2019

Species Distribution Model of Cetaceans in Relation to Environmental Factors at Two Locations in the North Atlantic

Adams, L.

Adams, L. (2019) 'Species Distribution Model of Cetaceans in Relation to Environmental Factors at Two Locations in the North Atlantic', The Plymouth Student Scientist, 12(1), p. 3-24.
http://hdl.handle.net/10026.1/14680

Species Distribution Model of Cetaceans in Relation to Environmental Factors at Two Locations in the North Atlantic

Laura Adams

Project Advisor: Jill Schwarz, School of Biological and Marine Sciences, Plymouth University, Drake Circus, Plymouth, PL4 8AA

Abstract

Cetaceans in the North Atlantic are under threat from the increasing pressure and demand placed on the sea by humans. In order to conserve cetacean populations, legislative acts such as the EC Habitats Directive (Council Directive 92/43/EEC) are calling for protected areas to be established. Species distribution models have been used in this study as a tool to outline habitat preference of cetaceans in the Celtic Sea and eastern North Atlantic, helping to determine the boundaries for protected areas. Remotely-sensed environmental data was analysed against visual boat survey data using boosted regression tree modelling. A review of the literature emphasised the need for site specific analysis due to the variability in physical oceanographic processes. This was reflected in the results with the most important variable being chlorophyll at Baltimore and water depth at Penzance. The optimal depth was calculated at 30 – 50 m for both sites. The optimal range of chlorophyll was between 1.8 – 2.5 mgm^{-3} with an increase in cetacean sightings towards the upper limit. A preferred sea surface temperature of 14.6 – 15 °C was found at the Baltimore site and 15.5 – 15.6 °C was found at the Penzance site. The need for fine scale analysis of oceanographic properties and their relation to species distribution is recognised, as well as an expansion of this study to capture the full range of environmental variability at these locations.

Introduction

The North Atlantic is inhabited by twenty-five species of cetaceans (Reid et al., 2003); including the humpback whale *Megaptera novaeangliae*, fin whale *Balaenoptera physalus*, minke whale *Balaenoptera acutorostrata*, common dolphin *Delphinus delphis*, harbour porpoise *Phocoena phocoena*, common bottlenose dolphin *Tursiops truncatus* and Risso's dolphin *Grampus griseus*. Individuals of each species are threatened by anthropogenic activities. These threats have led to diminishing population numbers resulting in legislation being put in place to conserve cetaceans. There are many frameworks and directives aimed at protecting cetaceans in the North Atlantic including the EC Habitats Directive (Council Directive 92/43/EEC) which lists both the harbour porpoise and bottlenose dolphin on Annex II of the directive obliging member states to establish Special Areas of Conservation for these species (Berrow et al., 2010). Species distribution models that outline habitat preference can aid in the planning for potential protected area locations (Hoyt, 2011).

The aim of this project was to provide site specific habitat preference for the cetacean species observed by charitable conservation trusts. Species occurrence data from visual boat surveys and environmental data from publicly accessible online resources were used to generate boosted regression tree models, from which habitat preferences were deduced.

Threats

Cetaceans in the North Atlantic are under threat. Fin whales are currently considered to be endangered and harbour porpoises and humpback whales are listed as vulnerable by the International Union for the Conservation of Nature (National Parks and Wildlife Service, 2009). Anthropogenic activites such as historical whaling, entanglement in fishing gear, ship strikes, vessel harassment, pollution and habitat degradation are contributing to the decline of cetacean populations (Clapham et al., 1999). The impact from fisheries seems to be the predominant detrimental activity for cetaceans, mainly due to the considerable number of bycaught cetaceans. Off the south-west coast of Ireland five different cetacean species were bycaught by Dutch mid-water trawlers (Couperus, 1995). Blanchard et al. (2005) also found that the impact of fisheries in the Celtic sea has influenced the fish community structure: Results showed that between 1987 – 2003 there had been a decrease in the abundance of larger fish and an increase in the abundance of smaller fish, this can in turn have a knock-on effect for the cetaceans feeding on these fish (Blanchard et al., 2005).

Rising sea surface temperature could impact cetaceans negatively because increasing water temperature coincides with a northward shift in the distribution of fish (Beare et al., 2004) as well as a shift in the ranges and community composition of phyto- and zooplankton (Barton et al., 2016). This shift affects pilchard and herring spawning, two species that are preyed upon by harbour porpoise and common dolphin (Rogan and Berrow, 1996). Therefore, increasing sea surface temperature can lead to the displacement of cetacean habitat.

Disturbance can be caused to cetaceans through whale watching which has become an important economic contributor in Ireland and the United Kingdom, with direct revenues worth millions (Hoyt, 2000). Any vessels not adhering to a strict code of

conduct issued by the Maritime Safety Directorate in 2005 have the potential to disturb cetacean species, causing them stress (Berrow and Holmes, 1999).

Legislation

Numerous international and regional legal frameworks, agreements and treaties have been established with the intent of providing legal footings for conserving and protecting cetaceans and their habitat. At an international level all of the species observed in this study are protected by the Convention on the Conservation of Migratory Species of Wild Animals (CMS; Bonn Convention; 1979; implemented 1983) which currently has 124 member parties (Evans et al., 2003). The convention arranges transboundary protection for cetaceans that frequently cross over national boundaries on their migratory routes, providing a legal framework for conserving cetacean species and aiming to restore or maintain a favourable conservation status. Cetacean protection is split regionally, encompassed by the Agreement on the Conservation of Small Cetaceans of the Baltic and North Seas (ASCOBANS), which includes the study area, and the Agreement on the Conservation of Cetaceans of the Black Sea, Mediterrenean Sea and Contiguous Atlantic Area (ACCOBAMS) (Prideaux, 2003).

At a regional scale the Convention on the Conservation of European Wildlife and Natural Habitats (Bern Convention; 1979; implemented 1982) aims to protect the endangered cetaceans listed in Appendix II and III and their habitats. Appendix II protects species such as the common dolphin, harbour porpoise and humpback whale strictly, whereas Appendix III allows for exploitation of other cetacean species as long as the population is maintained (Reid et al., 2003). The 1992 EC Habitats Directive (Council Directive 92/43/EEC) was implemented in order to adhere to the obligations of the Bern Convention, offering protection to all listed species within the 200 nautical mile Exclusive Economic Zone of member states (Hammond et al., 2013). Under this directive Natura 2000 was created, which refers to an ecologically coherent network of protected areas that conserve breeding and resting sites for endangered species. If the United Kingdom leaves the European Union in 2019 it will remove itself from its obligations under the EC Habitats Directive. However, the habitats and species that inhabit British waters will still contribute to European biodiversity. This means they will continue to be managed but under new legalities allowing scope for better designation; this is an important consideration because our understanding of populations and distributions has improved with the use of species distribution models (Kirkham & Shepherd, 2016) since the current Special Areas of Conservation were finalised.

Species Distribution Models

Species distribution models can be created to highlight locations for designating protected areas as they outline regions of preferred habitat. They are a numerical tool used in many fields to model the geographic distribution of species by correlating the recorded presence of a species and the environmental variables at the site of occurrence (Gomes et al., 2018). Species distribution models can be used to provide ecological understanding of a species and predict their distribution within an environment. The accuracy of each model can be determined by its sensitivity, which is the model's ability to correctly predict the presence of a species (true positive rate), and its specificity, which is the model's ability to correctly predict the absence of a species (true negative rate) (Lalkhen & McCluskey, 2008).

High quality species occurrence data and environmental variable data - which is non-biased, of high resolution and a large enough data set, is required to build a model. This can be obtained either by collecting data in the field, which ensures quality but can be time consuming, or through secondary resources such as GIOVANNI, an online database of ocean colour data, and the Global Biodiversity Information Facility (GBIF), an online database of species occurrence. However, these easily available sources can present problems such as coarse resolution and gaps in data (Tyberghein et al., 2011).

Models can be built using presence-only data, referred to as a naïve model, or using presence-absence data. For common species, bias is created in the presence-only model as the background data includes true presences as well as true absences (Ward et al., 2009). To combat this, pseudo-absences can be generated. Using a random approach to generate psuedo-absences has shown good discrimination power, and the capacity to exclude contributing variables form the input predictor data set (Cerasoli et al., 2017).

