Figure extracted from Fernandez et al. 2017. 70

Figure 4.12 Niche overlap (Schoener's D index) between the theoretical "real" niche and that predicted for an area outside the one used to build the model. Results for the 2 modelling approaches used (GLM and GBM), for the 3 temporal scenarios (daily, weekly and monthly) and for the 3-virtual species (static, dynamic and pseudoreal). Figure extracted from Fernandez et al. 2017. 71

Figure 4.13 Niche overlap (Schoener's D index) between the theoretical "real" niche and that predicted for an area outside the one used to build the model. Results for the 2 modelling approaches used (GLM and GBM), for the 3 temporal scenarios (daily, weekly and monthly) and for the 3-virtual species (static, dynamic and pseudoreal). 71

Figure 5.1 Minimum Sampled Area (grey grid) for a randomly selected trip in the Central Group of the Azores Archipelago. The black line represents the Minimum Convex Polygon, the circles represents all the sightings for that trip. The grid represented is in a 2x2 km scale. 82

Figure 5.2 Sperm whale group detectability function for the distance to the closest lookout, with sea state and company as covariates. 83

Figure 5.3 Baleen whale group detectability function for the distance to the closest lookout, with sea state and company as covariates. 84

Figure 5.4 Big dolphin group detectability function for the distance to the closest lookout, with sea state and company as covariates. 84

Figure 5.5 Small dolphin group detectability function for the distance to the closest lookout, with sea state and company as covariates. 85

Figure 5.6 Comparison of general AUC test results for all the potential scenarios: background (targeted background, T, and non-targeted background, NT) versus temporal scales of environmental variables (8-day and 1 month) versus spatial scale of environmental variables (2km and 4km). Significant differences calculated using a Kruskal-Wallis with a Nemenyi post-hoc test are noted with letters. 88

Figure 5.7 Cumulative response curves for the main environmental variables influencing the distribution of sperm whales. Mean values and standard deviation for 10 runs are presented. 93

Figure 5.8 Mean monthly suitability values for sperm whales around São Miguel Island, Azores. 94

Figure 5.9 Monthly standard deviation of suitability predictions for sperm whales around São Miguel Island, Azores. 95

Figure 5.10 Mean monthly suitability values for sperm whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 96

Figure 5.11 Monthly standard deviation of suitability predictions for sperm whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 97

Figure 5.12 Cumulative response curves for the main environmental variables influencing the distribution of short-finned pilot whales. Mean values and standard deviation for 10 runs are presented. 98

Figure 5.13 Mean monthly suitability values for short-finned pilot whales around São Miguel Island, Azores. 99

Figure 5.14 Monthly standard deviation of suitability predictions for short-finned pilot whales around São Miguel Island, Azores. 100

Figure 5.15 Mean monthly suitability values for short-finned pilot whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 101

Figure 5.16 Monthly standard deviation of suitability predictions for short-finned pilot whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 102

Figure 5.17 Cumulative response curves for the main environmental variables influencing the distribution of Risso's dolphins. Mean values and standard deviation for 10 runs are presented. 103

Figure 5.18 Mean monthly suitability values for Risso's dolphins around São Miguel Island, Azores. 104

Figure 5.19 Standard monthly deviation of suitability predictions for Risso's dolphins around São Miguel Island, Azores. 105

Figure 5.20 Mean monthly suitability values for Risso's dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 106

Figure 5.21 Monthly standard deviation of suitability predictions for Risso's dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 107

Figure 5.22 Cumulative response curves for the main environmental variables influencing the distribution of striped dolphins. Mean values and standard deviation for 10 runs are presented. 108

Figure 5.23 Mean monthly suitability values for striped dolphins around São Miguel Island, Azores. 109

Figure 5.24 Monthly standard deviation of suitability predictions for striped dolphins around São Miguel Island, Azores. 110

Figure 5.25 Mean monthly suitability values for striped dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 111

Figure 5.26 Monthly standard deviation of suitability predictions for striped dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 112

Figure 5.27 Cumulative response curves for the main environmental variables influencing the distribution of Atlantic spotted dolphins. Mean values and standard deviation for 10 runs are presented. 113

Figure 5.28 Mean monthly suitability values for Atlantic spotted dolphins around São Miguel Island, Azores. 114

Figure 5.29 Monthly standard deviation of suitability predictions for Atlantic spotted dolphins around São Miguel Island, Azores. 115

Figure 5.30 Mean monthly suitability values for Atlantic spotted dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 116

Figure 5.31 Monthly standard deviation of suitability predictions for Atlantic spotted dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores.....117

Figure 5.32 Cumulative response curves for the main environmental variables influencing the distribution of common dolphins. Mean values and standard deviation for 10 runs are presented. 118

Figure 5.33 Mean monthly suitability values for short-beaked common dolphins around São Miguel Island, Azores. 119

Figure 5.34 Monthly standard deviation of suitability predictions for short beaked common dolphins around São Miguel Island, Azores. 120

Figure 5.35 Mean monthly suitability values for short beaked common dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 121

Figure 5.36 Monthly standard deviation of suitability predictions for short beaked common dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 122

Figure 5.37 Cumulative response curves for the main environmental variables influencing the distribution of bottlenose dolphins. Mean values and standard deviation for 10 runs are presented. 123

Figure 5.38 Mean monthly suitability values for bottlenose dolphins around São Miguel Island, Azores. 124

Figure 5.39 Monthly standard deviation of suitability predictions for bottlenose dolphins around São Miguel Island, Azores. 125

Figure 5.40 Mean monthly suitability values for bottlenose dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 126

Figure 5.41 Monthly standard deviation of suitability predictions for bottlenose dolphins around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 127

Figure 5.42 Cumulative response curves for the main environmental variables influencing the distribution of blue whale. Mean values and standard deviation for 10 runs are presented. 128

Figure 5.43 Mean monthly suitability values for blue whales around São Miguel Island, Azores. 129

Figure 5.44 Monthly standard deviation of suitability predictions for blue whales around São Miguel Island, Azores. 130

Figure 5.45 Mean monthly suitability values for blue whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 131

Figure 5.46 Monthly standard deviation of suitability predictions for blue whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 132

Figure 5.47 Cumulative response curves for the main environmental variables influencing the distribution of sei whale. Mean values and standard deviation for 10 runs are presented. 133

Figure 5.48 Mean monthly suitability values for Sei whales around São Miguel Island, Azores. 134

Figure 5.49 Monthly standard deviation of suitability predictions for Sei whales around São Miguel Island, Azores. 135

Figure 5.50 Mean monthly suitability values for Sei whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 136

Figure 5.51 Monthly standard deviation of suitability predictions for Sei whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 137

Figure 5.52 Cumulative response curves for the main environmental variables influencing the distribution of fin whale. Mean values and standard deviation for 10 runs are presented. 138

Figure 5.53 Mean monthly suitability values for fin whales around São Miguel Island, Azores. 139

Figure 5.54 Monthly standard deviation of suitability predictions for fin whales around São Miguel Island, Azores. 140

Figure 5.55 Mean monthly suitability values for fin whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 141

