A comparison of the predictive abilities of four approaches for modelling the distribution of cetaceans

By

Laura Mandleberg

Thesis submitted in partial fulfilment of the requirements

for the degree of

Mres. Marine and Fisheries Science

University of Aberdeen, Aberdeen, U.K

2004

"I hereby declare that this thesis has been composed by myself and has not been accepted in any previous application for a degree. Information drawn from other sources and assistance received have been duly acknowledged."

Laura Mandleberg

23rd August 2004

Contents

Abstract

This study compared the ability of three presence-only techniques (Principal Component Analysis, Ecological Niche-Factor Analysis and Genetic Algorithm for Rule-set Prediction) and one presence-absence technique (Generalised Linear Modelling) to model and predict the distribution of harbour porpoises on the West Coast of Scotland. The application of these techniques for studying habitat preferences of marine mammals has great potential benefits yet this study is the first attempt to compare their predictive abilities for any marine species.

Accurate absence data for cetaceans are not always available, for logistic reasons (expensive surveys at sea), and ecological reasons (cetaceans spend most of their time underwater remaining undetectable to visual observers until they return to the surface). Presence-only techniques may be advantageous over traditional presence-absence approaches because a) they avoid potential bias associated with absence data (i.e. uncertainty with the separation of 'true' absences, where cetaceans are actually absent from an area, from 'false' absences, where cetaceans are present but undetectable to observers on the surface, and b) they enable the analysis of a wider range of data sources which cannot be analysed using techniques such as GLM and GAM (i.e. data *not* collected from dedicated effort-based surveys at sea e.g. public sightings databases). Presence-absence data for this study were collected at a relatively low expense using commercial ferries to carry out small scale surveys. Model predictions were evaluated both individually, using the evaluation methodologies for each technique, and then in relation to the other techniques by applying an *independent* dataset, consisting of presence-absence data, to test the predictions of each technique. ROC plots were used to assess overall model accuracy. The presence-only techniques performed as well as the presence-absence technique in terms of predictive ability and all techniques produced maps of predicted occurrence that identified key areas corresponding to current knowledge of harbour porpoise distribution in the study area. However, a composite map combining the predictions of all four techniques provided the best overall representation of actual harbour porpoise distribution in the study area. Therefore, for future modelling, a combination of different techniques may be preferential for predicting species occurrence since the limitations of any one modelling technique can be compensated by the strengths of another model.

Introduction

Background

The distribution of cetaceans is thought to be primarily influenced by the aggregation of suitable prey species (Baumgartner, 1997; Davis *et al*., 1998; Payne, 1986). The distribution of prey species is often linked to a number of oceanographic variables. For example, depth and slope play an important role in directly limiting the distribution of benthic or demersal prey species (Gil de Sola, 1993). For other cetacean prey species, such as pelagic fish and cephalopods, oceanographic variables could influence their distribution more indirectly. For example, topographically induced up-welling of nutrients, or convergence of surface waters may locally increase primary production and aggregation of zooplankton, leading to the aggregation of suitable prey species for cetaceans (Rubin, 1997). Therefore, it is likely that the distribution of cetaceans is also related to such variables.

Early whalers recognised that relationships existed and later came to realise that a greater understanding of any relationship between whale distribution and environmental factors would be greatly beneficial to them in locating whaling grounds more easily (Townsend, 1935; Uda, 1954). More recently, scientists have been investigating the relationship between cetacean distribution and oceanographic variables, but with very different objectives. By investigating cetacean-environment relationships, scientists can increase knowledge of cetacean distribution by predicting areas where cetaceans are more likely to occur. The availability of such information would have many potential benefits. For example, a greater knowledge of cetacean occurrence in an area would assist in the designation of Special Areas of Conservation for the protection of vulnerable cetacean species; such information would also be useful in the development of environmental impact assessments.

Whilst in the past, whalers based their understanding of the distribution of cetaceans in relation to oceanography on past judgement and experience, scientists today use statistical models to investigate and describe them. Such models attempt to quantify species-environment relationships by statistically relating the geographical distribution of a species with environmental factors, and can be used to predict species distribution by identifying habitat requirements. Traditionally, approaches for modelling cetacean distribution such as Generalised Linear Models (GLMs) and Generalised Additive Models (GAMs) have relied on the collection of presence-absence data. However, these methods assume that the absence data is accurate. Obtaining reliable absence data for cetaceans is problematic. This usually involves expensive surveys over long time periods involving many dedicated vessels and

trained observers, and even then it can not be guaranteed that absences represent 'true' absences. Due to the mobility of marine mammals and their ability to spend long periods of time underwater (and therefore undetectable to observers on the surface), there is always a degree of uncertainty associated with cetacean absence data. The separation of 'true' absences, where animals are actually absent, from 'false' absences, where animals are present but not detected, is difficult and leads to uncertainty when interpreting results. Hirzel *et al.* (2002b) suggest that inclusion of these types of 'false' absences in predictive modelling could substantially bias analysis and propose the use of alternative approaches to modelling species' potential distributions when there is no reliable absence data.

Modelling techniques requiring exclusively presence-only data have recently been developed, and provide an alternative approach to the modelling of species distributions. Presence-only techniques were originally created to predict fauna distributions that are especially susceptible to 'false' absences due to an animal's ability to disperse or 'hide' during field surveys (Hirzel et al. 2002b). These techniques base their model predictions on data where a species has been recorded as present and therefore avoid the inclusion of potentially inaccurate absence data. Presence-only modelling techniques are increasingly being used to study the distribution of many different organisms (Robertson *et al.*, 2001; Hirzel *et al.*,2001; Ortega-Huerta and Peterson 2004; Schweder, 2003). In addition to avoiding potential bias associated with 'false' absence data, these techniques provide scientists with the opportunity to take advantage of data sources that cannot be analysed using techniques such as GLM and GAM. For example, sightings databases or museum records, which lack associated absence/effort data, or similarly, datasets that are assembled from a variety of different sources, which have inconsistent sampling methodologies.

The application of these techniques for studying habitat preferences of marine mammals has great potential benefits allowing a wider range of 'opportunistic data' to be included in statistical analysis, thus maximising the use of available data resources. Furthermore, opportunistic sightings are often the only source of information for a species in an area. Opportunistic data may be collected from many different vessels, for example fishing vessels, commercial ships or whale-watching boats. Collecting data from such vessels can be a cost-effective way of obtaining sightings data from many areas where coverage from dedicated surveys has not been possible.

Despite the problems associated with collecting accurate absence data, very few studies have used presence-only techniques to study cetaceans (Schweder, 2003; Mendes, 2004). It would therefore be beneficial to compare the performance of presence-only techniques to presence-absence approaches to gain a greater understanding of their ability to predict the occurrence of cetaceans. This study aims to investigate the ability of three presence-only modelling techniques and one presence-absence technique (GLM) to predict the distribution of harbour porpoises (*Phocoena phocoena*) in relation to oceanographic variables on the West Coast of Scotland. Whilst previous studies have compared the predictive abilities of presence-only techniques with presence-absence approaches, they have focussed on modelling terrestrial species (Zaniewski *et al*. 2002; Brotons *et al*, 2004). The present study is the first attempt to compare modelling techniques for any marine species.

The three presence-only techniques are Ecological Niche Factor Analysis (ENFA), Genetic Algorithm for Rule-set Prediction (GARP) and a PCA-based approach. All three techniques are based on the ecological niche theory (Hutchinson, 1957) and use the environmental variables of locations where animals have been recorded as present to identify the niche occupied by a species. This niche can then be used to predict species distribution within the area under investigation. Both ENFA and the PCA-based approach compare the spatial distribution of the ecogeographic variables for localities where the species has been recorded as present with the variable distribution of the whole study area. These variables are then summarised into a few uncorrelated factors retaining most of the information. Both approaches account for most of the information in the first few factors from which probability values for habitat suitability for the whole area are then derived. Whilst in PCA the components are purely statistical, in ENFA these components have an ecological meaning: the first factor is a measure of the marginality of the species. *Marginality* is defined as the ecological distance between the species mean and the mean habitat in the study area. Therefore a high marginality value (close to one) indicates that the species occupies a niche that is very different from the overall study area. Marginality plus the remaining factors represent the *specialisation* of the species. Specialisation is a ratio comparing the range of variables over the whole study area to the range that the species occupies. A high specialisation (value over 1) indicates that the species occupies a narrow range of variables compared to the overall range of variables within the study area.

GARP differs from ENFA and PCA in that it is a machine-learning approach to modelling ecological niches of species. GARP develops a set of rules through an evolutionary

refinement procedure by generating and testing a diverse range of possible solutions based on various rules from different statistical approaches including categorical and logistic regression models. It does this until a 'best set' of rules is selected. The difference between a rule and the more familiar regression model is that a rule has pre-conditions that determine when it can be applied; when these conditions are not met, the rule is not used. In this way, GARP reduces error in predicted distributions by maximising significance and predictive accuracy of the rules (Stockwell and Peters, 1999). The optimal model is selected (Anderson *et al.* 2003) and can be projected onto a map of the study region showing the species potential geographic distribution.

Comparison of predictive ability between modelling techniques

In order to have confidence in any predictive model, or in the approach used to build it, the model's predictions should be assessed by some objective means. There are many different approaches for evaluating model performance. A common strategy for evaluating model quality has been to divide the available data into a *training* dataset to construct the model and an independent *testing* dataset to evaluate model quality. The predictive ability of a model can then be assessed by applying the *test* dataset to the model's predictions. Two types of prediction error are possible: false positives (model has predicted presence where species has been observed as absent) and false negatives (model has predicted absence where the species was observed as present). The relative proportions of these errors are typically summarised in a confusion or error matrix (Fielding and Bell, 1997). Four elements are present in a confusion matrix (Table 1). Element *a* represents the cells correctly predicted as present. Likewise *d* reflects areas where the species has not been observed and that are classified by the model as absent. Thus *a* and *d* are considered correct classifications; in contrast, *c* and *b* are usually interpreted as errors. Element *c* denotes a false negative and *b* denotes a false positive.