Boosted Regression Trees

A range of different modelling techniques exist. A comparison by Reiss et al. (2011) analysed the robustness of nine different models. Maxtent v3.3.3a software and R modelling were used. Evaluating the models was done by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC). The ROC represents the relationship between sensitivity and the proportion of false positives (1 – specificity). The AUC value calculated is between 0 and 1, where 0.9 and above shows excellent prediction, 0.7 to 0.9 shows good prediction, 0.5 to 0.7 shows poor prediction and 0.5 and below shows a prediction no better than random (Hosmer and Lemmeshow, 2000). The boosted regression tree had one of the highest AUCs of 0.840 in Reiss et al.'s (2011) analysis, showing best predictive power and accuracy. Other models tended to over predict species distribution, whereas the boosted regression tree recreated distribution patterns more accurately. The advantages of boosted regression trees were summarised by Friedman & Meulman (2003) to be that they accommodate multiple types of predictor variables (numeric, categorical, binary, independent and non-independent). They also accommodate for missing values, ignore extreme outliers and inclusion of irrelevant predictors and they fit interactions between predictors allowing for 'robust' analysis of multiple variables. Absence data can be ignored and the weighted mean of fitted values in relation to each non-factor predictor is given (Friedman, 2001). 'Boosting' allows trees to be progressively added to the model whilst previous trees data is reweighted to emphasise poorly predicted cases. The robustness of the model comes from the ability to fit a large number of simple trees together.

Cetaceans in the Area

The Celtic Sea bordering Baltimore and the islands of west Cork is an area of ideal habitat for cetaceans with deep waters, offshore banks and a continental shelf that runs parallel with Ireland's South West coast (O'Brien et al., 2009). Charif et al. (2001) recognised the West of Ireland seaboard as an important migratory corridor for large baleen whales, including fin and humpback. The harbour porpoise and common dolphin are the two most commonly sighted cetaceans in Irish waters with their abundance being centred around the South West coast (Reid et al., 2003). The

coastal waters of County Cork including Roaringwater Bay, designated as an SAC under Natura 2000, have shown to be valuable habitat for the harbour porpoise (Evans et al., 2003) due to the shallowness of the water in which they predominantly feed on dermersal fish (National Parks and Wildlife Service, 2009).

Nearly all UK records of dolphin sightings have been recorded in the Western channel waters surrounding Cornwall, demostrating habitat importance (McClellan et al., 2014). Leeney et al. (2012) indicated that cetaceans occur frequently in coastal waters south of the Cornish peninsula in close proximity to the edge of the coastal shelf. It was also concluded that minke whales regularly utilize both the offshore and coastal waters around Cornwall.

Physical Oceanography of the Area

The physical oceanography of the south west of Ireland has been reviewed by National Parks and Wildlife Service (2009). The area is largely sheltered from strong winds and currents from the North Atlantic. Dominated by shallow continental shelf waters that extend up to 200 km from the coast, most of it shallower than 100 m, making it preferable habitat for harbour porpoises. The seafloor drops off at the edge of the continental shelf sloping down into a deep oceanic basin. Along this slope are high levels of productivity generated by cold, nutrient rich waters being upwelled by the gulf stream. This upwelling triggers a bloom of phytoplankton resulting in the congregation of high densities of feeding fish such as herring, which are the typical diet of common dolphins (O'Brien et al., 2009) This process could be linked to the frequent sightings of common dolphins over the shelf-edge. The International Council for the Exploration of the Sea (2008) describes the Irish Shelf front that runs along the south west of Ireland and occurs all year round as a boundary between offshore North Atlantic waters and the tidally mixed shelf waters.

Evidence for coastal upwelling along the south west of Ireland is poorly researched. However, Raine et al. (1990) recorded temperature and salinity in the summer months in Bantry Bay, an inlet just north of the survey area. Highly variable measurements, in both space and time, were observed and this is suggestive of periodic upwelling. A study of waters neighbouring the survey area concluded upwelling of cool dense shelf waters was present in July and by early August downwelling occurred causing the shelf waters to recede, being replaced by warmer coastal water. Smaller scale events lasting a few days were also observed that were caused by the passage of Atlantic depressions (Edwards et al., 1996). These and other upwelling events often trigger a bloom of diatoms especially in areas which were previously highly stratified. As the water begins to re-stratify, conditions become optimum for dinoflagellate blooms. The blooms occur between April – October, the same months as the survey, and attract aggregations of zooplankton and fish, including species preyed upon by cetaceans.

The physical oceanography of the Western Channel, off Penzance, has been summarised by Pingree (1980). The area is influenced by the pressure system of the North Atlantic Ocean, with a gradual increase in pressure from the North to the South. Sea temperatures are influenced by the North Atlantic Drift and the Gulf Stream current which supplies warmer water. The winds generally blow from the west and are stronger in winter than summer due to the pressure gradient being steeper in winter. The wind speed effects the rate of water flow resulting in greater

net movement of water in winter. McClellan et al., (2014) described the Western Channel being significantly deeper than the Eastern Channel. It slopes from the coastline steeply to its deepest point, an undersea basin known as the Hurd Deep which exceeds 100 m depth. Frontal zones often form around archipelagos, headlands, thermal boundaries and density boundaries. Water has different salinity and oxygen carrying capacities at different temperatures (Nybakken, 2000). When different masses of water with different temperatures meet oceanic fronts are formed. Composite front mapping has shown fronts that are observed in the channel are persistent occuring all year round (Miller and Christodoulou, 2014). These front aggregate organisms such as phytoplankton trapping them in the surface layers (Franks, 1992) providing food for some cetaceans and their prey species.

Cetaceans and Environmental Variables

The presence of cetaceans can be linked to environmental factors, with specifics varying between species and site locations. Chlorophyll concentration has been found to be the most important variable for the distribution of common dolphins because of its direct relation to aggregations of food (Moura et al., 2012). Position in the spring-neap cycle and tidal current have also been found to be significant, for example harbour porpoise detection rate was found to decrease with increasing tidal speed in the Hebrides. Areas of low tidal speed were also found to have a higher proportion of mud in the sediment (Embling et al., 2010). Depth is also a determining factor with minke whales found to prefer a depth of 20-50 metres in the Moray Firth, Scotland (Robinson et al., 2009). However, the species were shown to inhabit waters beyond 50 metres in Cornwall (Leeney et al., 2012) showing a variation between sites. This difference can be explained by the reasoning that topography alone is not a determining factor. Sediment type showed a strong correlation with minke whale distribution as certain sediment types provide habitats for prey species of feeding minke whales (Robinson et al., 2009).

A seasonal variation was observed by McClellan et al., (2014) when reviewing the presence of six different dolphin species, including bottlenose and Risso's. It was deduced that sea surface temperature and chlorophyll were significant variables in spring, salinity and distance to shore were significant in summer and distance to shelf and bathymetry were significant in autumn. Age may also play a role in determining cetacean distribution; in the Moray Firth, Scotland 60% of minke whales sighted were juveniles suggesting they are utilising this area more frequently than adults (Robinson et al., 2009). Ceteacean behaviour must also be considered when surveying; feeding or socialising animals are much easier to spot and stay in the same area for longer compared to animals that are travelling.These studies demonstrate the variability of the relationship between environmental variables and cetacean sightings, highlighting the need for site specific and season specific surveys and analysis.

Methodology

Study Area

The survey area is made up of two locations; the waters surrounding Baltimore and the islands of west Cork, Ireland (49°48' - 50°6' N, 5°18' - 5°48' W) and Penzance, Cornwall (51°18' - 51°30' N, 9°0' - 9°48' W).