Figure 5.56 Monthly standard deviation of suitability predictions for fin whales around the central group islands (Pico, Faial, Sao Jorge and Terceira), Azores. 142

LIST OF TABLES:

Table 3.1 Static Environmental variables (modified from Fernandez et al. 2017). 33

Table 3.2 Dynamic environmental variables (modified from Fernandez et al. 2017). 36

Table 3.3 Sightings and group size features during the 2013 survey 39

Table 3.4 Sightings and group size features during the 2014 survey 40

Table 3.5 Detailed description of the fields on the observations dataset. 44

Table 3.6 Species selected for further analysis, with the number of total sightings for the different combinations of spatial resolutions. The last column represents the number of presence grids available after filtering for no-data pixels when using Chlorophyll as co-variate. 47

Table 4.1 Formulas used to build the suitability values for each virtual species according to the environmental variables. Dcoast: distance to 200m bathymetric line; SST: Sea Surface Temperature; Fdist: distance to frontal major areas. Table extracted from Fernandez et al. 2017. 58

Table 4.2. Number of sampled cells with suitability over and below the selected threshold for presences (H=0.6). Table extracted from Fernandez et al. 2017. 59

Table 5.1 Islands sampled each year and number of companies providing data for each location. 87

Table 5.2 Deep-diving species: permutation importance (P) and jacknife training gain (J) indexes of the variables selected for model predictions. 89

Table 5.3 Remaining delphinid species (striped, Atlantic spotted, common and bottlenose dolphins): permutation importance (P) and jacknife training gain (J) indexes of the variables selected for model predictions. 89

Table 5.4 Baleen whales (blue, fin and sei whales): permutation importance (P) and jacknife training gain (J) indexes of the variables selected for model predictions. 90

Table 5.5 Test AUC values (± standard deviation) obtained when testing predictive capacity of models at 2km spatial resolution with no chlorophyll variables. AUC was obtained using a spatio-temporal masked cross-validation approach and an independent dataset. 91

Table 5.6 Test AUC values (± standard deviation) obtained when testing predictive capacity of models at 4km spatial resolution with chlorophyll variables included. AUC was obtained using a spatio-temporal masked cross-validation approach. 92

ACKNOWLEDGEMENTS

Marc Fernandez was supported by grant M3.1.2/F/028/2011 from the Fundo Regional para a Ciência e Tecnologia (Azores Government). This research was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT - Foundation for Science and Technology, under the project "PEst-C/MAR/LA0015/2013", by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT - Foundation for Science and Technology and European Regional Development Fund (ERDF), in the framework of the program PT2020 and by cE3c funding (Ref: UID/BIA/00329/2013). It was also partly supported by CIRN (Centro de Investigação de Recursos Naturais, University of the Azores), and CIIMAR (Interdisciplinary Centre of Marine and Environmental Research, Porto, Portugal).

The data used in this thesis was obtained through the MONICET platform from Azorean whale and dolphin watching companies. I am very thankful to all these companies and the guides, skippers and lookouts that joined and helped us during this process. The setup of the MONICET platform was funded by grant ref. M5.2.2/I/005/2008 from the Azores Government. The continuous operation of MONICET, however, would not be possible without the support of the contributing companies and the dedication of a large group of volunteers and interns. The companies involved at some stage in the project (and their contact persons) were: Picos de Aventura (Pedro Miguel and Carla Coutinho), Terra Azul (Miguel Cravinho), Horta Cetaceos (Pedro Filipe), Ocean Emotion (Paulo Fernandes and Soraia Martins), SeaColors Expeditions (Jasmine Zereba and Paulo), Dream Wave Algarve (Raul Domingos Correia and Sara Galego), Futurismo (Ruben Rodrigues). From the long list of volunteers, photographers, contributors and interns who collaborated in the project: Adam Parker, Afonso Prestes, Anne-Maria Naapuri, Arianna Cecceti, Arne Kiss, Bruno Sampaio, Clara Sardà, Dick Hoyer, Filipe Lourenço, Harkema, Inge Rekveld, João Manuel Brum, João Quaresma, José Azevedo, Maggie L. Gamble, Maria Cruz, Marie Guilpin, Mario Nelson, Marta Aparici, Martin Kurtsson, Miranda van der Linde, Monserrat Cassas, Mr Jensen, Nélio Saldanha, Paulo Pacheco, Pedro Madruga, Rafaela Moniz, Raquel Soley, Ricardo Cordeiro, Rickard Melkersson, Rui Rodrigues, Tim Farland Took, Wolfgang Hantel, Kris Fuerstenau, Milou van Mulken, Tiago Batista, Cesar Medeiros, Karin Hartman, Peter Paul van Maasakkers, Rui Santos, Adomas Lukas, Aurelie Golin, Jorge Canet, Marc Perrussel, Ricardo Fernandes, Rebecca Walker, Sofia Mendoça, Stephanie Almeida, Fernando Coelho, Susana Simião, Maggie O'Connor, Laura Roig, Nuno Costa, Alexander Zwincknagl, Patrick Venturi, Natalia Peña, Manuel Fernandes, Elisabet Badosa, Maria Jaen,

Bea Olveira, Axel Martinez, Elena Tiebas, Gloria Marin, Denis Janampa, Lola Renard, Caludia Melville, Laura Tojeiro, Ramon Soto, Jennifer Coulon, Maria Salvador, Fatima Gonzalez, Jordi Galofré, Marine Attard, Leia Navarro, Marina Gardoki, Marilia Orilio, Rita Ferreira. I made every effort to have a comprehensive list of the volunteers, and apologize for any omission.

The sampling data collection for the present thesis was only possible through the collaboration of the Groupe de Recherché sur les Cétacés (GREC), who provided the infrastructure needed. I would like to thank also the effort of all the volunteers who participated in the survey detection campaigns. All the sampling campaigns were done under permission from the local authorities: licenses 30/2013/DRA and SAI/DRA/2014/1260. I thank Cláudia Faustino for the shapefiles to simulate the sampling transects used in this thesis.

I would like to thank to all the contributors to this work, however due to my poor memory probably I will forget many of them. Nevertheless I will try to acknowledge as many of them as I can remember. This is not only my work, but also the result of their help and support.

Thanks to my parents (Jesus and Pilar), thanks for the giving me the opportunity of a precious human life. Thanks for taking care of me for so many years without any other goal than my happiness.Thanks to all my family who always supported me and encouraged me to follow my dreams.

Thanks to José Azevedo who trusted me, a young crazy “espanhol” who came to the Azores 11 years ago with crazy ideas and trying to change few things. Thanks for being patient and trust on my ideas, even if sometimes may seem quite strange. Thanks to Ana Neto for all the trust and support provided throughout all these years.

Thanks to Arianna Cecchetti, my Phd colleague and friend, who had to stop me in difficult moments and listen my crazy theories many times. Thanks to Karin Hartman for many silly moments and conversations. Thanks to some of these talks I managed to really understand what I am doing.