		Observed	
		+	**-**
Predicted	**+**	***a***	***b***
	-	***c***	***d***

Table 1 Elements of a confusion matrix with element d representing both 'true' and 'apparent' absence

Once the misclassification errors of the predictions have been calculated, there are a number of measures available to assess prediction accuracy. The simplest and most widely used measure of prediction accuracy is the number of correctly classified cases (Fielding and Bell, 1997). However, this measure is dependent on the specification of an arbitrary threshold of misclassification error, which can lead to bias when interpreting results (Altman *et al.* 1994). A more powerful approach to assessing model performance is derived from a Receiver Operated Characteristic, or ROC plot (Fielding and Bell, 1997). Unlike many assessment measures, a ROC plot is independent of any threshold of misclassification error. This is advantageous because it avoids the arbitrary judgement of thresholds, which often lack ecological justification. A ROC plot is obtained by plotting all *sensitivity* values (true positive fraction : a / a+c) on the y-axis, against their equivalent 1- *specificity* values (false positive fraction : d / b+d), for *all* available thresholds on the x-axis (Figure 1). The area under the ROC function (AUC) provides a single measure of overall accuracy for a range of misclassification thresholds. The value of the AUC ranges between 0.5 and 1, where 0.5 indicates randomness and 1 a perfect fit (no overlap in the distribution of the group scores). For example, an AUC of 0.8 means that for 80% of the time, a random selection from the sensitivity (positive) group will have a greater score than a random selection from the 1-specificity (negative) group (Deleo, 1993).

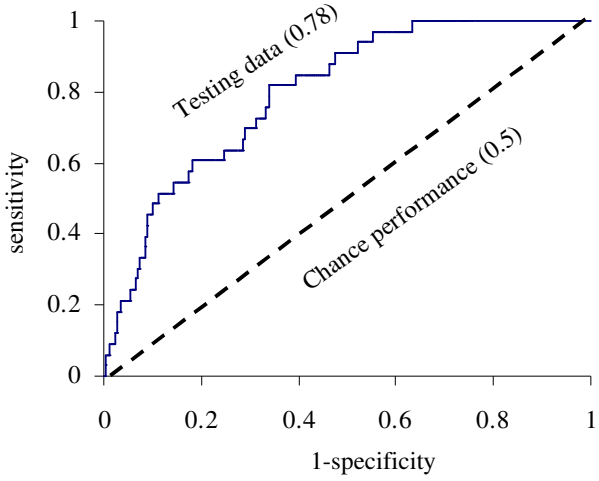


Figure 1 Example of a ROC plot (figures in parentheses are the AUC values and straight line indicates randomness).

'False' Absences

Due to uncertainty associated with recorded absences for cetaceans, harbour porpoises in particular (Palka, 1996), there is the potential for inclusion of 'false' absences in the *testing* dataset. 'False' absences would have a negative impact on the accuracy of the evaluation procedure, if present in the testing dataset, by apparently increasing the proportion of false positives, element *b* from the confusion matrix. For example, a *test* point of observed absence, falling into an area of predicted presence could be interpreted, either as a 'true' prediction error of the model, or an 'apparent' prediction error, due to the inaccurate sampling of absence data. In order to try and minimise 'apparent' prediction errors due to 'false' absences in the *testing* dataset, stricter criteria were applied to the *testing* datset before establishing absence.

Study Area

This study was conducted off the West Coast of Scotland. The study area included the Inner Hebridean Islands of Mull, Coll, Tiree, Colonsay, Kerrera, the Small Isles (Rum, Eigg, Muck, Canna) and also the outer Hebridean Island of Barra (Figure 2). The area encompasses a variety of possible cetacean habitats: shallow, inshore areas (e.g. the Sound of Mull), coastal areas (e.g. the Garvallach islands), offshore islands (e.g. Barra) and deeper, open water (e.g. the Sea of Hebrides). The oceanographic features of the West Coast of Scotland are complex and are largely determined by the properties and movements of waters of three main systems: the North Atlantic, the Irish Sea and freshwater inputs from neighbouring sea-lochs. The predominant current flowing through the waters of the west coast is the Coastal Current, which is formed by the mixing of water leaving the Irish Sea, through the North Channel, with Clyde water to form a low-salinity stream. The Coastal Current travels north along the Scottish coast, past the Inner Hebrides into the Little Minch. (Ellett, 1979). Strong tidal streams and currents induce the complex mixing of waters, especially around headlands and islands, making the West Coast of Scotland an area of high biological productivity with a rich species biodiversity. This is evident in the high numbers of cetaceans recorded in the region. The most commonly sighted species in the study area are, the harbour porpoise (*Phocoena phocoena*), bottlenose dolphin (*Tursiops truncatus*), common dolphin (*Delphinus delphis*), minke whale (*Balaenoptera acutorostrata*), Risso's dolphin (*Grampus griseus)*, white-beaked dolphin (*Lagenorhynchus albirostris*), long-finned pilot whale *Globicephla melaena,* and the killer whale *Orcinus orca* (Shrimpton and Parsons, 2000).

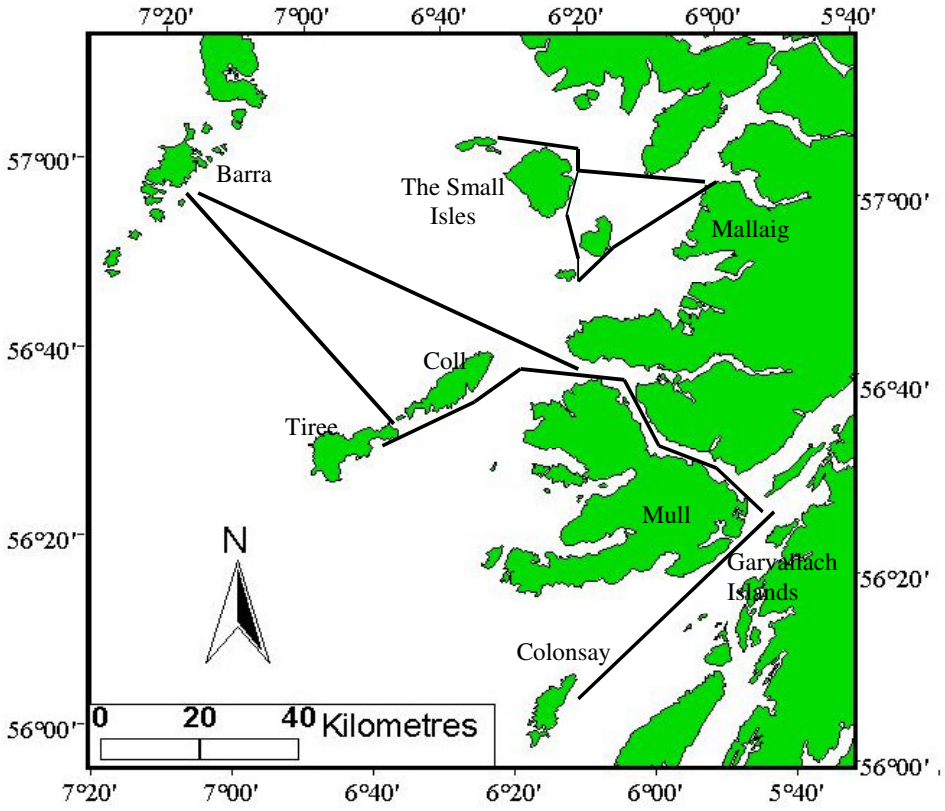


Figure 2 Map of the study area and survey routes

Harbour Porpoise (*Phocoena phocoena*)

Harbour porpoises are common and widely distributed in most coastal wasters of the study area (Shrimpton & Parsons 2000). Areas where harbour porpoises are regularly observed include the Outer Hebrides; Sea of Hebrides; North Minch; Sounds of Sleat and Raasay; the Small Isles and the coastal waters of Skye; (Evans, 1997b). The Sound of Mull; the coastal waters of the Firth of Lorn; the Sound of Jura; Kilbrannan Sound; Kyles of Bute; Coll; Tiree; the Treshnish Isles; Colonsay; Oronsay; Islay and Arran (Evans, 1997a). They also occur in many of the mainland lochs.

Despite being the most frequently encountered cetacean off the West Coast of Scotland, harbour porpoises are not easy to detect during surveys. This is partly due to their small body size (typically less than 2m in length), small group sizes (average group size of two; Pollock *et al*. 2000) and often 'cryptic' behaviour at the surface. Furthermore, increasing sea state can have a negative effect on 'detectability' of harbour porpoises at the surface and it is recommended that surveys of harbour porpoises are not carried out in sea states greater than Beaufort three (Palka, 1996)

Materials and Methods

Data collection

The data used for the model analysis were collected from commercial passenger ferries (run by Caledonian MacBrayne). The ferries make regular crossings throughout the study area (Figure 2) along set routes enabling an area to be consistently surveyed on multiple occasions at a relatively low cost. Cetacean sightings data were collected during two consecutive summer periods: From the 5th of May to 9th of July 2003 as part of a previous MSc project (Schweder, 2003), and from the 7th of May to 25th of July 2004, specifically for the present study. Three different vessels were used for the surveys: MV *Clansman* (99m length, eye height for surveying: 14.7m) was used to survey the routes to Coll, Tiree and Barra, and MV *Lord of the Isles* (85m length; eye height: 16.6m) was used to survey the Colonsay route. A third, smaller ferry, MV *Loch Nevis* (49m length; eye height: 11m) was used to survey the Small Isles area. The survey technique used was specifically developed for conducting surveys from passenger ferries on the West Coast of Scotland by Bannon (2003) and requires an observer to be situated on one side of the ship's bridge. These platforms allowed a 135º view of the surrounding area from the side where the observer was situated, round to 20º from the ship's bow.