Data Collection

Secondary data, provided by Whale Watch West Cork and Marine Discovery Penzance, has been utilised for this project. Two teams working out of Baltimore and Penzance used wildlife watching vessels as platforms of opportunity for sightings of cetaceans. The vessels would follow randomly chosen straight line transects until the end of the transect section was reached, usually determined by the sighting of marine megafauna as described by de Boer et al. (2018). Each sighting has two plots; the first is the position of the boat when the animal is initially sighted and for the second, once the animal has moved the boat will move to the position the animal was sighted and the position of the animal in relation to the boat recorded. Sea state and boat speed were also recorded for each sighting. The data spans the period from April to October for the year 2015. The survey was effort-based and the boats tracks were also recorded, however for the purpose of this study it will be interpreted as presence-only data due to time constraints.

The environmental data for the sea surface temperature (SST) and chlorophyll a was also downloaded from GIOVANNI using NASA's MODIS Aqua level 3 browser at 4 km resolution for spatial analysis. Seasonal summer data was downloaded for the year 2015.

The Penzance base map was created by Duncan Jones (de Boer et al., 2018), combining General Bathymetric Chart of the Oceans (GEBCO) data at 1 km resolution for the Celtic Shelf with Channel Coastal Observatory data at 1 m resolution for depths up to 30 m. A vector file was produced in ArcGIS (version 10.2.1) using the contour extraction tool to extract contours from the GEBCO. Depth data from a *MV Shearwater II* survey, where spot depths were collected using a depth sounder, was also utilised. The vector file was then converted to a geotiff in GRASS (version 6.4.3). Raster contours were produced using the vector to raster tool and finally the vector was interpolated using the *r.surf.contour* tool to make the depth file. This method is described by de Boer et al. (2018). The Baltimore base map was created by myself in a previous project. This raster was then interpolated using the *r.surf.contour* tool to smooth the areas in between the point depths, producing a base map with depth values at 5 km resolution.

Data Preparation

The cumulative frequency of number of sightings was plotted against sea state using Excel to remove bias and false negatives. Sightings with sea state 4 and above were eliminated from the Baltimore dataset and sightings with sea state 5 and above were eliminated from the Penzance dataset based on the cumulative frequency graphs.

Initially, the level 2 chlorophyll a concentration and SST data was prepared for use in the boosted regression tree using Matlab (version R2016a). Each variable and site was processed individually by looping through the files, arranged in monthly folders, one at a time. A latitude/ longitude grid was created, bounded by 6°W 48°N 3°W 50°N for the Penzance area and 11°W 50°N 5°W 52°N for the Baltimore area. The level 2 chlorophyll a concentration and SST were re-gridded and outputted to the newly defined grid at 4km resolution and saved as a *.mat* file. To time match the environmental variables with the sighting data points the sighting data was read into Matlab using *xlsread*. The gridded array was defined and chlorophyll a concentration, SST, longitude and latitude were read in from the newly created *.mat*

file. The sighting data was looped through and the pixel corresponding to the sighting location identified, so that chlorophyll a concentration and SST values could be assigned to each sighting. At this point the level 2 data were found to be too cloudy, with many missing values, to be used in the study resulting in switching to the level 3 data (monthly composites). A random set of data points was created using the *rand* command in Matlab to allow for presence-absence analysis of the data. This is more robust than a presence-only analysis, referred to as a naïve model, which is often highly biased (Ward et al., 2009). The data points were assigned environmental variables using the same method as for the level 3 data. Depth was extracted for each sighting using ArcMap by overlaying the sightings onto a bathymetry base map (Figure 1) and using the Extract Multi Values to Points tool. The number of species seen was visualised using graduated symbols to give an overview of species density distribution.

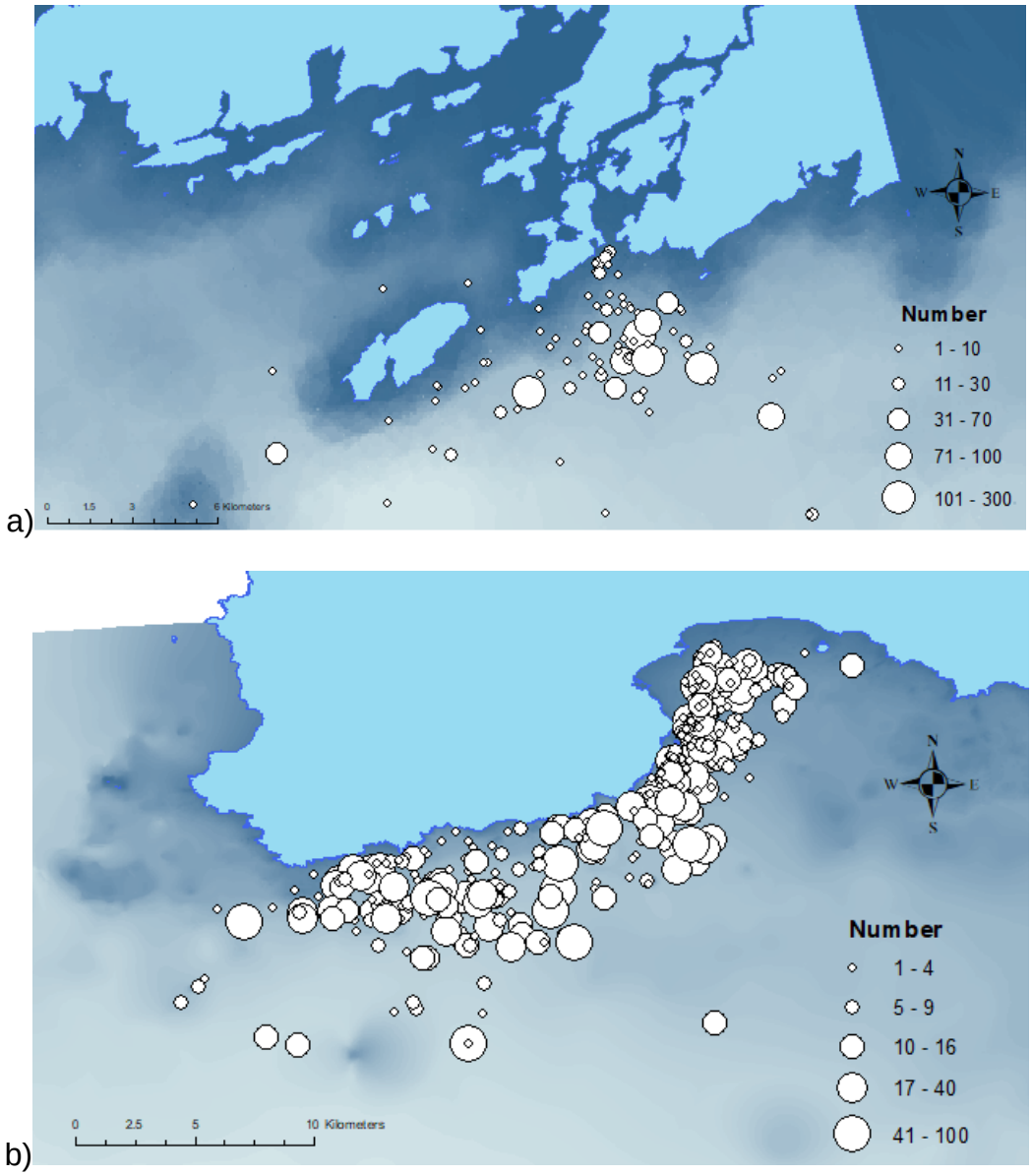

Figure 1: Number of individuals sighted May – October, Baltimore (a) April – October 2015, Penzance (b). Depth is represented by colour with light blue being deeper waters and dark blue shallower areas

The sightings were arranged in a csv file with their assigned chlorophyll a concentration, SST and depth ready to be inputted into the model.

Statistical Analysis

Spearman's Rank correlation coefficient was used for rank correlation on all three environmental variables for both sites to test for correlation amongst the predictor and response datasets ($\alpha = 0.05$).

Boosted regression trees were created in the statistical software R (version 3.4.1) using the *gbm* package with reference to Elith and Leathwick's (2008) tutorial. To initially look at the presence-only data a model was created using *gbm.step* with number of individuals sighted set as the response variable.