Thanks to Alexandre and Odile Gannier, for trusting me and providing me with the opportuninty of learning many things about the cetacean world. Thanks to the GREC to trust the Azorean project, allowing the use of the sailing boat Anacaona in the Azores for sampling cetaceans’ distributions. Thanks to Chris Yesson who trusted me and helped me in many analyses, R codes, discussions, English corrections and much more.

Thanks to the Azorean whale watching companies who trusted me, and listened. Especially I would like to thank Miguel Cravinho (Terra Azul) for many insightful and inspirational conversations.

Thanks to all my friends who had been patient to listen many boring conversations about cetaceans and distributions.

Thanks to all my colleagues of the University: Afonso, Artur, Emanuel, Eunice, Eva, Gustavo, Isadora, João, Nuno, Pedro, Ruben and many others.

Thanks to all the volunteers, Eurodysees, Erasmus and others who helped us during these years, without them MONICET wouldn't be alive nowadays. Thanks to all the people who I found during this life, who, somehow, gave me the opportunity to learn and grow. I am very grateful to all of them, without any exception. May all of them find happiness and success on this life.

RESUMO

O uso de modelos de distribuição é hoje em dia uma prática muito comum e bem estabelecida, com grande aplicação nos campos de gestão e conservação da natureza. Existem, no entanto, muitas preocupações sobre a validade dos resultados obtidos com estes procedimentos. Este facto é ainda mais preocupante quando os trabalhos são realizados no meio marinho. Este meio pode apresentar um grande dinamismo, com mudanças muito súbitas das condições ambientais, o que pode influenciar a distribuição das espécies duma forma muito rápida. Nesta tese de doutoramento são apresentados estudos destinados a melhorar a construção de modelos de distribuição para cetáceos em habitats oceânicos. Num primeiro capitulo é feita uma introdução teórica aos conceitos da modelização do nicho ecológico, conjuntamente com um estado da arte. Uma breve caracterização oceanográfica da área onde foram feitos os estudos (Arquipélago dos Açores, Portugal) é apresentada no capitulo 2. Seguidamente é feita uma descrição detalhada dos dados que foram usados para a elaboração desta tese (capitulo 3), sendo classificados em dados de ocorrências e dados ambientais. Duas tipologias de dados de ocorrências foram usadas: (1) dados obtidos com uma amostragem sistemática a traves de transectos dedicados e (2) dados recolhidos duma forma oportunista nas atividades comerciais de observação de cetáceos, a traves da plataforma MONICET. As metodologias de recolha e armazenamento de dados são descritas neste mesmo capitulo. Duas tipologias de variáveis ambientais foram usadas: (1) variáveis que não apresentam grande variabilidade temporal ou estáticas (relacionadas com características batimétricas) e (2) variáveis que apresentam uma grande variabilidade temporal ou dinâmicas (relacionadas com variáveis oceanográficas). No capitulo 4 é apresentado um estudo teórico para testar qual é a melhor resolução temporal das variáveis ambientais e dos dados a ser usada na construção de modelos de nicho ecológico no meio marinho. Foram construídas espécies virtuais baseadas na bibliografia existente o que permitiu testar os efeitos de usar diferentes resoluções temporais (diária, semanal e mensal) das variáveis ambientais. Dois algoritmos de modelização foram usados:(1) “Generalized Linear Models” e (2) “Generalized Boosted Models”. Os resultados mostraram que o uso de médias de 8 dias das variáveis ambientais oceanográficas produz os melhores resultados para as espécies cuja distribuição é condicionada por elas. Os dados armazenados na plataforma MONICET foram usados no capitulo 5 para testar métodos de modelização do habitat. Um total de 7 anos de dados de avistamentos para a ilhas de São Miguel, Pico, Faial e Terceira (Açores, Portugal) foram usados para construir os modelos, usando para esta finalidade o algoritmo MAXENT. Foi proposto um novo método para corrigir o enviesamento associado a estes dados, baseado num índice de detetabilidade e numa área mínima amostrada.

No mesmo estudo foram testadas diferentes resoluções espaciais e temporais. Os resultados obtidos demonstram que para a maioria das espécies testadas (8 de 10) é possível obter modelos de distribuição bons ou mesmo excelentes. A correção do enviesamento aplicada provou ser útil, melhorando a capacidade preditiva dos modelos em todos os casos. A resolução temporal demonstrou ser importante, obtendo diferentes efeitos dependendo da espécie estudada. Desta forma, para as espécies altamente dinâmicas tais como os golfinhos oceânicos, o uso de médias de 8 dias produziram melhores resultados, enquanto que para as espécies mais dependentes de variáveis estáticas, tais como os mergulhadores profundos, a diferença entre os cenários temporais foi mínima. Os resultados desta tese serão úteis para melhorar os modelos de distribuição para cetáceos, assim como para outras espécies marinhas com grande mobilidade.

ABSTRACT

The use of species distribution models (SDMs) is a well-established practice for management and conservation procedures. However there are still many concerns regarding the applicability of the results obtained following these procedures. Specifically, marine ecosystems can present a high dynamism rate, with very fast changes of environmental conditions which correspondingly might influence the distribution of many species. This thesis presents studies analysing and testing hypothesis related with SDMs for marine cetaceans in oceanic habitats. In the first chapter a general introduction of the ecological niche modelling field is presented with a review of current methods and new directions for cetacean niche modelling. Following the introduction, a short characterization of the study area (Azores Archipelago, Portugal) is presented in chapter 2. Chapter 3 presents the environmental and occurrences data used in the subsequent chapters of the thesis. Two kinds of occurrences data were used: (1) systematic data collected on transects and (2) opportunistic data collected from commercial whale watching activities, stored in the MONICET platform. Data collection methodologies and storage are presented in this chapter. The environmental data used in this thesis can also be divided into two main categories: (1) 'static' variables with little or no temporal variability (generally related with bathymetric characteristics) and (2) 'dynamic' variables with high temporal variability (related with oceanographic processes). In the chapter 4 a theoretical analysis is used to test the potential effects of different temporal resolutions for selected environmental variables. Virtual species were constructed based on existing knowledge for several cetacean species, which allowed a test of the effects of grouping environmental variables into different temporal grains (daily, weekly and monthly). Two modelling algorithms were used: (1) Generalized Linear Models and (2) Generalized Boosted Models. Results showed that an 8-day resolution for the oceanographic variables produced the best results for species whose distribution is most affected by them. Data stored at MONICET was used in chapter 5, to test different niche modelling techniques. Using seven years of sightings in São Miguel, Pico, Faial and Terceira islands (Azores, Portugal) niche models were constructed using the MAXENT algorithm. A sampling bias correction for presence-background modelling procedures was applied, based on a bespoke detectability index. Different spatial and temporal resolutions were also tested. Results exceeded expectations, as the models showed good or excellent predictive capabilities for 8 of the 10 cetacean species studied. The sampling bias correction method improved the predictive capability in all cases. The importance of temporal grain size selection varied, depending on the species studied. Highly dynamic species (such as small oceanic delphinids) produced the best results with 8-days grain size, while for more "static" species (such as deep-diving species) the

difference between using a monthly and an 8-day temporal resolution was small. The findings of the present thesis will improve the development of species distribution models for cetaceans and other highly mobile marine species.