The following data were recorded every 15 minutes throughout the duration each survey: date, time, position of ship (latitude and longitude from GPS), direction of travel (bearing) and speed (km/h). Environmental variables were also recorded: sea state (Beaufort Scale), swell height (m), visibility (km), cloud cover (octaves) and precipitation (nil, slight, moderate, heavy). The area was scanned constantly with the naked eye and with 7 x 50mm range-finder binoculars. The binoculars had a reticulated scale that was used to estimate distances to cetaceans and a compass rose for measuring relative bearings. Each time a group of cetaceans was sighted, the following information was recorded: time, ship's position (latitude and longitude from GPS), direction of travel (bearing), relative bearing of cetacean from ship's bow (degrees 90º, 270º etc) and number of reticules vertically from cetacean(s) either to land or horizon. Cetaceans were identified to lowest taxonomic group possible. Approximate group size and presence or absence of seabirds was also recorded.

Estimating position of sightings

In order to calculate the actual position of each sighting, the following information was required:

1) Position of the observer

The position of the observer is effectively the position of the ship at the time of sighting (latitude and longitude from GPS).

2) Distance to the group of animals

The distance to the group can be calculated from the angle from the horizon or land to the group (measured in reticules from the binoculars), and was calculated using the formulae from Lerczak and Hobbs, 1998 (Figure 3).

3) Bearing to the animals

The true bearing to a sighting was calculated by adding the relative bearing (compass rose) to the ship's course (subtracting 360 if the total exceeded 360).

Once all three measurements had been calculated, the position of each group of cetaceans was then estimated using the *waypoint editor* function of Garmin PCX5 software (version 2.09).

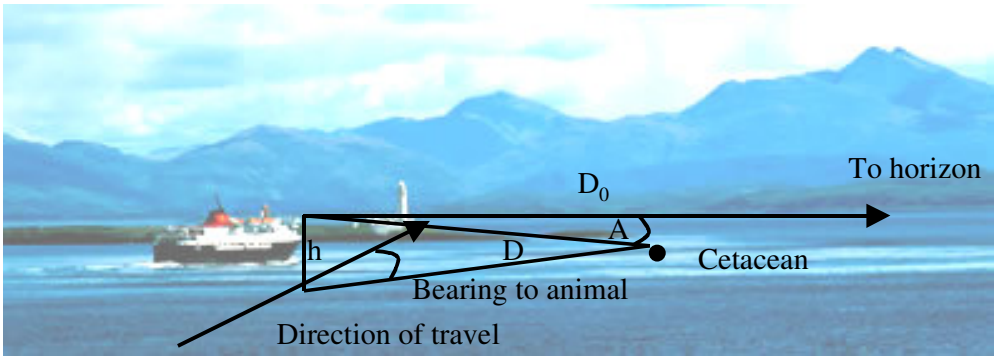


Figure 3 Survey platform and measurements used to calculate estimated position of a sighting. D_0 is the line-of sight distance to the cetacean, D is the distance to the cetacean along the surface of the water, A is the angular drop from the horizon (or land) to the cetacean, h is the height of the observer above sea level (see Lerczak and Hobbs, 1998).

Track and surveyed grid cells

A base grid was created in MapInfo® and imported into Arcview® 3.2. This enabled the entire study area to be divided into 1km x 1km grid cells (15520). The ship's position, recorded at 15 minute intervals for each survey, was plotted and the points were connected as a straight line course. The sea state at the start of each line was assigned to that segment.

Segments were then joined together for each trip, removing those segments that were recorded in sea states three or more. This was to minimise missed sightings due to lower detectability of harbour porpoises in poor conditions (Palka, 1996). In order to determine the grid cells that had been covered by the surveys, buffer zones were plotted on either side of each track line using the buffer function. The buffer limit was 750m on the side where the observer was situated and 250m from the ship's bow due to the restricted angle of vision, giving a total buffer width of 1000m. (Figure 4a). These distances were considered the maximum distances within which all porpoises present at the surface could be detected with a high certainty.

A grid cell was defined as 'surveyed' if more than half of the buffer edge fell within a grid cell (Figure 4b). Cells falling outside the buffer zones were classed as 'un-surveyed' cells. If less than 500m of buffer edge fell within a grid cell, it was classed as 'un-surveyed' (Figure 4b).

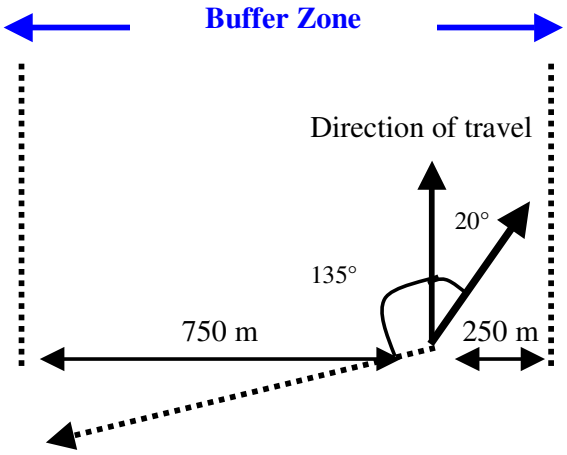


Figure 4a Buffer zones used to determine surveyed grid cells.

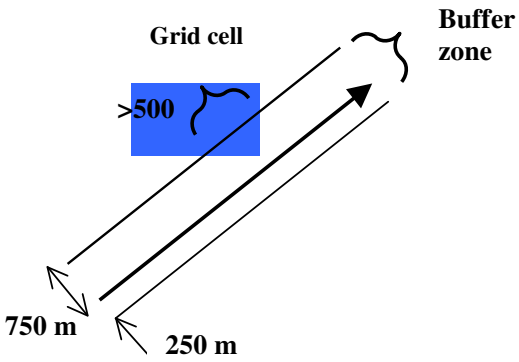


Figure 4b Schematic of track with buffer zone and grid cell

Presence and Absence cells

Surveyed grid cells were then assigned to one of two categories:

Presence - where the species was recorded at least once in a 'surveyed' grid cell

Absence – where the species was not recorded in a 'surveyed' grid cell on any occasion

Any sightings that fell within 'un-surveyed' grid cells were removed from the dataset.

Data partitioning

The dataset was randomly divided into two subsets: a *training* dataset to construct each of the models, and a *testing* dataset for independent evaluation of the models's predictions (Figure 5). This was achieved by assigning each grid cell in the study area a random number using the random grid function in Arcview®. The lowest third of the random numbers were set aside for the *testing* datset and the remaining two-thirds were used for the *training* dataset. In this way, both the 'surveyed' grid cells and the ***presence*** cells could be randomly partitioned into the testing and training datasets simultaneously.

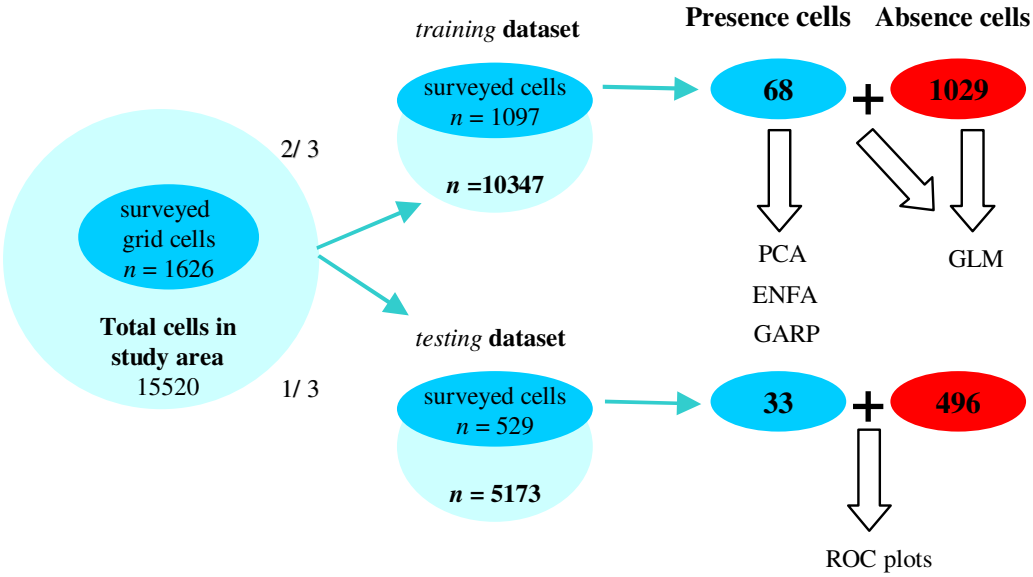


Figure 5 A summary of the random data partitioning procedure

Ecogeographic Data

The central depths for each grid cell in the study area were interpolated from the ETOP02 Global 2' elevation dataset (ETOP02 2001). The depth values were then used to calculate slope. Standard deviation of slope was calculated by comparing the central value of each square with the surrounding 24 grid squares (i.e. 5km x 5 km area) using the neighbourhood statistic. Distances to nearest coastline and latitude from the centre of each grid cell were calculated using the spatial analyst function in Arcview®. Aspect of the seabed was converted from a circular variable (i.e. degrees) into two linear components. Since aspect is measured in degrees from north, circular variables would give misleading results. For example, a cell with an aspect of 359º would give a very different value to a cell with an aspect of 1º even though in reality the two cells would be facing in a very similar direction. Therefore, aspect was separated into an easting and a northing, defined as sine (easting) or cosine (northing) of the original aspect value. Both eastings and northings ranged from –1 to 1, with a positive value

indicating a tendency to face east or north and a negative value indicating a tendency to face west or south respectively. This gave seven separate grids of ecogeographic variables (EGV) matching exactly the dimensions of the cells in the base grid (15520 1km x 1km grid cells).