The presence-absence data was divided up; one portion was used to train the model and the other witheld portion was used to validate the model. This is known as holdout cross valiadation (Yadav and Sanyam, 2016). Models were generated for Baltimore, Penzance and a combined data set which included the data from both sites. Baltimore's training data had 110 sites and the validation data 70 sites. Penzance's training data had 303 sites and the validation data 174 sites. The combined data set had a training data size of 413 sites and validation data size of 244 sites. Each model was generated with 3 predictor variables. Using *gbm.fixed* to create a model with the training data, different learning rates (0.1 – 0.0001) and tree complexities (1 – 10) were tested. Then *predict.gbm* was used to predict onto the validation data. This calculated the predictive deviance of the different learning rates and tree complexities to allow for selection of optimal parameters. Learning rate shrinks the contribution of each tree in the model (Elith et al., 2008) and tree complexity indicates the interaction between predictor variables. For example a tree complexity of 2 represents a two way interaction (Pour et al., 2016).

Using the identified tree complexity and learning rate models were generated with *gbm.step* using Elith and Leathwick's (2008) online tutorial as a guide. Partial dependency plots of fitted functions, which show the average fitted value of a variable, were created using *gbm.plot*. The fitted value of individual data points was also plotted using *gbm.plot.fits*. Pairwise interactions between the environmental variables was found with *gbm.interactions* displaying the strength and rank of interactions. These interactions were then plotted using *gbm.perspec* to produce a three dimensional graph of the interaction between two variables. The model data was predicted onto the validation data and deviance was calculated with *calc.deviance* and the AUC with *roc*. Finally, predictive deviance of the model was calculated using the same method as described above with *gbm.fixed* and *predict.gbm*. This allowed for representation of robustness of the model.

Robustness of Methodology

A robust method is required for any study to make it insensitive to variation (Arvidsson & Gremyr, 2008). Several steps have been taken to ensure this: The survey method for collecting species observation data was the same for both site locations. Although pseudo-absences were generated in place of data collected in the field, the number of pseudo-absences selected was equal to the number of presences recorded to improve the accuracy of the model (Barbet-Massin et al.,

2012). Pseudo-absences were randomly generately from within the boundaries of the study area to maximise specificity.

Environmental data was obtained for the summer season only (June, July and August) providing data that could be analysed spatially. Although a small portion of the sighting observations occur outside of this these months the application of the model was constrained to the months in which the highest number of sightings was collected, allowing for an accurate representation of variability. In order to prevent overfitting and find the most suitable parameters for the model, holdout cross validation was applied. This simple technique is more reliable than evaluating the data set as a whole, as it avoids overfitting the model, and is the least time consuming. However, it makes inefficient use of the data, limiting the number of instances the model can be trained on. This can result in the model not being trained robustly, giving poor results when the model is used for testing (Yadav and Shukla, 2016), especially in data sets with small sample sizes of n<1000 (De'ath, 2007).

Discussion

Rank Correlation

The strongest relationship was shown between chlorophyll and SST for the environmental variables. It is clear that for both data sets there is only a correlation between SST and chlorophyll (p = 0.0002 and 0.0007, Table 1; Figure 2).

Figure 2: Linear regression between environmental variables at Penzance study area (a,b,c) and Baltimore study area (d,e,f)

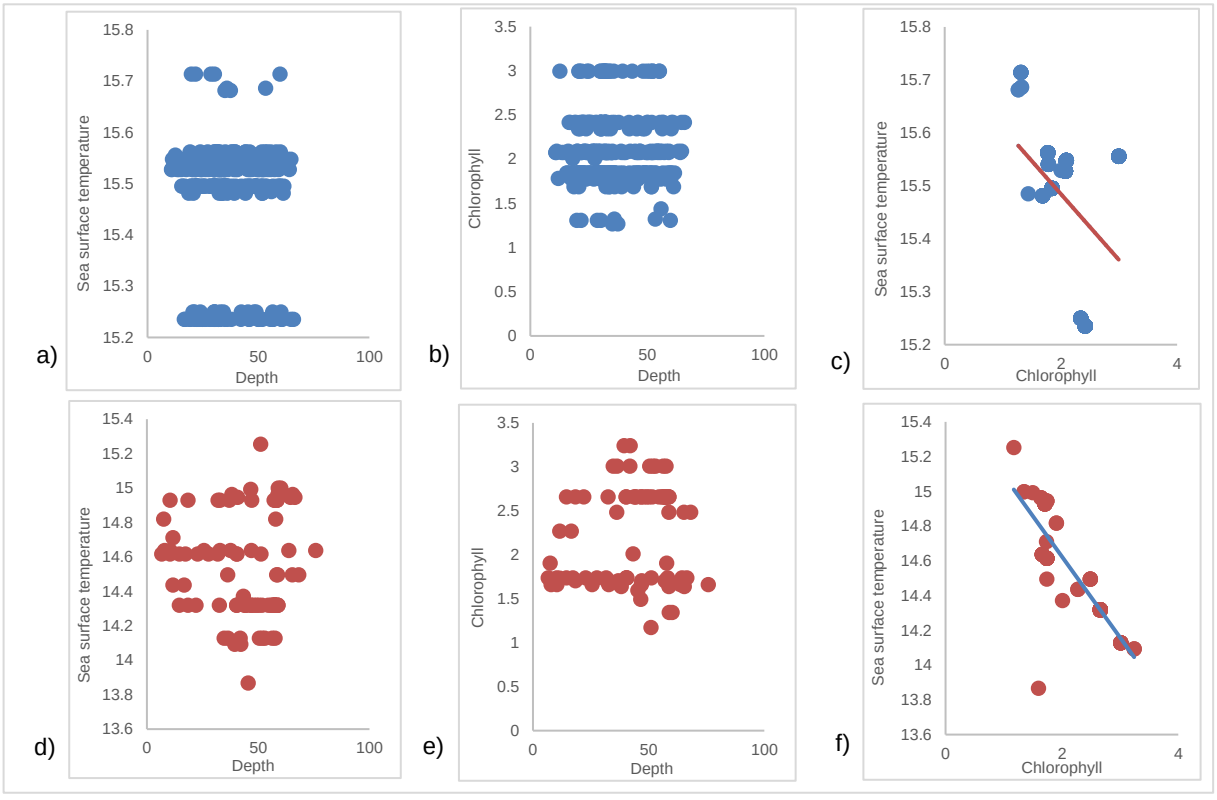

Table 1 Associated regression coefficients (R), p values (P) and number of samples (N) for figure 2

Graph	**a**	**b**	**c**	**d**	**e**	**f**
R value	0.050	-0.042	**-0.207**	0.029	0.017	**-0.898**
P value	0.3683	0.4491	**0.0002**	0.7866	0.8708	**0.0007**
N	320	320	320	90	90	90

The relationship between these two variables is a weak negative correlation with a regression coefficient of -0.207 for Penzance and a strong negative correlation with a regression coefficient of -0.898 for Baltimore.

The negative correlation can be explained by the physical processes happening in these areas over the seasons. In winter, temperatures are low and there are high levels of nutrients in the euphotic zone because of increased mixing from strong winds and low primary productivity levels because of limited sunlight, meaning replenished nutrients are not taken up. In spring, temperatures rise, sunlight hours increase and stratification occurs as the sun warms the upper layers of the ocean. These conditions, combined with the presence of readily available nutrients from the winter, trigger a phytoplankton bloom (Hartman et al., 2014). Over the spring this increase in primary productivity depletes the nutrients in the surface layers, leaving excess nutrients trapped below the newly formed thermocline. In summer, temperatures are still rising and sunlight hours extending, however primary productivity is now limited by the low levels of nutrients (Hydes et al., 2001; Hartman et al., 2014). Upwelling events, driven by wind forcing surface water away from the coast and ocean currents interacting with topographic features, occur frequently in certain areas due to the dynamic yet predictable behaviour of winds and currents. These events bring cold, nutrient rich water from the deeper layers which can support localised primary productivity over the summer. These restricted nutrient levels and the cold temperature associated with upwelling explain the negative correlation between chlorophyll and SST.