CHAPTER 1

INTRODUCTION

1 INTRODUCTION

Knowledge of geographic distributions is important for conservation efforts in many fields. In the absence of comprehensive distribution data, estimates obtained from ecological models can be useful alternatives. Habitat models have been used for several applications such as identifying key habitats and areas of concern for vulnerable populations (Guisan et al. 1999), management of anthropogenic treats (Redfern et al. 2013) and evaluating climate change effects (Keith et al. 2014). In recent times this field has experienced enormous growth (Peterson et al 2011), in part driven by easy access to biodiversity records through opportunistic datasets and citizen science programs.

This introduction starts with a general description of niche modelling procedures, including a description the niche theory and its different applications; the second part introduces and analyses the specific problems of the marine environment, with a special focus on cetaceans.

1.1 SPECIES DISTRIBUTION AND NICHE MODELING: THEORETICAL FRAMEWORK

WHAT IS THE NICHE?

The term "niche" is used in several fields and corresponds to different concepts. Hutchinson's (1957) defined ecological niche as a space in a hypervolume of environmental variables such that "every point (…) corresponds to a state of the environment which would permit the species to exist indefinitely". Peterson et al. (2011) have stated that most of the differences between authors on the definition of niche rely on 3 main topics: (1) the meaning of "exist indefinitely", (2) what kinds of variables constitutes the hypervolume, and (3) the nature of the feedback loop between a species and the variables composing the hypervolume.

From a population ecology point of view, a species might exist for a period of time in a place *g* of the world if its total instantaneous growth rate in that place is on average non-negative (Vandermeer 1972, Maguire 1973, Hutchinson 1978). However the effects of the environment on the growth rates can vary due to different factors (Holt 2009), such as: (1) the Allee effect (reduced per-capita growth rate at small population densities), (2) environment modification in areas of high population growth (Chase & Leibold 2003), and (3) stochastic variations in growth rate (Lande et al. 2003).

Regarding the variables constituting the niche hypervolume it is essential to differentiate between the variables not affected by the presence of the species and those that might be changed by the species. Peterson et al. (2011) based on Hutchinson (1978) define these two kinds of variables as scenopoetic (those which are unlinked and can be used for the construction of multivariate environmental space in which there are different niches, constituted by simple subsets) and nonscenopoetic (linked to the population levels of the species in question). Generally scenopoetic variables are abiotic, related to things like climate and geomorphology, while nonscenopoetic variables are related to biotic factors, such as nutrients.

Understanding the impacts of the species on environmental variables can improve niche definitions, however in some cases it might be extremely complicated. Chase & Leboid (2003) defined the niche using the knowledge of the full instant growth rate in a space that is constantly modified by the species of interest. While this is a very fruitful approach, obtaining such knowledge requires a huge effort and may be almost impossible to gather in most cases. Therefore Peterson et al. (2011) recommend the definition of an environmental space in terms only of non-interactive variables, which might be a plausible hypothesis at relatively coarse spatial resolutions.

GRILLENIAN VS ELTONIAN NICHE

The previous discussion provides a background to understand the meaning of the term "niche". The main meaning is explicitly geographic, and is based on an environmental space (E-space) composed of scenopoetic variables. These niches have been called "Grillenian niches" (James et al. 1984). Other potential term to be used to describe a niche is the term "Eltonian niche", which is oriented to community–ecology questions and is defined at local scales. Ideally, to fully understand the geographic distribution of a species both niches are required.

Hutchinson (1957) defined two sub-types of environmental niches: the fundamental and the realized niche. The first was defined as "all the states of the environments conditions which would permit the species **S** to exist". The "realized niche" can be defined as the subset of the fundamental niche corresponding to environmental conditions under which a species **S** is a superior competitor and can persist (Hutchinson 1957). Therefore it is possible to say that the fundamental niche is an expression of the physiology and behavior of an organism, and its definition can be obtained independently from the localities where a species is observed (Peterson et al. 2011).

The fundamental niche may include combinations of environmental variables currently missing in the existing E-space (Jackson & Overpeck 2000), which is constricted by geography and changes continuously across evolutionary time periods (Manning et al. 2009). Therefore, it is possible to

define a "potential niche" (Jackson & Overpeck 2000) as an intersection of the existing E-space (or existing environmental space) with the fundamental niche – or the portion of the fundamental niche that exists somewhere in the region at the time of the analysis (Peterson et al. 2011).

By defining a geographic space (G-space) and intersecting it with digital data layers of environmental variables it is possible to extract subsets of the existing E-space that correspond to different regions of the G-space (Peterson et al. 2011). These can be also expressed as the Grilleninan niche dimension and it will be the basis for all the chapters of the present work.

NICHES AND GEOGRAPHIC DISTRIBUTIONS

The distribution of a given species can be defined as the set of all grid elements in which, within a given sampling time period, the probability of recording an individual of such species exceeds some given threshold (Peterson et al. 2011).

There are several expressions and ways to explain the relation between distributions and niches. I use here the BAM diagram (Fig.1.1) as depicted by Soberón & Peterson (2005) to describe some of the results of the interactions that determine a species distribution. Set **A** represents regions in **G** where scenopoetic variables are favourable for the species, allowing a positive growth rate. The **B** area represents the geographic area where the biotic factors (e.g. interactions with other species) are favourable for the existence of the species. Finally **M** corresponds to the accessible area for the species in a given period of time. Two further areas from this graph need to be defined, $\mathbf{G_0}$ and $\mathbf{G_I}$. $\mathbf{G_0}$ is defined as the distributional occupied area, and it is the sub-set of accessible geographic areas which are suitable for the species. It can be linked to the "realized niche" concept defined by Gaston (2003). On the other hand, $\mathbf{G_I}$ is the invadable distributional area, or the area that the species would occupy if distributional barriers or movement limitations were to be overcome. The union of occupied and invadable areas can be defined as the potential distributional area (Gaston 2003) for the species.

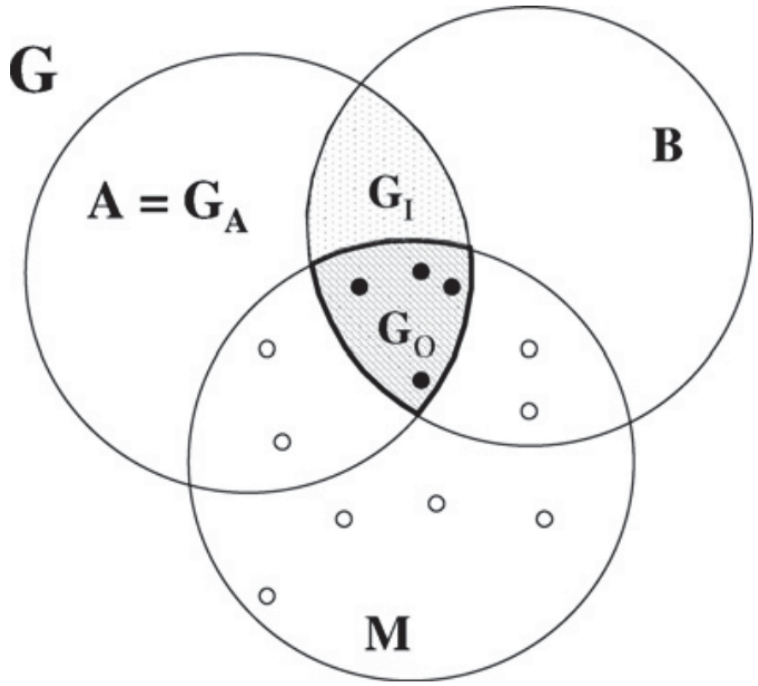


Figure 1.1 The BAM diagram (Soberon & Peterson 2005), depicting the interaction between abiotic (A), biotic (B) and movement (M) factors. G represents the entire geographic area under consideration, G_A the abiotically suitable area, G_0 the occupied distributional area, and G_I the invadable distributional area. Black circles indicate presences, white circles indicate absences.