1) PCA-based technique

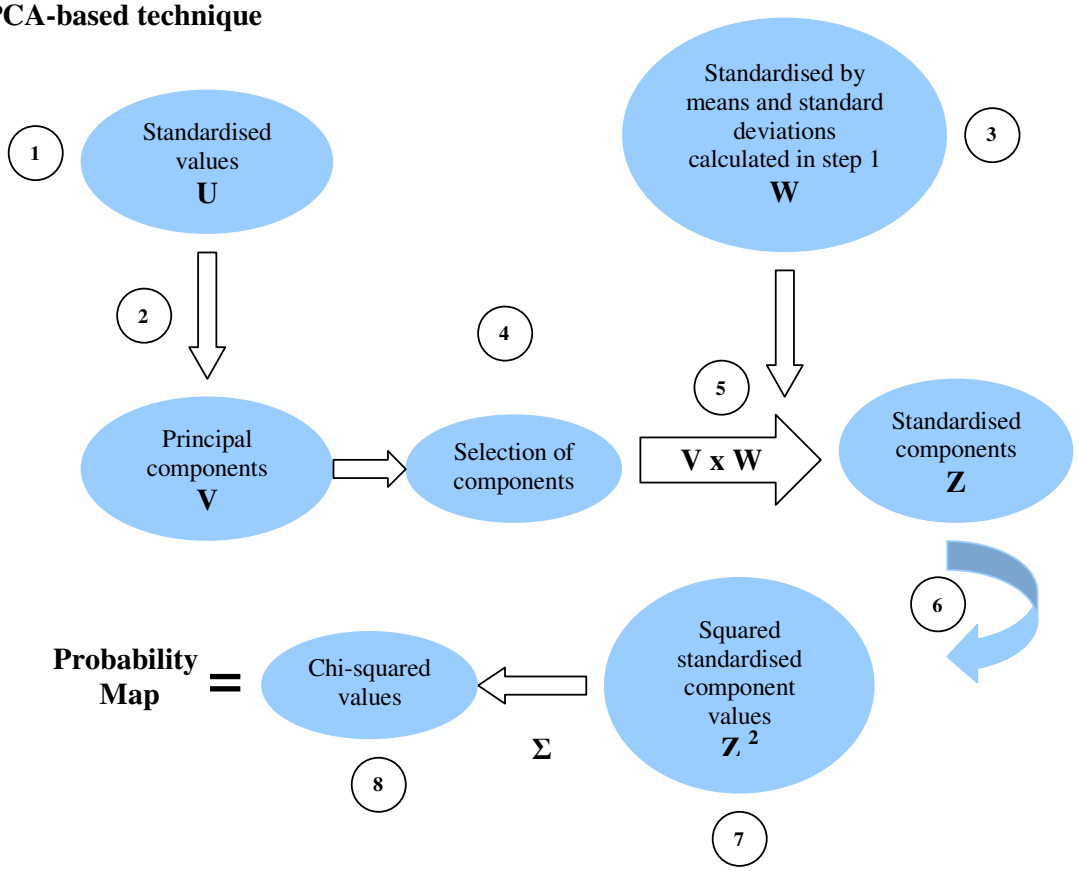


Figure 6 A summary of the steps involved in Principal Component Analysis (see Appendix 1 for description of steps 1-8). The full methodology for constructing PCA-based predictive models was followed as described by Robertson *et al*. (2001).

Ideally, all combinations of variables would have been compared to determine the combination that produced the most accurate model. However, due to time constraints for this study, it was only possible to carry out PCA on nine different combinations of ecogeographic variables (Table 2). The combinations of variables were chosen partly based on those used by Schweder (2003) and partly based on consideration of which variables might play an ecological role in influencing distribution.

PCA Intra-Model Evaluation

PCA models were evaluated using the *testing* dataset which was reserved for overall model evaluation. AUC values from the ROC plots were compared to determine the model with the highest predictive ability. To investigate whether false positives were resulting in an 'under-assessment' of model accuracy, further ROC plots were created using cells that had been surveyed more than three times and five times respectively. A map using probability of occurrence values was created in Arcview® for the best performing model.

Model	Variables
1	depth, latitude, distance to coast
2	all variables – no latitude
3	all variables – no latitude, no aspect (north)
4	depth, slope, distance to coast
5	latitude, distance to coast, depth, slope
6	all variables
7	all variables – no aspect (north or east)
8	distance to coast, depth, slope, standard deviation of slope
9	distance to coast, depth, standard deviation of slope

Table 2 Variables used in the construction of PCA models

2) Ecological Niche-Factor Analysis (ENFA)

Occurrence data and grids for all seven ecogeographic variables were imported into BIOMAPPER 3 software in IDRISI raster format. Ecological niche-factor analysis was performed and McArthur's 'broken stick' method was used to determine how many factors to include in the calculation of a habitat suitability map (Hirzel *et al.*2002). A habitat suitability map was created automatically in the BIOMAPPER 3 software, which assigns to each grid cell in the study area a habitat suitability index (HSI). The HSI ranges from zero to 100, and is a re-scaled version of the initial probabilities of occurrence calculated from the ENFA factors. Cells with low HSI scores indicate low habitat suitability, whilst high HSI scores indicate high suitability. A second ENFA was performed using all the variables except for latitude since it was suspected that this variable was causing bias in predictions due to most survey effort being concentrated in the central part of the study area (Figure 10).

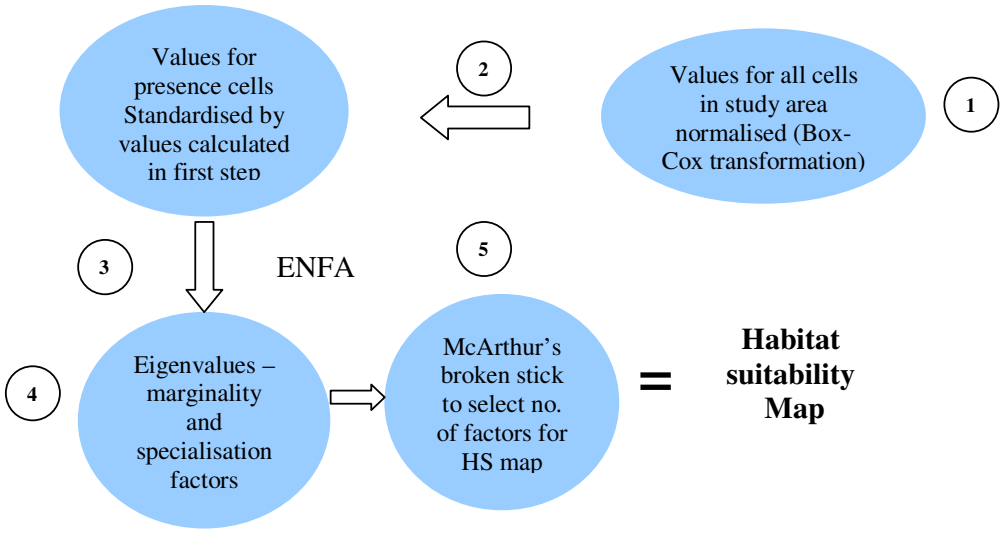


Figure 7 Summary of the steps involved in Ecological Niche-factor Analysis

ENFA Intra-Model Evaluation

The predictive accuracy of the two habitat suitability maps was evaluated using the cross-validation 'jack-knife' function in the software. The presence cells (68) were randomly partitioned into 5 subsets and a model was re-constructed five times. On each occasion, one of the subsets was left aside to serve as an independent set to validate the data. For each jack-knife run the resulting HSI values were then grouped into one of four bins (Table 3).

HSI value	**0-24**	**25-49**	**50-74**	**75-100**
Classification	unsuitable	poor	suitable	core

Table 3 Classification of ENFA cross-validation bins

Each bin covers a proportion of the study area and a proportion of the validation points. In order to take into account of uneven representation of variable classes in the study area, each bin class was standardised as follows by calculating its area adjusted frequency:

$$\mathbf{F}i = \mathbf{N}i \,/\, \mathbf{A}i$$

N*i* = the proportion of *validation points* falling in the *i*th bin class compared to the whole study area

A*i* = the proportion of cells falling in the *i*th bin class compared to the whole study area

The means and standard deviations of each jack-knife run were calculated and plotted as a histogram. If the HS map is completely random, one expects the area adjusted frequency value to equal 1 for all the bins (Figure 8a). However, if the model is good, low habitat

suitability should have a low value (below 1), and high habitat suitability, a high value (above 1) (Figure 8b).

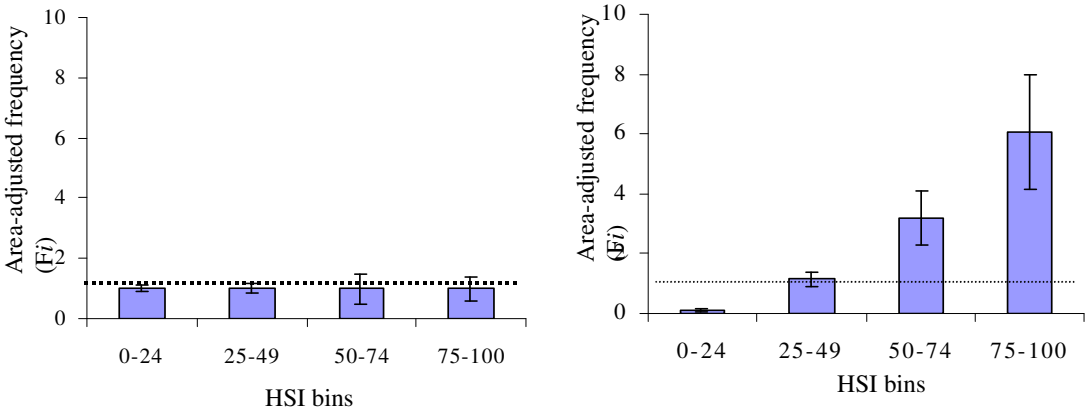


Figure 8 ENFA area-adjusted frequency histograms **a)** example of even distribution of points in each bin indicating species has a random distribution in terms of HSI within the study area and **b)** majority of validation cells are classed in HSI > 50 indicating species has a different distribution in terms of HSI from the all of the cells in the study area (the global distribution).