A learning rate of 0.005 and tree complexity 2 was selected for Baltimore. Figure 3 c, d show a learning rate of 0.0001 and tree complexity 2 to be optimal for the Penance dataset. However, with the full model creation procedure, these values were found to underfit the model leading to too much bias. The next viable learning rate of 0.001 was chosen instead. Figure 3 e, f show a learning rate of 0.0001 and tree complexity 1 being optimal for the combined dataset. However, when inputted into R these were not viable as the model was underfit and the next viable options of learning rate 0.005 and tree complexity 2 were chosen instead. The Penzance data set and the combined sites data set did not yield to the model, this isshown by the non-conformative plots in figure 3 c,d, e and f which would be expected to have a more similar distribution to a and b where the curve reaches the lowest predictive deviance early on and doesn't soar to a high predictive deviance afterwards but remains at a fairly stable level or gradually increases.

Model Design

Figure 3 shows the predictive deviance results used to select optimal learning rate and tree complexity.

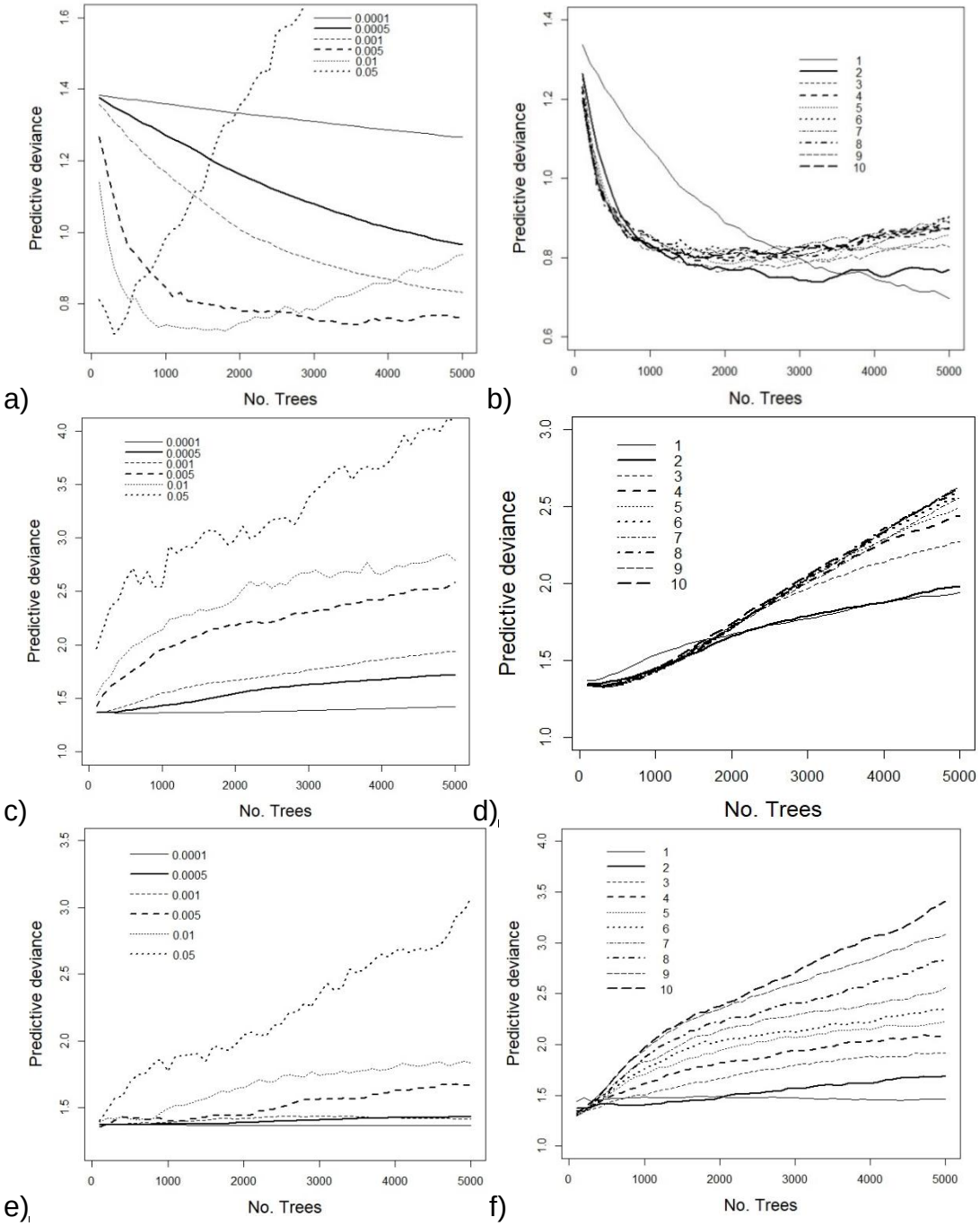

Figure 3: Predictive deviance of different learning rates and tree complexities for Baltimore (a) tree complexity 2 & b) learning rate 0.005), Penzance (c) tree complexity 1 & d) learning rate 0.001) and commbined (e) tree complexity 2 & f) learning rate 0.005)

This may be indicative of the Penzance and combined data sets being too small or having a variance that is not well represented by the data set. Relatively small data sets of n<1000 (De'ath, 2007) increase the predictive deviance as the model may not

be trained correctly, reducing its robustness when used for testing. To combat this in future studies multiple years of sightings data would be utilised.

Fitted Functions

The fitted functions of Baltimore, Penzance and the combined data set are displayed in Figure 4. The logit scale refers to probability, a logit reading of 0 corresponds to a probability of 0.5 and the relationship is symmetrical along the axes (Pampel, 2000). The fitted functions of a model do not represent the relevance of the variables flawlessly, especially if there is a strong correlation between predictors, but they do provide a basis for understanding (Friedman, 2001; Friedman & Meulman, 2003).

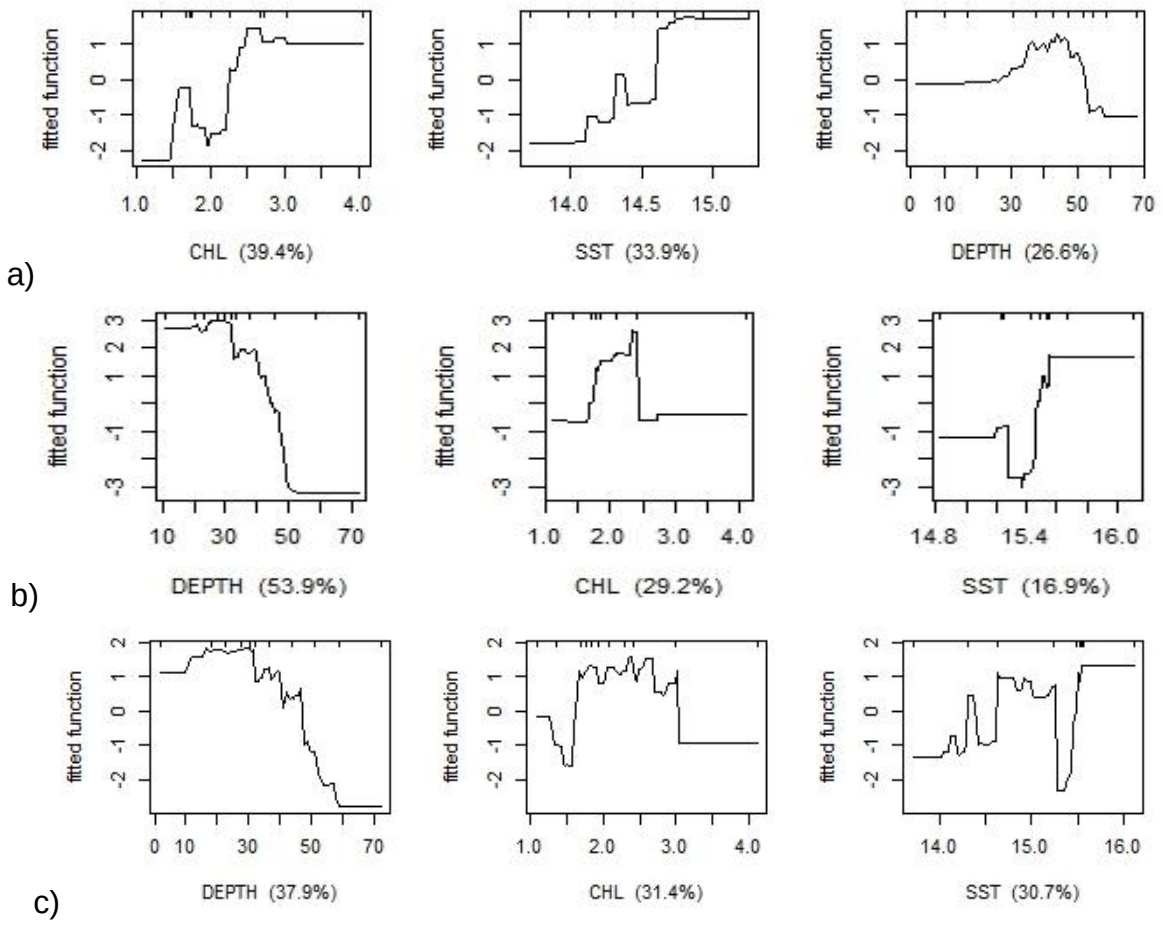

Figure 4: Partial dependency plots of Baltimore (a), Penzance (b) and combined (c) indicating the effect of each variable on cetacean presence/absence. Y-axes are on the logit scale and the contribution of each variable to cetacean presence is shown in brackets