While source populations can only occur within the $\mathbf{G_0}$ area, sink populations may occur outside $\mathbf{G_0}$ but within the area of movement (Peterson et al. 2011). Following the same criterion, it is possible to define $\mathbf{G_A}$ as the abiotically suitable area, representing the set of grids with the favourable scenopoetic variables for a species. Peterson et al. 2011 also defines this term as the "existing fundamental niche" representing the subset of **E** (environmental space) that is within the tolerance limits of the species (fundamental niche). The authors define the Grillenian fundamental niche ($\mathbf{N_F}$) as the "scenopoetic fundamental niche". While the fundamental niche concept is a physiological characteristic of the species, the existing fundamental nice ($\mathbf{E_A}$) can be defined as the existing subset of **E** that is within the tolerance limits of a species (Peterson et al. 2011). The $\mathbf{E_A}$ may be estimated using non-physiological techniques and even correlative methods such as "ecological niche modelling" (ENM).

Several techniques of ENM allow the use of a limited number of presences in a restricted geographical region, and corresponding to the observed E-space, to a larger one (e.g. Brotons et al. 2004, Guisan & Thuiller 2005, Soberon & Peterson 2005, Araújo & Guisan 2006, Elith et al. 2006, Peterson 2006) representing different aspects of the niche. If unbiased absence data is available, correlative models can be used to estimate the

probability a species being present, conditioned to the environmental features (Guisan & Zimmermann 2000, Keating & Cherry 2004, Phillips et al. 2009). This would provide an estimate of $\mathbf{G_0}$. However, if data is not truly unbiased, correlative methods won't be able to estimate $\mathbf{G_0}$ (Ward et al. 2009). With presence-only or presence-background methods, one can only estimate the sets of environmental variables similar to the localities with presence data (Soberón 2010). Therefore, without dispersal limitations or biotic information, results will estimate some area between the actual area of distribution and $\mathbf{G_A}$ (Jiménez-Valverde et al. 2008).

1.2 NICHE MODELLING

SPECIES OCCURRENCE DATA

The concept of occurrence data might seem simple; however it is essential to obtain good distribution estimates. One can define an occurrence as an observation, or detection, of an individual (or group) on a specific site of the geographic space. There is a series of factors that need to be considered, which according to Peterson et al. (2011) can be divided in two groups: biological factors (e.g. mobility, abiotic suitability and biotic suitability) and factors related to exploration, detection and data (e.g. species must have been identified correctly). For a presence to occur all the factors should be fulfilled simultaneously.

Even when species are able to access specific areas there are other factors to be taken into consideration. If an area is accessible to the species and has the right abiotic and biotic conditions, the occurrence data should reflect a presence. However, as depicted in Fig. 1.2, this is not always the case, as due to detection problems or biases we can obtain an absence. Therefore, the probability distribution of presence records depends on the combined effects of the five components. Occurrence data inevitably contains information that may mask the true objects of interest, so results must be interpreted carefully (Peterson et al. 2011).

A first division of occurrence data can be made between primary (directly obtained from the species observations) and secondary (processed from primary data). Primary occurrence data often provides a distorted representation of true distributional patterns, even when large amounts of data is available (Peterson et al. 2011). One of the main reasons for this is detectability; some species might be more easily detected than others. Another source of misrepresentation of species distributions is the bias regarding the distribution of the sampling effort. Data collected might be biased towards some specific locations (such as roads or cities). Therefore, it is essential to look at the different potential biases associated with data

collected (either in geographic space or in environmental space) to be able to apply the necessary corrections.

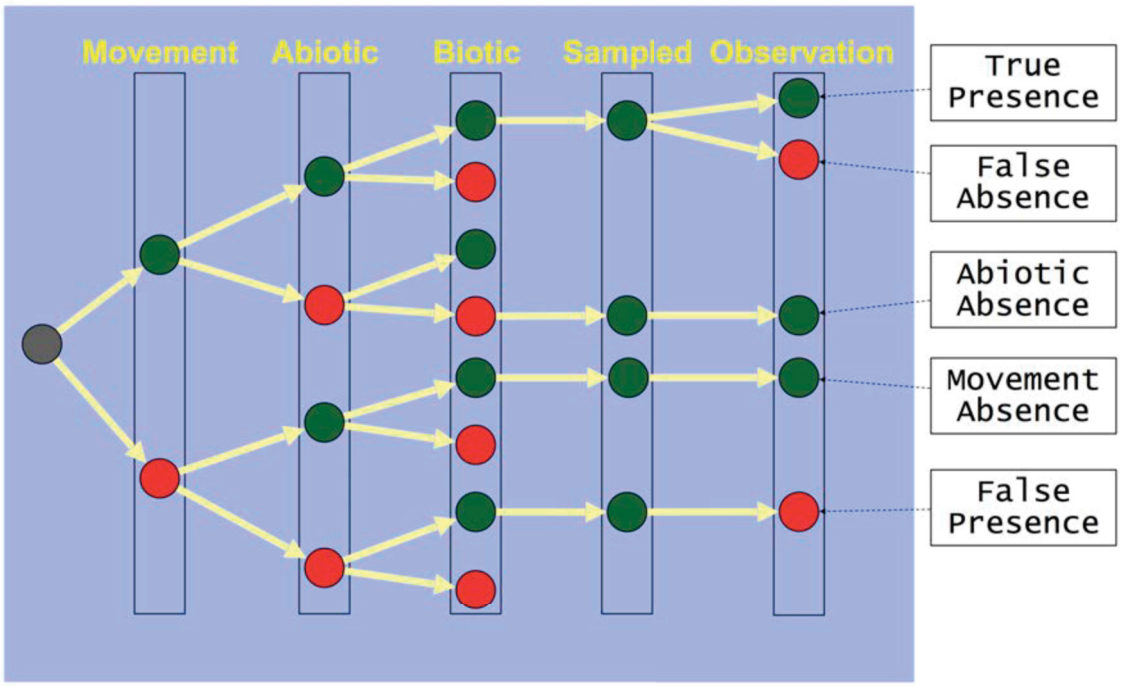


Figure 1.2 Probabilistic events leading to a presence or absence, modified from Peterson et al. (2011). Each bar represents a choice, with red filled circles representing "no" and green representing "yes". The first column represents the movement area, the Abiotic and Biotic columns represent the areas suitability for both cases, the Sampled columns represents if the area is sampled or not. Finally the Observation column reflects whether an observation is correctly performed or not.