3) Genetic Algorithm for Rule-set Prediction – GARP Modelling System

The occurrence data and the seven ecogeographic variables were imported into the GARP software. The percentage of points to be used for 'training' i.e. model construction was specified as 67%. These points were randomly selected by GARP leaving the remaining *test* points for subsequent model evaluation. Test points (33%) were withheld completely from GARP's model building process and were only used for the internal evaluation process. Due to time constraints, it was not possible to test for all possible combinations of the variables; therefore, all possible combinations of three variables were tested. This created models for thirty five different combinations of variables, with twenty runs for each model. Each model was run twenty times, to reduce the impact of a chance relationship. The model with the highest mean predictive accuracy out of the thirty-five models was selected using the evaluation procedure described below. Another model was constructed using all of the variables to investigate whether including all seven variables produced a more accurate model. A final model was constructed using all variables except latitude, since it was suspected that latitude was causing bias in predictions due to most survey effort being concentrated in the central part of the study area (Figure 10).

GARP Intra-model evaluation

The presence data points set aside for model evaluation (33%) were used to test the model predictions automatically in the GARP software. To assess model performance, the model with the lowest *omission* error was selected. Omission error is the proportion of *test* points falling outside the predicted area (out_{test}/n_{test}), where out_{test} = the number of test points falling outside predicted areas and n_{test} = the number of test points. The *commission* index is the proportion of the study area predicted presence. The mean omission error and mean commission indices were plotted for all 35 combinations of variables and the combination of variables with the lowest mean omission error, was selected as the best model. The mean omission error was also calculated for the models constructed using all variables, and all variables except for latitude. A map of predicted occurrence was produced for the model with the lowest omission error. A composite map was produced by summing all twenty runs for each model, using a value of 1 for grid cells with predicted presence, and 0 for predicted absence. The resulting map of occurrence contained grid cells with values ranging from 0-20, representing the number of runs predicting harbour porpoise presence (20 in cells where all runs of the model predicted occurrence and zero where no runs predicted occurrence).

4) Generalised Linear Modelling (GLM)

In order to analyse presence/absence (binary) data, a binomial regression was applied (*presence:* n = 68 and *absence*: n = 1029). All linear and quadratic terms were included as potential predictors in the building of the model. This excluded latitude as preliminary analysis of the variables using GAM showed latitude to have a complex, non-linear relationship with species occurrence. To select the model that explained the most variation using the fewest number of variables, a 'backwards stepwise' procedure was used using the BRODGAR software (Highland Statistics Ltd). A stepwise model builds a model by eliminating different variables and investigates how much they improve the fit. In this way variables that do not reduce the fit by a significant amount are 'dropped' from the model. The statistic used to select the final linear model was the Akaike Information Criterion (AIC, Chambers and Hastie 1997).

In order to maximise the quality of the absence data, a second model was constructed using only cells that had been surveyed on three or more occasions (presence cells: n = 64 and absence cells: n = 679). The same method was used as the first model to select the variables producing the best 'fitted model'. The AIC from this model was compared with the AIC

calculated from the first model and a map of predicted occurrence was produced for the model with the lowest AIC value.

Probabilities of occurrence were calculated for all grid cells in the study area by substituting the intercept value and the coefficients for each of the variable into the following equation:

$$Y = e^{g(xi)}$$

Where $g(xi)$ is the equation of a straight line ($y = m\,x + c$).

Inter-model evaluation

The *testing* dataset was used to assess the relative performance of all modelling techniques.

In order to minimise potential 'false' absences in the *testing* dataset, the model's predictions were evaluated first, using all of the surveyed grid cells in the study area, and then to grid cells that had been surveyed more than three and five times respectively. ROC plots were created for all models using the Analyse-it Laboratory Software Ltd and Area Under the Curve (AUC) values for each technique were compared to determine the 'best' model.

Presence-absence thresholds

In order to make direct comparisons between the models and their predictions of occurrence within the study area, a method for setting the threshold between predicted presence and absence was required. This was achieved by finding the point on the ROC plot which maximised both sensitivity and 1-specificity i.e. the point at which the true positive fraction ceased to increase at a greater rate than the rate of increase in 1-specificity (false positive fraction). This point was found by applying a line with the same slope as a random model (i.e. AUC of 0.5) to the top left of the ROC graph and moving this line until it touched the ROC curve. The point at which the line first touched the ROC curve identified a particular sensitivity/1-specificity pair, the value for which gave the desired presence/absence threshold (Figure 9). Thus predictive maps were produced that enabled the predictions of presence and absence produced by each technique to be directly compared. Finally, the predictions of presence-absence for all four models were summed together to produce a composite map so that areas where all four models predicted harbour porpoise presence could be identified. Values for this map ranged from 0 – 4. Zero indicating areas where none of the models predicted occurrence and four indicating areas where all models predicted occurrence.

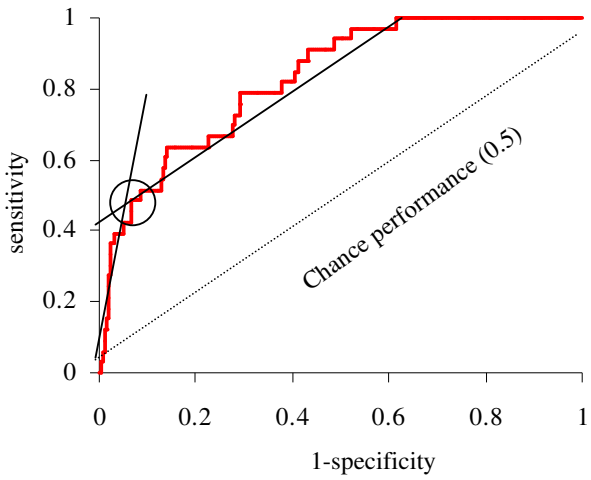


Figure 9 Method used to set the presence-absence threshold for each technique. Threshold derived from ROC curve as the point at which the straight line (with same slope as line indicating chance performance i.e. AUC = 0.5) first touches the ROC curve.

Results

Sightings

Between the 5th of May and 9th of July 2003 and 7th of May and 25th July 2004, an area of 1626 km^2 was surveyed over 61 days and a total of 159 harbour porpoises were recorded in Beaufort scale 3 or less within the study area (86 in 2003, 73 in 2004).

Survey tracks passed through a total of 1626 different grid cells. Many of these grid cells were surveyed multiple times, with some grid cells being surveyed on 83 occasions (Figure 10). After random data partitioning of the surveyed cells, the *training* dataset contained 68 ***presence*** cells (surveyed cells in which harbour porpoises were recorded) and 496 ***absence*** cells (surveyed cells in which no harbour porpoises were seen), and the *test* dataset contained 33 ***presence*** cells and 1029 ***absence*** cells (Figure 5).

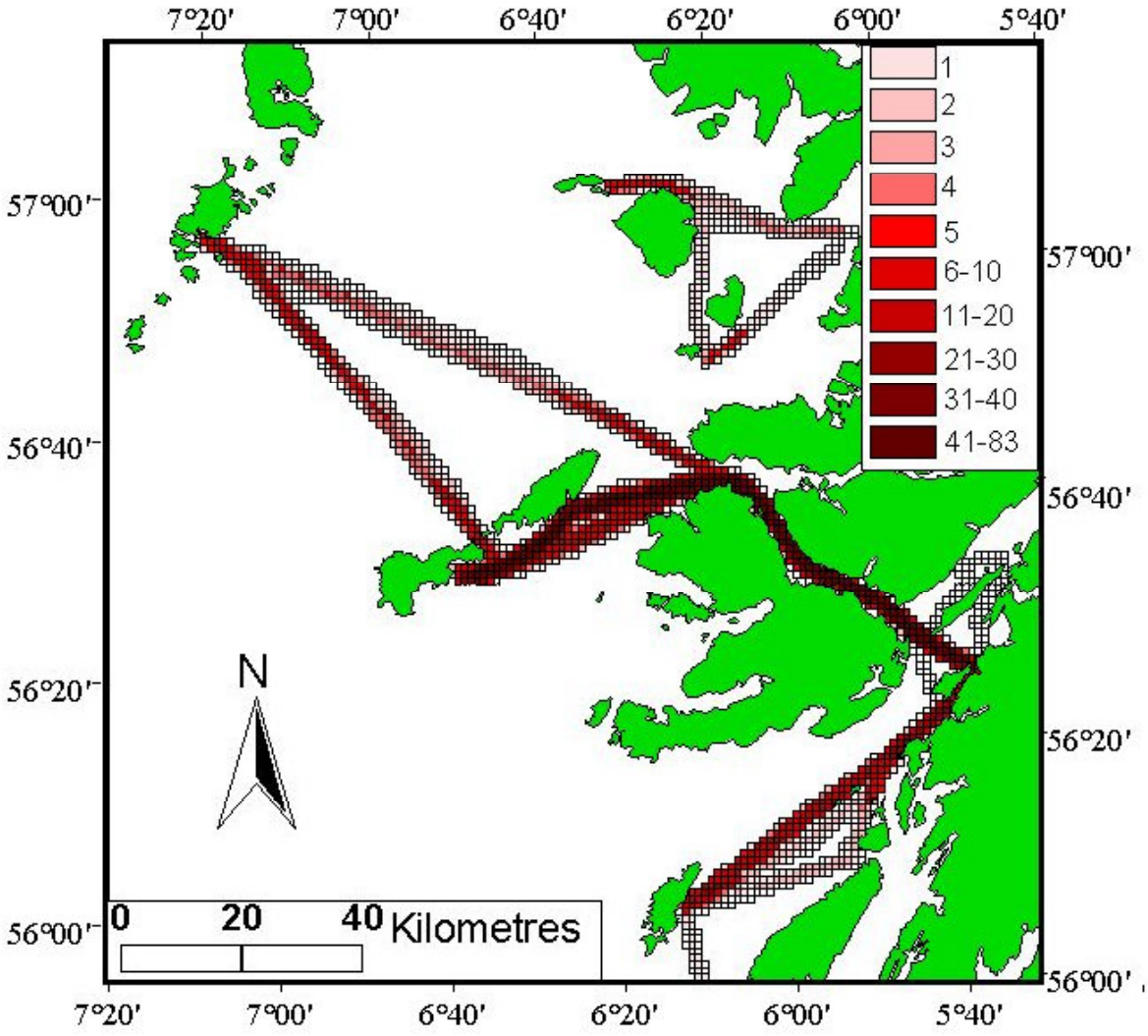


Figure 10 Map showing number of surveyed grid cells within the study area (grid cells that were surveyed at least once and covered by a buffer zone of at least 500m). Colour scale indicates number of times grid cells were surveyed.