For Baltimore, the variable with the highest significance was chlorophyll a concentration. The range at which chlorophyll a concentration had a positive relationship with cetacean presence was between 2.4 and 3.0 mg/m^3. Below this the probability of sighting a cetacean was low, excluding an anomaly at 1.5 mg/m^3. For Penzance chlorophyll a concentration was the second most significant variable.

There was an increase in the probability of seeing a cetacean between 1.8 and 2.5 mg/m^3. For the combined data set, chlorophyll a concentration was also the second most important variable with cetacean presence more likely to occur between 1.7 and 3.0 mg/m^3. This corresponds to the ranges found at each site indicating a minimum chlorophyll a concentration value of 1.7 mg/m^3 and a maximum of 3.0 mg/m^3. This result suggests that there is a high degree of coupling between phytoplankton and the higher trophic level prey species of the cetaceans in these areas. Cetacean abundance has before been linked to higher chlorophyll a concentrations signifying more productive areas (Smith et al., 1986).

At Baltimore the second most important variable was found to be SST with an increase in cetacean presence seen between 14.6 – 15 °C. At Penzance SST was the third most significant variable. The likelihood of cetacean presence peaked at 15.5 -15.6 °C, indicating a narrow range of preferred temperature. The results of the combined data set concluded two different ranges for two different sites. The difference between the two sites may be explained by the impact SST has on physical and biological processes. Penzance and Baltimore have different bathymetry and currents influencing the formation of oceanic fronts which could lead to the two distinct temperature ranges. The minimum temperature cetacean were found was 14.6 °C and the maximum temperature was 15.6 °C. The limits of these temperatures may be indicative of the fact species of cetaceans are strongly confined to ranges of temperatures (Hind & Gurney, 1997). At an oceanographic level these temperature ranges may be found at the oceanic feautures that are associated with high prey occurrence.

Depth was the least significant variable at Baltimore with cetacean presence within the range of 30 – 50 m depth. Between 30 – 40 m the likelihood of presence increases and between 40 – 50 m depth the likelihood decreases but is still probable. At Penzance depth was the most significant variable; the highest probability of cetacean presence was at 30 m with a gradual decrease in likelihood up until 50 m. This resulted in the same overall pattern as Penzance for the combined data set, which may be because the ratio of Penzance data points to Baltimore data points was 3:1 resulting in Penzance data points being more influential on the combined sites results.

The distribution of prey is directly linked to bathymetry in coastal areas and the chlorophyll a maximium is often found in shallower waters above the thermocline (Agusti and Duarte, 1998). Bathymetry impacts the mixing of water masses and the circulation of nutrients (Kimura et al., 1997); this could explain the significance of depth. The shallow, coastal waters surrounding Baltimore and Penzance can be more abundant in productivity due to the influence of tidal mixing, which replenishes nutrients in the surface layers, justifying the preference for shallower waters (Robinson & Tetley, 2007). The gradual decrease in cetacean sightings and depth at the Penzance site may be indicative of a sloping sea floor. All of the species studied were shallow water species (excluding the Risso's dolphin; Canadas et al., 2002) explaining the relatively shallow depths that were preferred.

Interactions

The interactions between each variable were tested at both sites (Figure 5).

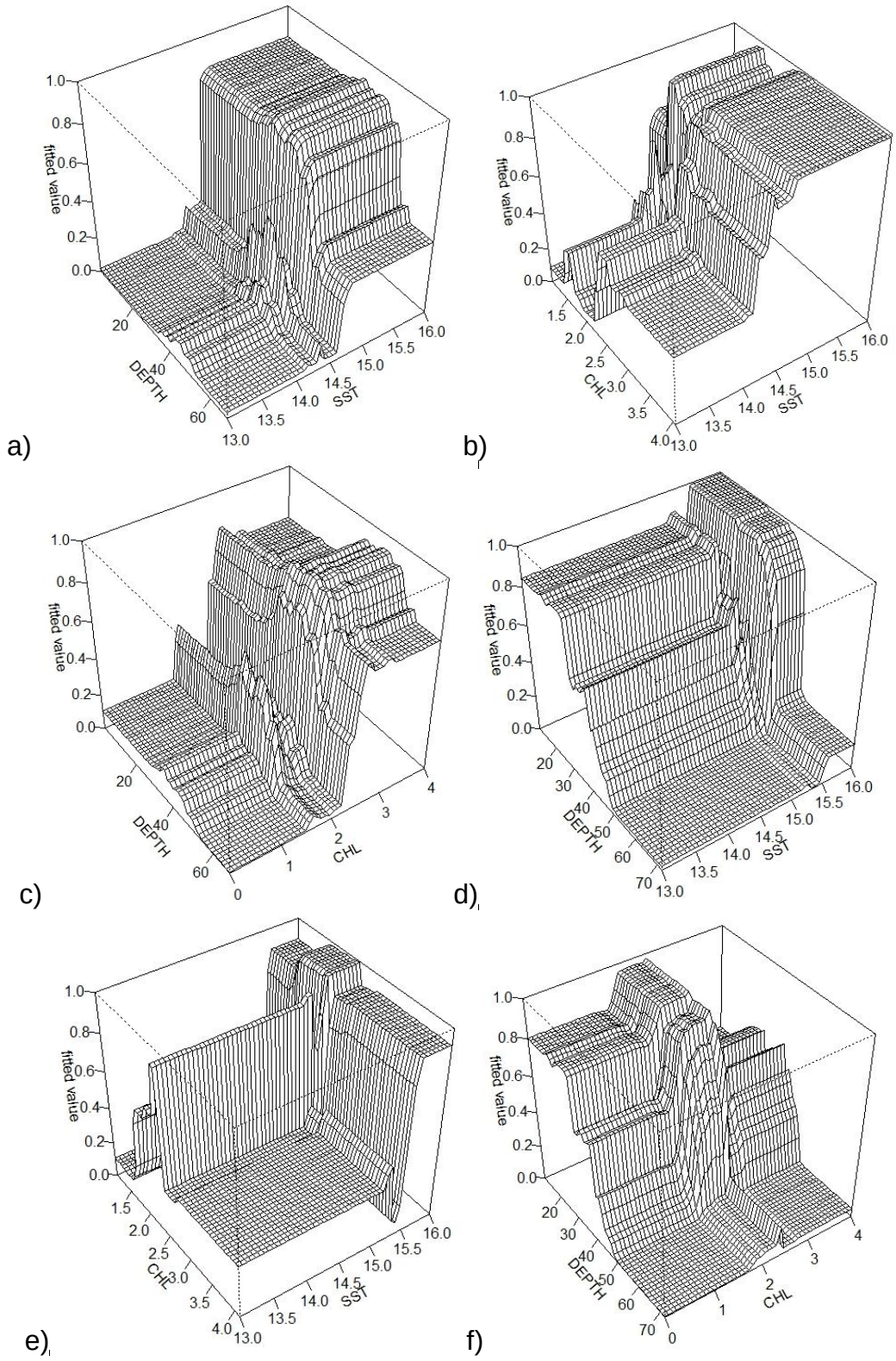

Figure 5. Three dimensional partial dependency plots for the interactions between all variables at Baltimore (a, b & c) and Penzance (d, e & f)

Table 2. Strength of pairwise interactions of predictor variables for Baltimore (a) and Penzance (b)

	CHL	**SST**	**DEPTH**
CHL		107.5	153.9
SST			265.71
DEPTH			

a)

	CHL	**SST**	**DEPTH**
CHL		44.14	519.49
SST			166.64
DEPTH			

b)

The strongest interaction (265.71, Table 2) occurred for depth and SST for the Baltimore site. A depth of 40 m and shallower with a SST of 14.5 °C and above was predicted as cetacean preference (Figure 5a). The strongest interaction at the Penzance site was found to be between depth and chlorophyll a concentration with a (519.49). This score was significantly larger than the other two interactions which may be explained by the minimal variance in SST shown in the partial dependency plots. Depths of 40 m and shallower with chlorophyll a concentrations at 1.8 – 2.5 mg/m^3 were determined to be optimal conditions (Figure 5f).