Absence data, or something similar, is required by several modelling techniques that use some sort of absence, pseudo-absence or background data for model calibration (Hirzel et al. 2002, Elith et al. 2006). According to Peterson et al. (2011) there are 3 main approaches to assemble occurrence data: (1) presence-only data; (2) presence-absence data; and (3) presence data and a sample of the background or "pseudoabsence" data. Background sampling is used to characterize the environmental conditions present across the study region, while pseudoabsence sampling entails selecting from areas or sites where the species has not been detected. The good quality of the absence data is essential to produce good models: as shown in Fig. 2 it is possible to obtain false absences, which will cause problems in the modelling process. Ideally presences should be sampled from $\mathbf{G_0}$ whereas absences would be selected from areas outside $\mathbf{G_P}$ or potential distributional area, and within **M** and **B**. Absence data, if to be used at all, must be considered carefully with respect to the probability tree and to the BAM diagram (Peterson et al. 2011).

ENVIROMENTAL DATA

In order to estimate the niche, species occurrences must be complemented with environmental data. Besides the classification of environmental variables provided by Hutchinson (1978) into scenopoetic and binomic, the classifications of Austin (2002) are also useful:

1. *Idealized variables* – Based on the level of direct physiological effects on the species distributions.
    a. *Indirect* – No direct effects on the species, but can affect species distribution due to correlation with other variables (e.g. latitude, elevation).
    b. *Direct* – Direct effects on the species, but not consumed by them (e.g. temperature)
    c. *Resource* – Consumed by the organisms (e.g. food resources)
2. *Distal/Proximal variables* – Based on the degree of causality of a species responses to environmental factors
    a. *Proximal* – Organisms respond directly to variables
    b. *Distal* – Indirect responses due to additional causal links to the variables

It is possible to say that different scales of geographic distributions will be more influenced by different kinds of variables. When working at fine spatial resolutions a combination of scenopoetic and binomic variables will be required (Soberón 2007) while at coarser scales (regional to global) the signature of scenopoetic variables is expected to be dominant (Pearson & Dawson 2003).

Careful should be taken when using indirect variables (such as elevation as a proxy of temperature); while this assumption might work in a certain latitudinal range it won't be true across global scales (Austin 2002). Temperature at high altitudes in tropical areas might be equal to the temperatures at low altitudes in northern areas. Therefore, elevation won't be a valid environmental variable as a temperature proxy across a global region.

When selecting the variables to take into consideration, two main approaches can be used: using a few preselected and relatively uncorrelated variables that correspond directly to known physiological rules (e.g. Huntley et al. 1995, Baselga & Araujo 2009); or use large datasets, even if highly correlated variables are present (e.g. Stockwell 2006). According to Peterson et al. 2011 both extremes have drawbacks. When using few variables there is the risk of underestimating the niche, producing overly broad potential distributional areas (Barry & Elith 2006). But the use of too many variables can seriously overfit the models, particularly when few occurrence records are used (Peterson & Nakazawa, 2008).

The resolution of the environmental variables refers to the size of the subdivisions (in time or in space) that are applied to the datasets under consideration (Peterson et al. 2011). The spatial resolution of the analysis should match the spatial resolution of the biological phenomenon studied. Thus when studying distribution of terrestrial mammals a spatial resolution of 1-2 squared km might be logical choice; conversely this might be not useful when studying soil nematodes (Peterson et al. 2011).

The temporality of the occurrences and environmental data must also coincide (Peterson et al. 2011), although additional factors must be taken into consideration. For instance, the environment dynamism must be evaluated to select the adequate temporal resolution to be used. Thus, when studying the distribution of small mammals in tropical climates a seasonal scale might be enough, however when looking at temperate areas, where the variability is higher, it could be more useful to use a monthly scale.

MODELLING TECHNIQUES

According to Peterson et al. (2011) modelling the ecological niche can be defined as the task of characterizing every cell within a region in terms of quantitative values related to the probability of occurrence of a given species, as a function of the environmental conditions presented in that cell.

According to their objective, models can be classified as predictive or exploratory. Predictive models use methods that aim to produce optimal predictions in geography, yet provide little in the way of interpretable environmental information regarding the specific qualities of the niche being modelled. On the other hand an explanatory model will focus on characterizing the niche in understandable terms, even though it may not necessarily produce accurate predictions (Peterson et al. 2011). This classification is not exclusive: a single model can have at the same time good explanatory and predictive powers.

In order to produce an estimate of the distribution it is necessary to apply a modelling algorithm. This part can be considered as the core of the modelling procedure, however it should be noted that this is only a small part of the broader modelling process: factors, such as including selection the reference regions **G** and **M**, choice of environmental variables, characteristics of the occurrence data, model calibration and others are also key elements in the process (Peterson et al. 2011).

Here I will present a classification of the models available, according to Peterson et al. (2011), based on the data input:

1. *Presence-only:* Use only presences, without references to other samples or information of the study area (e.g. envelope models such as BIOCLIM, HABITAT, DOMAIN).

2. *Presence/absence:* Works contrasting sites where the species has been detected with sites where the species has been documented as absent (e.g. GLM, GAM, CART, BRT, ANN).
3. *Presence/background:* Assesses how the environment where the species is known to occur relates to the environment across the entire study area (e.g. ENFA, Maxent, Poisson point process, or modified GLM and GAM).
4. *Presence/pseudoabsence:* Aims to compare known occurrence localities with a set of localities having a bellow unity probability of constituting presence localities. Pseudoabsence data is sampled from sites where the species is not known to occur. In general, the same methods used with presence/absence data can be used here.

Recent studies showed that even if using a specific kind of data there is no golden rule to select the best model algorithm to be used. Several algorithms should be tested (Qiao et al. 2015). However, identifying the "best" model is not an easy task, since the method used to evaluate the model quality should depend on the aim of the modelling (Peterson 2006).

When calibrating and evaluating models two pools of occurrence data should be available: data for calibrating the models and for evaluating model predictions. One important point is the origin of the dataset used for testing purposes. According to Peterson et al. (2011) several issues are important when selecting this data: (1) degree of independence between calibration and evaluation data in space and time, (2) how best to divide occurrence records into calibration and evaluation datasets, and (3) whether records are divided in respect to spatial position. Ideally, calibration and evaluation data may be drawn from both different areas and time periods, but such cases are rare (Nogues-Bravo et al. 2008, Peterson et al. 2009). Theoretically, to allow a real model validation evaluation data should be fully independent from calibration data (Araújo et al. 2005). As technically it can be quite costly to generate this kind of datasets, most evaluation efforts are based on the partition of a single sample into a calibration and a testing data datasets. Different techniques (such as different kinds of k-fold cross-validation methods and bootstraps) are used for this purpose.