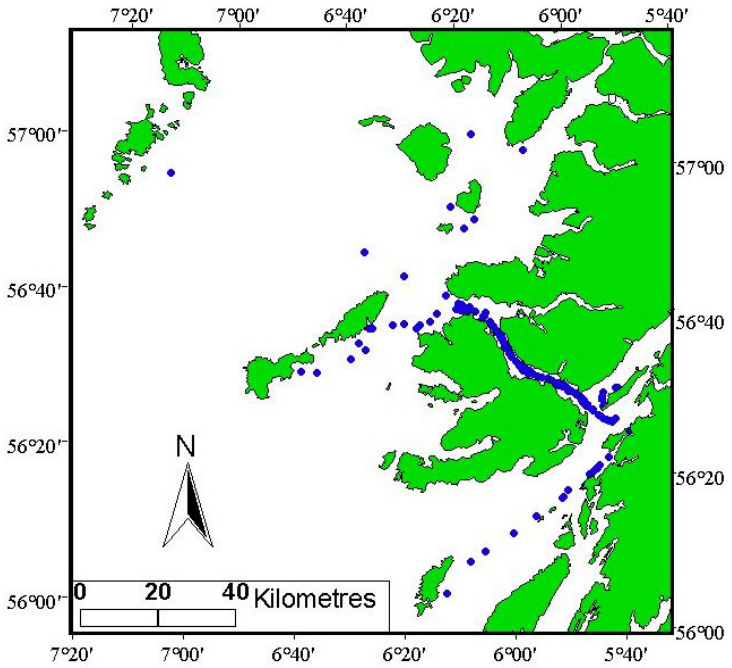


Figure 11a Harbour porpoise sightings recorded in Beaufort scale 3 or less within the study area (*n* = 159: 86 in 2003, 73 in 2004).

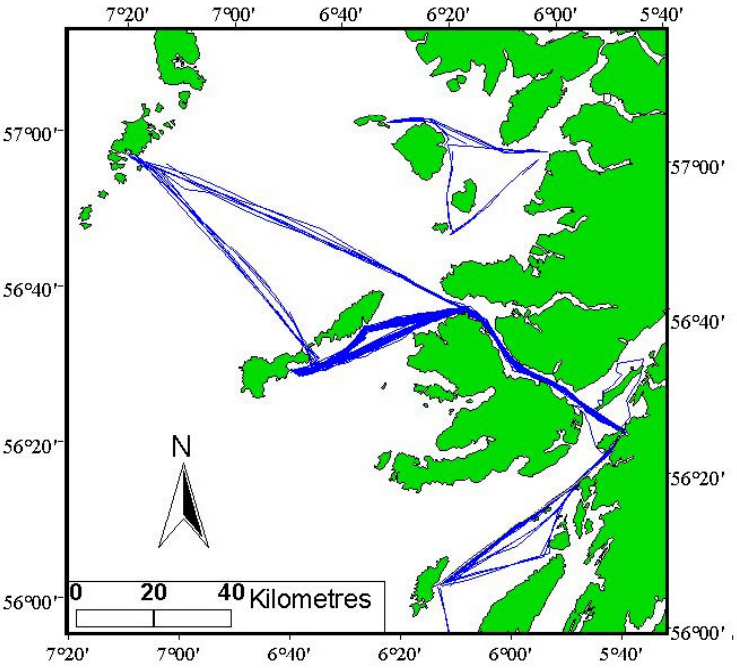


Figure 11b Map showing total track lines covered by ferry surveys during two consecutive summer periods (2003 & 2004)

1) PCA-based technique

AUC values for the PCA models ranged from 0.676 to 0.799. The two models with the highest AUC values, and therefore the best overall models in terms of predictive accuracy, were constructed using depth, latitude and distance from coast, and all variables except latitude and aspect (north). Table 5 & Figure 12

AUC	Model
0.799	**depth, latitude, distance from coast**
0.676	all variables – no latitude
0.787	**all variables – no latitude, no aspect (north)**
0.722	depth, slope, distance from coast
0.760	latitude, distance from coast, depth, slope
0.735	all variables
0.776	all variables – no aspect (north or east)
0.733	distance from coast, depth, slope, standard deviation of slope
0.726	distance from coast, depth, standard deviation of slope

Table 4 AUC values for all nine PCA models

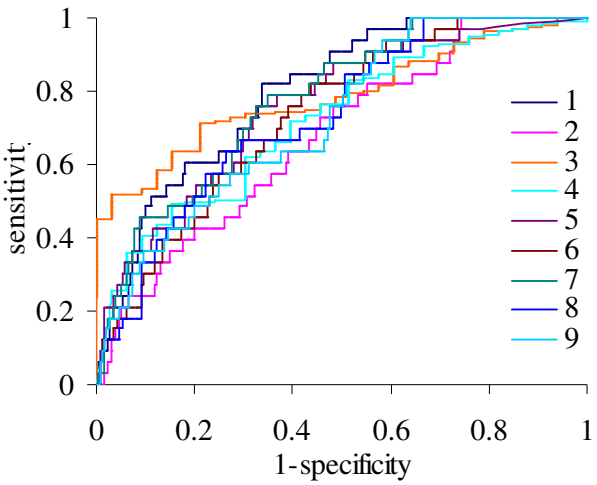


Figure 12 ROC plots for all nine PCA models

Although statistically the model constructed, using depth latitude and distance from coast, performed the best in terms of predictive ability, its predictions of harbour porpoise occurrence were limited to the central region of the study area (Figure 13a). The model failed to predict occurrence in higher or lower latitudes, for example around the Small Isles area or around the Garvallach Islands, where harbour porpoises are often sighted (Shrimpton & Parsons 2000). However, the second most accurate model (constructed using five variables; depth, distance from coast, slope, standard deviation of slope and aspect-east), produced a map which predicted harbour porpoises to occur in all areas where they would be expected to occur within the study area (Shrimpton & Parsons 2000; Evans, 1997b; Pollock et al. 2000) (Figure 13b). The 'under-prediction' of occurrence in the first model was suspected to be due to the inclusion of latitude, a variable where there was an obvious bias towards middle latitudes during data collection. As a result this model, despite its high AUC value, was rejected as biologically invalid. However, the model with the second highest AUC was identified as the most biologically sensible and was selected as the better model for inter-model testing.

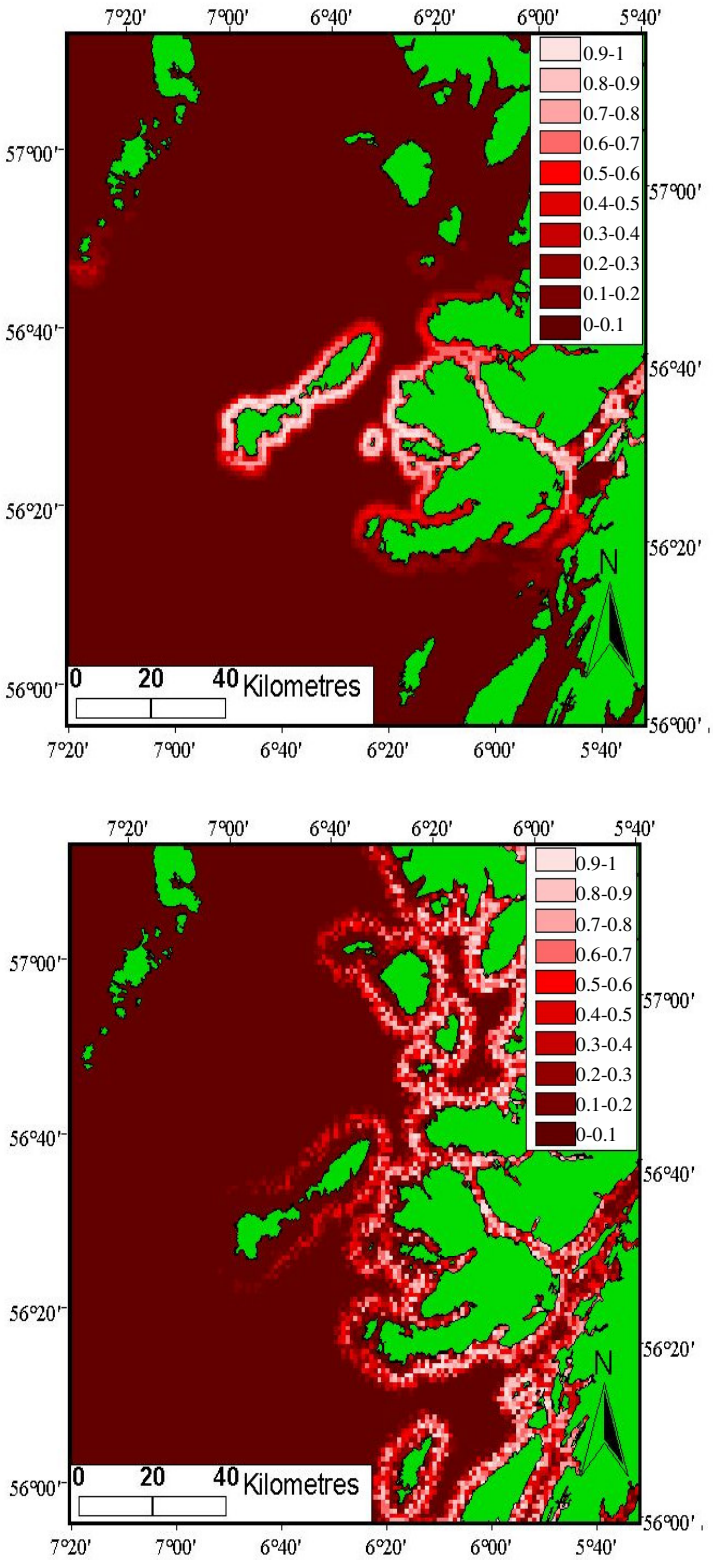


Figure 13 PCA models showing probabilities of occurrence for the Harbour porpoise (*Phoecena phocoena*) **a)** constructed using three variables; depth, latitude and distance from coast and **b)** constructed using five variables; depth, distance from coast, slope, standard deviation of slope and aspect (east)