The model predictive deviance was lowest for Baltimore, at 0.795, compared to 2.31 for Penzance. However, these presence-absence model deviances are both lower than that of the test data which was presence-only. For Baltimore presence-only the deviance was 60+ and for Penzance presence-only it was 9. This is in agreement with the other studies showing that the inclusion of pseudo-absences improved model robustness (Cerasoli et al., 2017).

To improve the robustness of this, study multiple years of sightings data could be collected and analysed to better represent habitat preference within a wider range of environmental predictor values. Data sets spanning numerous years are useful tools to detect decadal cycles of oceanic processes (Koslow & Couture, 2013). Another improvement to be made would be to analyse the sightings of each species separately. It has been shown that different cetacean species require different habitats due to the variation of their morphology, feeding techniques and breeding requirements. Healy et al. (2013) observed that the distribution of baleen whales in Irish waters was aggregated around the spawning grounds of prey species in

comparison to the distribution of common dolphins which was more scattered across the area. Additionally, the environmental data was analaysed at a seasonal scale in this study; this could be improved by using data with a daily resolution that is more representative of conditions at the time of each sighting. Finally, there are higher amounts of dissolved sediment in Baltimore because of the shallown nature of the shelf waters. This can result in remote sensing misidentifying the dissovled substances as chlorophyll a, leading to error (O'Reilly, et al., 1998).

Conclusions

This study concluded that the distribution of cetaceans in relation to temperature, depth and chlorophyll a values at Baltimore was not comparable to the distrubiton of cetaceans in Penzance, emphasising the variation between different locations. It is clear that different oceanographic processes are occuring at each site. In line with the aims, habitat preference of ceteacean was identified: Depths of 50 m and shallower were preferred at both locations but the relationship between depth and cetacean presence was different for each site. This could indicate that areas deeper than 50 m may not be preferable habitat for the cetacean species studied and should therefore not be considered when allocating areas of protection except as they affect other, related environmental drivers and/or response species. Chlorophyll a concentrations of 2.0 – 2.6 mg/m^3 were determined to be prefereable for cetaceans across both sites. Sea surface temperatures of 14.6 - 15 °C for Baltimore and 15.5 – 15.6 °C for Penzance were found to be optimal. These guidelines can be used when determing protected areas both for current conditions and using climate model predictions.

The variable influence of environmental variables on cetacean presence must be considered when assigning protected areas. The designation of these areas has previously been based on fixed variables such as sediment type and depth creating fixed areas of protection. However, for highly mobile species such as cetaceans this is not an adequate form of protection. Habitat shifts of cetaceans can make stationary protected areas redundant. Protected areas must therefore take into consideration the seasonal and temporal variances in species distribution by studying fine-scale parameters that influence distribution. This study outlines the specifity of each site in relation to cetacean presence and how much this varies for the same species across different sites.

Acknowledgements

I would like to acknowledge Duncan Jones, Marine Discovery Penzance and Nic Slocum, Whale Watch West Cork for providing the sightings data. I would also like to acknowledge Duncan for providing the bathymetry base map for Penzance. Finally, I would like to acknowledge Jill Schwarz, Plymouth University for supervision throughout my project and providing guidance for using Matlab and R.

References

Agusti, S., & Duarte, C. (1998). Phytoplankton chlorophyll a distribution and water column stability in the central Atlantic Ocean. *Oceanologica Acta*.

Arvidsson, M., & Gremyr, I. (2008). Principles of robust Design Methodology. *Quality and Reliability engingeering International*, 23-35.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, 327-338.

Barton, A. D., Irwin, A. J., Finkel, Z. V., & Stock, C. A. (2016). Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. *Proceedings of the National Academy of Sciences of the United States of America*, 2964-2969.

Beare, D., Burns, F., Greig, A., Jones, E. G., Peach, K., Kienzle, M., . . . Reid, D. G. (2004). Long-term increases in prevalence of North Sea fisheries having southern biogeographic affinities. *Marine Ecology Progress Series*, 269-278.

Berrow, S., & Holmes, B. (1999). Tour boats and dolphins: a note on quantifying the activities of whale watching boats in the Shannon Estuary, Ireland. *Journal of Cetacean Research and Management*.

Berrow, S., Murray, C., O'Connell, M., Wall, D., & Whooley, P. (2010). Monitoring cetaceans in Irish waters. *Irish Whale and Dolphin Group*.

Blanchard, J., Dulvy, N., Jennings, S., Ellis, J., Pinnegar, J., Tidd, A., & Kell, L. (2005). Do climate and fishing influence size-based indicators of Celtic Sea fish community structure? *ICES Journal of Marine Science*, 405-411.

Blum, A., Kalai, A., & Langford, J. (1999). Beating the Hold-Out: Bounds for K-fold and Progressive Cross-Validation. *COLT '99 Proceedings of the twelfth annual conference on Computational learning theory* (pp. 203-208). Santa Cruz: Association for Computing Machinery.

Canadas, A., Sagarminaga, R., & Garcia-Tiscar, S. (2002). Cetacean distribution related with depth and clope in the Mediterranean waters off southern spain. *Deep Sea Research*, 2053-2073.

Cerasoli, F., Innella, M., D'Alessandro, P., & Biondi, M. (2017). Comparing pseudo-absences generation techniques in Boosted Regression Trees models for conservation purposes: A case study on amphibians in a protected area. *Public Library of Science ONE*.

Charif, R. A., Clapham , P. J., & Clark, C. W. (2001). Acoustic detections of singing humpback whales in deep waters off the British Isles . *Marine Mammal Science*, 751-768.

Clapham, P., Young, S., & Brownell, R. (1999). Baleen whales: conservation issues and the status of the most endangered populations. *Mammal Review*, 37-62.

Clark, J., Dolman, S. J., & Hoyt, E. (2010). *Towards Marine Protected Areas for Cetaceans in Scotland, England and Wales: A Scientific Review Identifying Critical Habitat with Key Recommendations: a Report from the WDCS Scotland and Critical Habitat/MPA Programmes.* Chippenham: Whale and Dolphin Conservation Society.

Couperus, A. (1995). Interactions between Dutch mid-water trawler and Atlantic white-sided dolphins (Lagenorhynchus acutus) southwest of Ireland. *Journal of Northwest Atlantic Fishery Science*, 209-218.

Dalla Rosa, L., Ford, J. K., & Trites, A. W. (2012). Distribution and relative abundance of humpback whales in relation to environmental variables in coastal British Columbia and adjacent waters. *Continental Shelf Research*, 89-104.

De Boer, M., Jones, D., Jones, H., & Knee, R. (2018). Spatial and Temporal Baseline Information on Marine Megafauna-Data Facilitated by a Wildlife Tour Operator. *Open Journal of Marine Science*, 76-113.

De'ath, G. (2007). Boosted trees for ecological modeling and prediction. *Australian Institute of Marine Science*, 243-251.