Model evaluation can be divided into two different approaches: evaluation of performance and tests of significance (Peterson et al. 2011). The performance measures basically address how well the model achieves a particular goal. These metrics are generally based on omission and commission error (related with the false negatives/positives concept). The omission error represents the number of grid cells of known occurrence of the species predicted as absent by the model. The commission error indicates the proportion of localities of known or assumed absence for the species which are predicted as presences by the model. Some of these measures indicate

the ability of the model to rank presences or absences correctly, while others assess a model's goodness of fit to observed presences and absences. In contrast, tests of significance assess whether model predictions of records in the evaluation dataset are better than random regarding the prediction and evaluation data (Peterson et al. 2011). Some of these metrics are based in a modified receiver operating characteristic (ROC) curve, in which the axes are the lack of omission error in the y-axis versus commission error on the x-axis. The area under the curve (AUC) of the ROC curve represents an overall measure of model performance. The AUC ranges between 0-1, an AUC value of 0.5 corresponds with a random prediction. Care should be taken when using AUC values with presence-background models, as the index interpretation might not be straightforward. For example AUCs are not comparable among species (Phillips et al. 2006) or between regions (Lobo et al. 2008). Some methods relying on a partial AUC, emphasizing subsectors of the ROC proved to be an efficient tool to overcome the concerns raised (Peterson et al. 2008).

1.3 NICHE MODELING AND DISTRIBUTION STUDIES FOR CETACEANS

THE MARINE ENVIRONMENT SPECIFICITIES

Marine ecosystems are dynamic and fluid, which leads to an important habitat variability. While some changes may have a decadal (e.g. el Niño) or annual scale (e.g. general currents patterns), others might happen on a weekly, daily (e.g. frontal areas, Fig. 1.3), or even an hourly basis (e.g. tidal currents). However, there is no simple rule to describe habitat dynamism. For example, while some frontal areas can present seasonal patterns others can be short-lived, lasting only a few days (Belkin et al. 2009).

The strong temporal variability of oceanic features is also reflected on the spatial scale, where they can range from several meters to thousands of kilometres. These complex dynamics, with very fast changes of environmental conditions, can influence the distribution of many species. Different biogeographic provinces also present different dynamic conditions. Factors such as latitudinal gradients should be taken into consideration: tropical areas generally present more stable conditions than temperate ones. Other aspects may also be important, for example large pelagic areas tend to be more stable than coastal areas (Mann & Lazier, 2013).

Oceanic islands represent a particular modelling challenge, due to their strong impact on water circulation. For example, on the Azores archipelago (used as study area for this thesis) the oceanographic system is influenced by eastwards flows, the North Atlantic current (45-45˚N) and the warm Azores current (34-36˚N). Sangrà et al. (2009) identified two westward propagating

corridors of warm water eddies north and south of the Azores current. Previously Bashmachnikov et al. (2004) had suggested that some of these eddies could be trapped in specific areas due to the complex topography of the region. Moreover Sala et al. (2015) described the importance of the archipelago as an area for retaining incoming particles. An extended description of the Azores archipelago is given in chapter 2.

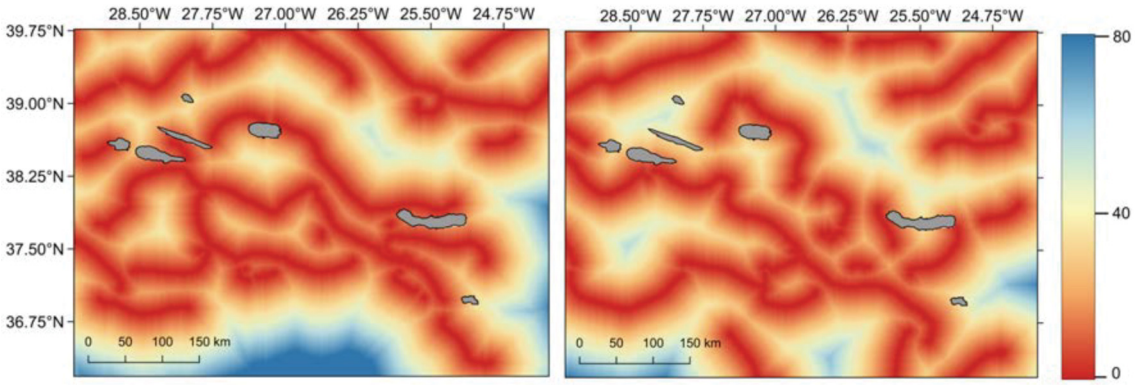


Figure 1.3 Distance to frontal thermal areas in the Azores archipelago area in two consecutive weeks (26/6/15 to 4/7/15 and 4/7/15 to 11/7/15). Distances are expressed in kilometers. Fronts were derived from MUR SST daily products.

DATA SOURCES FOR CETACEAN DISTRIBUTIONAL STUDIES

Ideally distribution data should result from a temporally consistent effort across the localities, biomes or countries they cover (Margules et al. 2002). However this is not often the case, especially when working in the marine environment where there are many biases associated even to designed surveys.

Two kinds of datasets can be defined according to their nature: opportunistic and non-opportunistic (or survey data). Non-opportunistic or survey datasets are those collected by a directed effort, following surveys designed to address specific questions. Opportunistic data are collected as a by-product of some other activity. In the case of cetaceans it can come from many sources, such as observers on fishing, ferry or cargo vessels, whale and dolphin watching companies, and other sea users.

Initially almost all the distributional studies were based on pre-designed transects or a in a directed effort. Ship cruises or plane flights were designed specially to detect cetaceans and to estimate their distribution and abundances (Hedley et al. 1999, Gomez de Segura et al. 2008, Forney et al. 2012, Manocci et al. 2014). Data obtained through platforms of opportunity were generally seen as low-cost alternatives, providing limited useful information to understand factors affecting distribution and abundance (Evans

& Hammond 2004). Recently, cetacean habitat-modelling data collected from platforms of opportunity has been considered to be almost equivalent to data collected using designed surveys if all the potential sources of bias are taken into consideration (Redfern et al. 2006). When looking at other research and conservation fields, the use of opportunistic data has been of great utility, especially the so-called citizen science data (Strein et al. 2013). In the cetacean field some studies used opportunistic data to understand habitat preferences (Isojunno et al. 2012), distributions (Rossi-Santos et al. 2006, Kiszka et al. 2007, Moura et al. 2012) or animal movements (Whitehead 2001), nevertheless the majority of the studies still rely on dedicated survey data.

MODELLING CETACEANS DISTRIBUTIONS

Since the early days of cetacean research efforts have been made to better understand cetacean distributions. Early studies, such as those of Hui (1979) or Au & Perryman (1985) used correlational analysis to find the relationship between delphinid observations and the environmental conditions registered for the study areas. In the following years several studies tried to identify cetacean habitats by plotting sighting locations from surveys and using a diversity of statistical methods to look for significant differences (Baumgartner et al. 2000, Jaquet & Whitehead 1996, Ballance & Pitman 1998, Kingsley & Reeves 1998, Waring et al. 1999). Others, such as Hooker et al. (1999), selected areas with high frequency of sightings. However, all these methods relied on the identification of habitat relations within the surveyed area, and were therefore unable to depict the species niche or its distribution.