2) Ecological Niche-Factor Analysis (ENFA)

The model that used all seven variables had an overall marginality value of 0.907 indicating that the niche occupied by harbour porpoises is quite different from the average within the study area. Specialisation was calculated at 1.674 indicating that harbour porpoises are somewhat restricted in the habitats they utilise within the study area. Three factors were retained from the seven calculated, and these three factors accounted for 100% of the marginality and 81.4% of the overall specialisation. The marginality alone accounted for 33.8% of this overall specialisation. The eigenscores from the marginality factor (the first factor) were weighted in the following order of importance: distance from the coast (-0.598), standard deviation of slope (0.575) and depth (0.522). The eigenvalue (6.628) indicated that randomly chosen cells were over six times more dispersed along this factor's axis than cells where harbour porpoises were present (Table 5).

The model that used all variables except latitude had an overall marginality value of 0.984 indicating a high marginality. Specialisation was calculated at 1.499, a slightly lower value than the previous model. Four factors were retained from the six calculated and these four factors accounted for 100% of the marginality and 88.4% of the overall specialisation. The marginality alone accounted for 50.5% of this overall specialisation. The eigenscores from the marginality factor were weighted in the same order of importance as the previous model: distance from the coast (-0.599), standard deviation of slope (0.576) and depth (0.523). The eigenvalue (6.805) indicated that randomly chosen cells were nearly seven times more dispersed along this factor's axis than cells where harbour porpoises were present (Table 5).

Model	Marginality	Specialisation	% of Marginality explained	Eigenvalue for Marginality
All variables	0.985	1.674	33.8	6.628
All variables – no latitude	0.984	1.499	50.5	6.805

Table 5 ENFA predictions of marginality and specialisation for both ENFA models

ENFA Model Evaluation

The majority of test points were re-classified into the bin representing core habitat (HSI: 75-100) for both models (Figures 14a and b) and were very different from a random distribution (F*i* = 1). Both models grouped the highest area-adjusted frequency values (F*i*) into the bin representing core habitat and the lowest F*i* values into the unsuitable habitat bin indicating that they are good models.

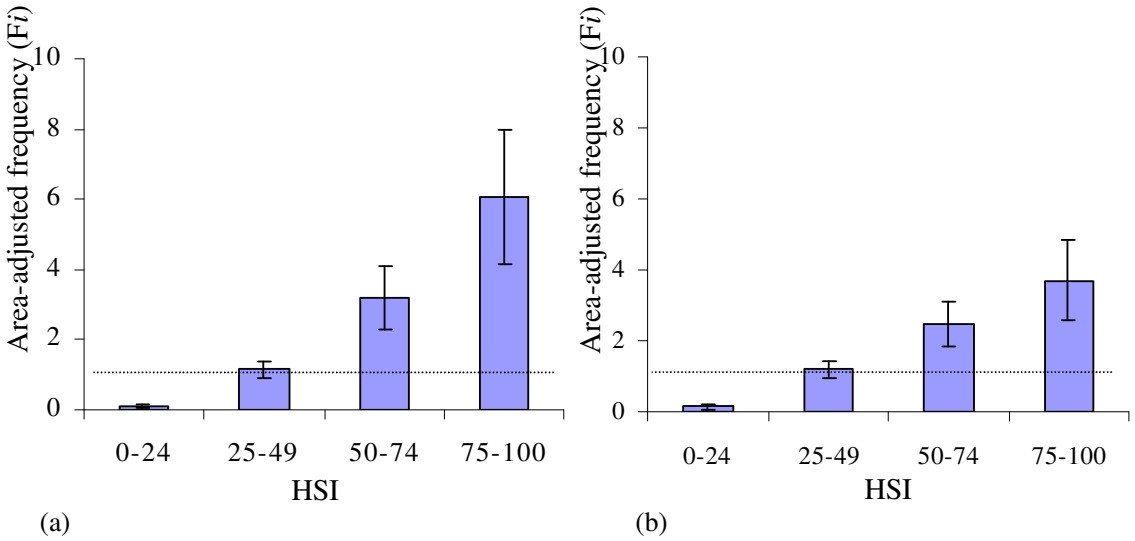


Figure 14 Graphs showing the classification of test points into habitat suitability bins for both ENFA models (a) using all variables and (b) all variables except latitude. (F*i* value of one indicates random distribution of test points)

The habitat suitability map which did not include latitude as a variable successfully identified areas of high habitat suitability in areas where harbour porpoises usually occur (Shrimpton & Parsons 2000; Evans, 1997b; Pollock et al. 2000) (Figure 15b). Whereas the model which included latitude, only identified areas of high habitat suitability in the central part of the study area (Figure 15a). Therefore the model excluding latitude was selected as the better model for inter-model testing.

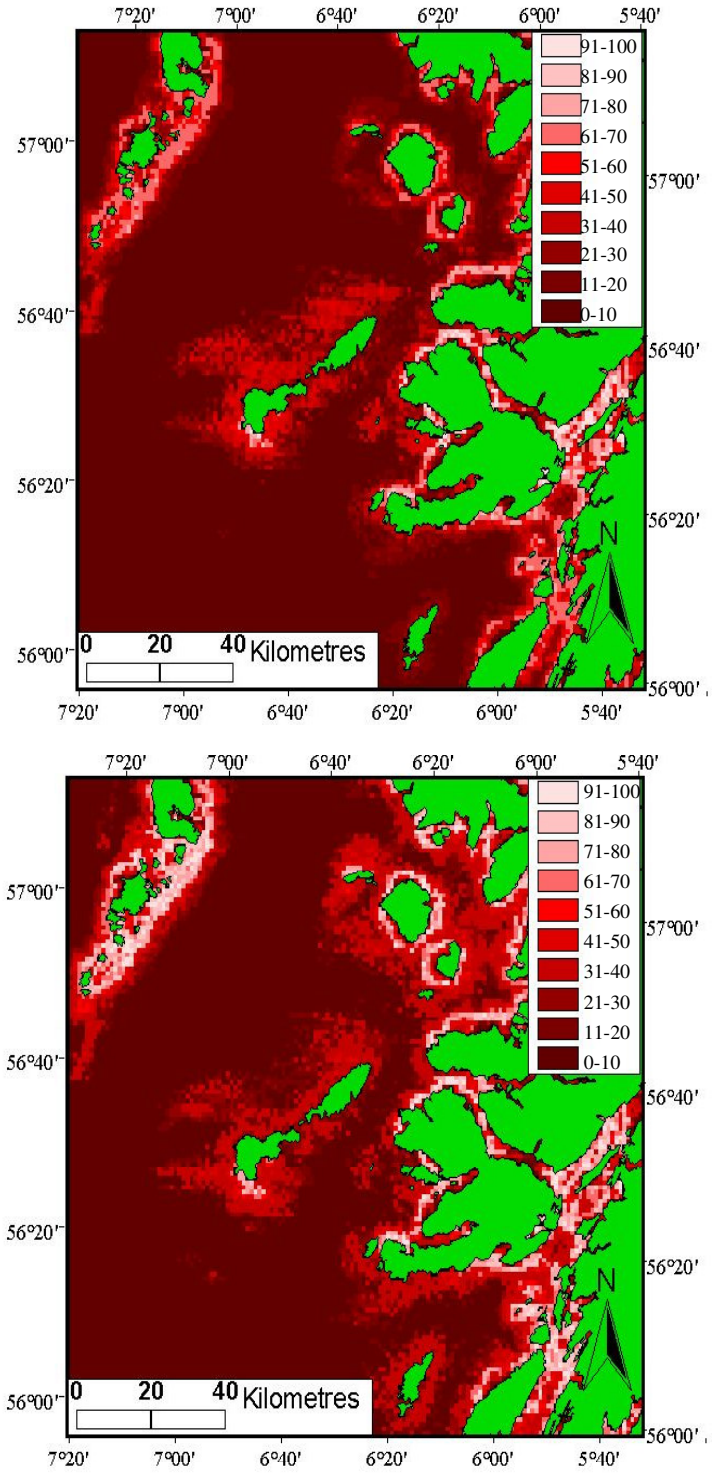


Figure 15 Habitat suitability maps for ENFA models **a)** constructed using all seven variables and **b)** constructed using all variables except latitude

3) Genetic Algorithm for Rule-set Prediction (GARP)

Out of the thirty five possible combinations of three variables, distance from coast, slope and standard deviation of slope had the lowest mean omission error (Figure 16).

The other two GARP models (using all variables and all variables *except* latitude) did not exceed the performance of the model GARP selected from all possible combinations of three variables in terms of omission error. These two models both had a higher mean omission error (Table 6 & Figure 17). The map of predicted occurrence for the model with the lowest mean omission error, highlighted all areas where harbour porpoises are expected to occur. Summing all twenty runs from this model produced a composite map which revealed consistent patterns of harbour porpoise presence, with a high agreement between the suite runs (>18) in most areas. (Figure18).