Edwards, A., Jones, K., Graham, J. M., Griffiths, C. R., MacDougall, N., Patching, J., . . . Raine, R. (1996). Transient Coatal Upwelling and Water Circulation in Bantry Bay, a Ria on the South-west Coast of Ireland. *Estuarine, Coastal and Shelf Science*, 213-230.

Elith, J., & Leathwick, J. (2008). Tutorial for running boosted regression trees. *Journal of Animal Ecology*.

Elith, J., Leathwick, J., & Hastie, T. (2008). A wokring guide to boosted regression trees. *Journal of Animal Ecologt*, 802-813.

Embling, C., Gillibrand, P., Gordon, J., Shrimpton, J., Stevick, P., & Hammond, P. (2010). Using habitat models to identify suitable sites for marine protected areas for harbour porpoises (Phocoena phocoena). *Biological Conservation*, 267-279.

Evans, P. G., Anderwald , P., & Baines, M. (2003). *UK cetacean status review.* Oxford: Report to English Nature and Countryside Council for Wales.

Franks, P. J. (1997). Sink or swim: accumlation biomass at fronts. *ICES Journal of Marine Science*, 161-167.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 1365-1381.

Gomes, V. H., & et al.,. (2018). Species Distribution Modelling: Contrasting presence-only models with plot abundance data. *Scientific Reports*.

Hammond, P. S., Macleod, K., Berggren, P., Borchers, D. L., Burt, L., Canadas, A., . . . Vazquez, J. (2013). Cetacean abundance and distribution in European Atlantic shelf waters to inform conservation and management. *Biological Conservation*, 107-122.

Hartman, S. E., Hartman, M. C., Hydes, D. J., Jiang, Z.-P., Smythe-Wright, D., & Gonzalez-Pola, C. (2014). Seasonal and inter-annual variability in nutrient supply in relation to mixing in the Bay of Biscay. *Deep Sea Research II*, 68-75.

Healy, H., Coilin, M., Wall, D., O'Donnell, C., & O'Connor, I. (2013). *Marine Mammals and Megafauna in Irish Waters - Behaviour, Distribution and Habitat Use.* Galway: Marine Institute.

Hind, A. T., & Gurney, W. S. (1997). The metabolic costs of swimming in marine homeotherms. *Journal of Experimental Biology*, 531-542.

Hosmer, D. W., & Lemmeshow, S. (2000). *Applied logistic regression.* New York: Wiley Interscience.

Hoyt, E. (2000). Whale watching 2000: worldwide tourism numbers, expenditures and expanding socioeconomic benefits. *A special report from the International Fund for Animal Welfare*.

Hoyt, E. (2011). *Marine Protected Areas for Whales, Dolphins and Porpoises.* Abingdon: Routledge.

Hydes, D. J., Le Gall, A. C., Miller, A. E., Brockmann, U., Raabe, T., Holley, S., . . . Orren, M. (2001). Supply and demand of nutrients and dissolved oragnic matter at and across the NW European shelf break in relation to hydrography and biogeochemical activity. *Dea Sea Reserach II*, 3023-3047.

International Council for the Exploration of the Sea. (2008). *Report of the ICES Advisory Committee.* Copenhagen: International Council for the Exploration of the Sea.

Kimura, S., Kasai, A., Nakata, H., Sugiomoto, T., Simpson, J. H., & Cheok, J. V. (1997). Biological productivity of meso-scale eddies caused by frontal disturbances in Kuroshio. *ICES Journal of Marine Science*, 179-192.

Kirkham, K., & Shepherd, P. (2016). Natura 2000 post Brexit in the UK - Challenges and Opportunities? *The Habitats Regulations Assessment Journal*, 16-19.

Koslow, A., & Couture, J. (2013). Follow the fish. *Nature*, 163-164.

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: sensitivity and specificty. *Continuing Education in Anaesthesia Critical Care & Pain*, 221-223.

Leathwick, J., Elith, J., Francis, M., Hastie, T., & Taylor, P. (2006). Variation in dermseal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Press Series*, 267-281.

Leeney, R., Witt , M., Broderick, A. C., & Godley, B. J. (2012). Marine megavertebrates of Cornwall and the Isles of Scilly: Relative abundance and distribution. *Journal of the Marine Biological Association of the UK*.

McClellan, C. M., Brereton, T., Dell'Amico, F., John, D. G., Cucknell, A., Patrick, S. C., . . . Godley, B. J. (2014). Understanding the Distribution of Marine Megafauna in the English Channel Region: Identifying Key Habitats for Conservation within the Busiest Seaway on Earth. *Public Library of Science*.

Miller, P. I., & Christodoulou, S. (2014). Frequent locations of oceanic fronts as an indicator of pelagic diversity: Application to marine protected areas and renewables. *Marine Policy*, 318-329.

Moura, A., Sillero, N., & Rodrigues, A. (2012). Common dolphin (Delphinus delphis) habitat preferences using data from two platforms of opportunity. *Acta Oecologica*, 22-32.

National Parks and Wildlife Service. (2009). *Conservation Plan for Cetaceans in Irish Waters.* Dublin: Department of the Environment, Heritage and Local Government.

Nybakken, J. W. (2001). Plankton and plankton communities. *Marine biology: an ecological approacj*, 38-93.

O'Brien, J., Berrow, S., McGrath, D., & Evans, P. (2009). Cetaceans in Irish Waters: A Recent Review of Research. *Biology & Environment: Proceedings of the Royal Irish Academy*, 63-88.

O'Reilly, J., Maritorena, S., Mitchell, G., Siegel, D., Carder, K., Garver, S., . . . McClain, C. (1998). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research*, 24937-24953.

Pampel, F. (2000). *Logistic Regression: A Primer.* SAGE.

Pingree, R. D. (1980). Chapter 13 Physical Oceanography of the Celtic Sea and English Channel. *Elsevier Oceanography Series*, 415-465.

Pour, A. T., Moridpour, S., Rajabifard, A., & Tay, R. (2016). Influence of social and economic characteristics on pedestrian crash severity at mid-blocks. *Australasian Transport Research Forum.* Melbourne: Australia.

Prideaux, M. (2003). *The Convention on Migatory Species and its relevant Agreements for Cetacean Conservation.* Munich: Whale and Dolphin Conservation Society.

Reid, J. B., Evans, P. G., & Northridge, S. P. (2003). *Atlas of cetacean distribution in north-west European waters.* Peterborough: Joint Nature Conservation Committee .

Reiss, H., Cunze, S., Konig, K., Neumann, H., & Kroncke, I. (2011). Species distribution modelling of marine benthos: A North Sea case study. *Marine Ecology Press Series*, 71-86.

Robinson, K. P., & Tetley, M. J. (2007). Behavioural observations of foraging minke whales (Balaenoptera acutorostrata) in the outer Moray Firth, north-east

Scotland. *Journal of the Marine Biological Association of the United Kingdom*, 85-86.

Robinson, K. P., Tetley, M. J., & Mitchelson-Jacob, E. G. (2009). The distribution and habitat preference of coastally occuring minke whales (Balaenoptera acutorostrata) in the outer southern Moray Firth, northeast Scotland. *Journal of Coastal Conservation*, 39-48.

Rogan, E., & Berrow, S. D. (1996). Review of harbour porpoise Phocoena phocoena L. in Irish waters. *Report of the International Whaling Commission*, 595-605.

Smith, R. C., Dunstan, P., Au, D., Baker, K. S., & Dunlap, E. A. (1986). Distribution of cetaceans and sea-surface chlorophyll concentrations in the California Current. *Marine Biology*, 385-402.

Stafford, K. M., Citta, J. J., Moore, S. E., Daher, M. A., & George, J. E. (2009). Environmental correlates of blue and fin whale call detections in the North Pacific Ocean from 1997 to 2002. *Marine Ecology Progress Series*, 37-53.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., & De Clerck, O. (2011). Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*.

Ward, G., Hastie, T., Barry, S., Elith, J., & Leathwick, J. R. (2009). Presence-Only Data and the EM Algorithm. *Biometrics*, 554-563.

Yadav, S., & Sanyam, S. (2016). Analysis of k-Fold Cross Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. *Advanced Computing (IACC)*.

Yadav, S., & Shukla, S. (2016). Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. *IEEX Xplore*.