The first cetacean habitat prediction models (Fiedler & Reilly 1994, Moses & Fin 1997, Hamakazi 2002, Hedley et al. 1999, Gregr & Trites 2001, Waring et al. 2001) attempted to overcome some of the limitations above. Cetacean habitat models allow predicting potential distributions (or abundances) for unsampled areas from estimates of the existing fundamental niche. With the increasing accessibility and improvements in software GIS software and data analysing packages, the number of studies focusing on cetacean habitat and distribution experienced an important growth. While some studies focused on habitat modeling of a few species (Booth et al. 2013, Cañadas & Hammond 2008, Esteban et al. 2014, Griffin & Griffin 2003, Gomez de Segura et al. 2008, Tepsich et al. 2014, Thorne et al. 2012) others looked at several species in a wide variety of habitats (Azzellino et al. 2008, Ballance et al. 2006, Correia et al. 2105, Kanaji et al. 2016, Kaschner et al. 2006, MacLeod et al. 2007, Skov et al. 2008, Manocci et al. 2014). Many different techniques have also been used, such as CCA (Griffin & Griffin 2003), binary logistic regressions (e.g. Azzellino et al. 2008), GLMs (Gordon et al. 2000, Gomez de Segura et al. 2008), GAMs (Hedley et al. 1999, Correia et al. 2015),

CART analysis (MacLeod et al. 2007), ENFA (e.g. Skov et al. 2008, Praca & Gannier 2007), MAXENT (Moura et al. 2012, Thorne et al. 2012) and other statistical learning methodologies (Torres et al. 2013, Kaschner et al. 2006).

Different sets of environmental variables were used on the construction of distributional and density estimates. In some cases in situ oceanographical measurements were used (e.g. Baumgartner et al. 2000, Griffin & Griffin 2003). However, the use of remotely sensed oceanographic data produced similar performances or outperformed, in some cases, the models constructed using in-situ data (Becker et al. 2010). The use of these data sources added more flexibility into the models, and allowed to test different effects related with spatial and temporal scales.

While on the terrestrial environment several studies have targeted the influence of different scales temporal and geographical scales on model accuracy (e.g. Pearson & Dawson 2003, Guisan & Thuiller 2005), the issue has only recently started to be addressed in the marine environment. Redfern et al. (2006) and Ballance et al. (2006) discussed the potential effects of different scales when building habitat models for cetaceans. Initial correlational studies (such as Jaquet & Whitehead (1996) for sperm whales) found a dependence of the distribution of some species on the spatial scale. Redfern et al. (2008) tested the effects of using different spatial scales for modelling dolphin distributions in the Eastern tropical Pacific, finding little or no effect of spatial scale variability.

Other authors started to include not only spatial but also the temporal resolution. In order to better understand the logic of these approaches it is important to have a basic understanding of the preferred environment and ecology of the target species. Marine cetaceans have a wide distribution, being present in a variety of ecosystems. As seen above, marine ecosystems can present very fast changes of environmental conditions, which correspondingly might influence the distribution of oceanic species. Some areas might present more stable conditions (like tropical areas) or different species might relate differently with their environment. For example a generalist species will adapt easily to different environmental conditions, due to its plasticity, with other factors (such as social structure or competition) potentially influencing their distribution.

Few studies have so far tried to understand how different habitat dynamics might influence model output. Becker et al. (2010) tested on the California current system environmental layers with different temporal resolutions, finding that in general the 8-day means provided the best estimates. Manocci et al. (2014), in contrast, found that climatic means of environmental conditions in the Madagascar area provided the best estimates. Recently Scales et al. (2017) found that climatological or seasonally based models can introduce strong biases when working with animal movement

predictions. While it seems that the temporal and spatial resolution might play a very important role on the modelling procedures of highly mobile marine species such as cetaceans, there is no real consensus on the best practices to follow. Consequently, there is still much work to be done in this scientific domain.

1.4 RESEARCH OBJECTIVES

The general objective of this PhD is to advance the understanding of the factors to take into consideration when modelling cetacean distributions. In order to achieve this goal I followed a logical structure, starting with a theoretical approach to end with a practical application. The specific goals were defined as:

1. Contribute on a theoretical level to the knowledge of the interplay between spatial and temporal scales in models of highly mobile species in a dynamic environment.
2. Provide a practical example of the selection of temporal scales when modelling cetacean distributions.
    a. Analyse the potential utility of an opportunistic dataset with high-temporal resolution for niche modelling.
    b. Describe the dynamic niche of cetacean species.

1.5 THESIS OUTLINE

This thesis is divided in 5 main chapters, together with an introduction. Chapters are organized in order to follow the research objectives and represent a logical path to the final results.

CHAPTER 2: THE AZORES: A COMPLEX GEOGRAPHICAL AND OCEANOGRAPHIC SYSTEM

This chapter presents a short description of the oceanographic characteristics of the study area. It is intended to provide a background to the complexity and dynamism of the area.

CHAPTER 3: ENVIROMENTAL AND OCCURRENCES DATASETS

This chapter introduces all the datasets used in chapters 4 and 5. A detailed description of the different sources of environmental and occurrences data is given. Information about the data origin, processing and mapping is given. On the occurrences data special attention is given to the MONICET dataset, as it is considered an essential part of the present work. This is an open access platform which collects, processes, and gives access to

distribution and photo-identification information from commercial whale watching operations.

CHAPTER 4: A THEORETICAL APPROACH TO MODELLING MOBILE SPECIES IN DYNAMIC ENVIRONMENTS

This chapter presents a theoretical analysis of how the selection of different temporal resolutions of environmental data might influence modelling performance. Three virtual species were created whose distribution was influenced by: (1) only static (topographic) variables; (2) only dynamic (oceanographic) variables and; (3) both dynamic and static variables. Three different temporal resolutions (or grain sizes) were tested for grouping the environmental variables: daily, weekly and monthly.

Results showed that the choice of temporal resolution is an important factor in modelling the distribution of mobile species when that distribution is influenced by dynamic variables.

CHAPTER 5: USING A HIGH TEMPORAL RESOLUTION OPPORTUNISTIC DATASET TO MODEL THE ECOLOGICAL NICHE OF CETACEANS

The knowledge obtained on the previous chapter is applied to models build from the MONICET dataset. A sampling bias correction method was used, combining spatial thinning together with a background selection function based on a detectability index. Different spatial resolutions were tested. A set of strict validation techniques were applied.

Results exceeded expectations, as the models developed for 8 of the 10 species showed good or excellent predictive capabilities. The sampling bias correction method improved the predictive capability in all cases. The importance of temporal grain size selection varied, depending on the species studied. Models for species primarily dependent from dynamic oceanic features produced best results with 8-days grain size; for species mainly dependent from static variables the difference between using a monthly and an 8-day temporal resolution was very small.

CHAPTER 6: DISCUSSION

This chapter summarizes all the work presented in this thesis, analysing the importance of each chapter and creating a general picture of the findings. Additionally it also presents recommendations for future work.