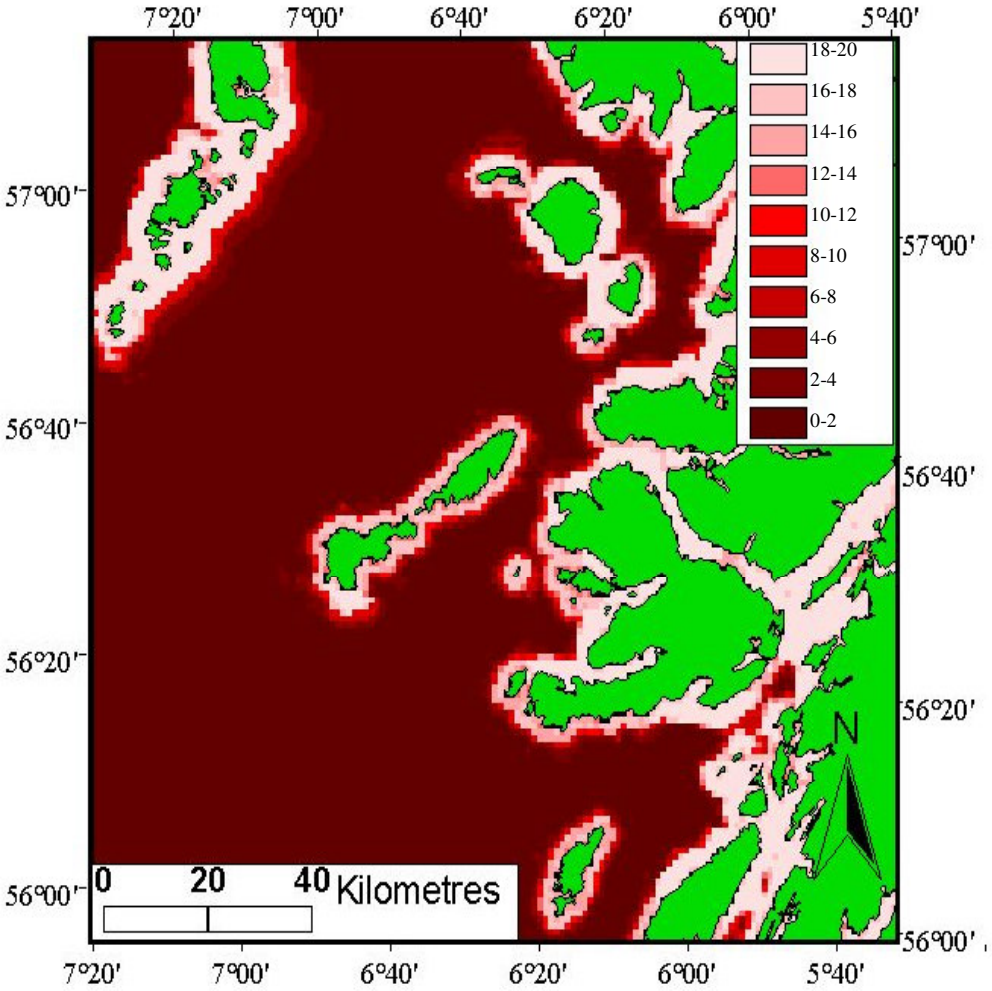


Figure 18 GARP model showing predictions of harbour porpoise occurrence (constructed using distance from coast, slope and SD slope)

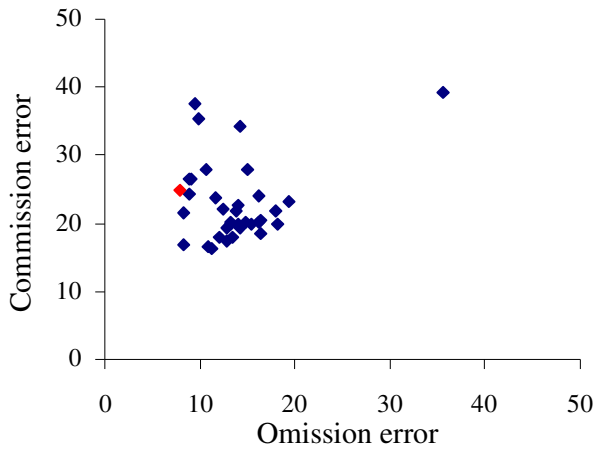


Figure 16 Mean omission/commission values for all thirty-five possible combinations of three variables. The combination with the lowest mean omission error is highlighted in red.

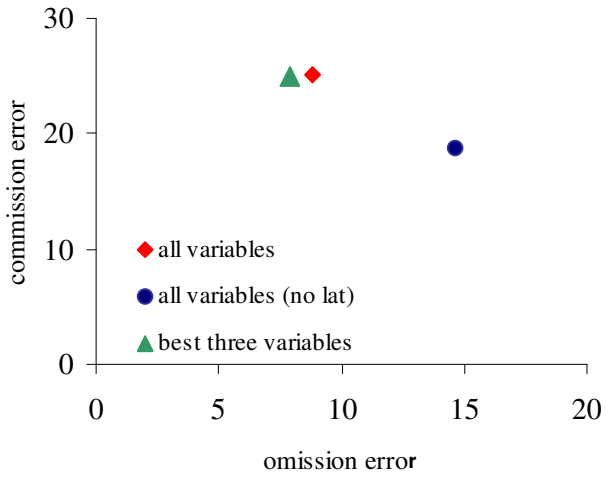


Figure 17 Mean omission/commission values for all three models

Model	Mean omission error	Mean commission error
distance from coast, slope, SD slope	7.89	24.99
all variables	8.83	25.19
all variables (no latitude)	14.6	18.82

Table 6 Mean omission/commission values for all three GARP models

4) Generalised Linear Modelling (GLM)

Using all available presence-absence cells from the *training* dataset, the model with the best 'fit' from the 'stepwise' procedure used the five variables: distance from coast ($p = 0.00532$), depth, aspect (north) and standard deviation of slope ($p = 0.00107$). Depth and aspect (north) were not significant, but including them created a better 'fit' i.e. lower AIC value (432.4), which justified including them in the model.

Using the dataset with the stricter absence criteria, i.e. with grid cells surveyed on three or more occasions, the model with the best 'fit' used three variables: distance from coast, standard deviation of slope and aspect (north). The AIC value was lower for this model than for the first GLM (363.6) suggesting that the full dataset does indeed contain a proportion of false absences. Distance from coast ($p = 0.004$) and standard deviation of slope ($p = 0.002$) were highly significant. The model with the lowest AIC value (i.e. the restricted dataset) was used to produce a map of probabilities of harbour porpoise occurrence. The map produced for this model was successful in identifying most of the key areas where harbour porpoises occur are known to occur within the study area (Figure 19).

Overall Inter-model evaluation

The *test* dataset was used to evaluate and compare the predictive ability of each of the four modelling techniques and to determine the best, biologically sensible model. ROC plots were created to assess the overall model success.

GLM had highest AUC value followed by PCA, GARP and ENFA (Table 7). Overall model accuracy estimated with the ROC method indicated that the models predicted significantly better than a random model in every case ($P < 0.05$). No significant difference was found between the techniques ($P > 0.05$) (Table 8). For each model, the *test* dataset restricted to grid cells surveyed on three or more occasions gave the highest AUC value (Figure 20a-d).

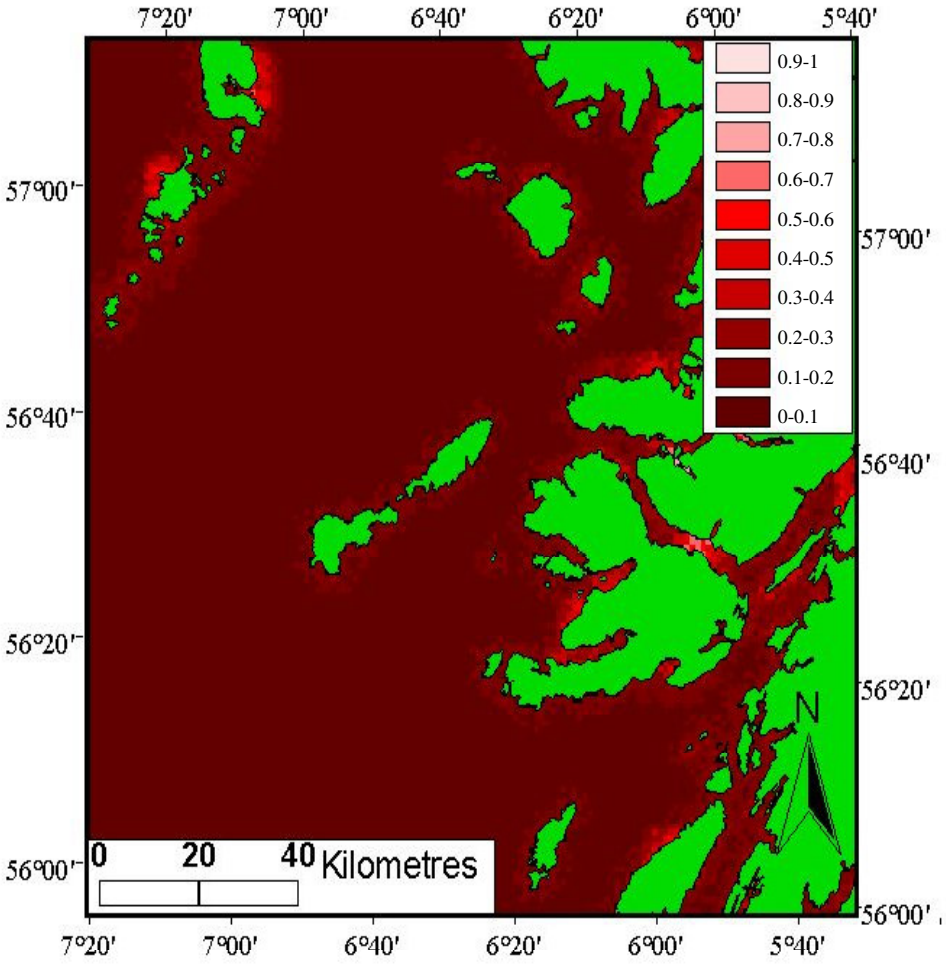


Figure 19 Probabilities of occurrence predicted by GLM technique three variables: distance from coast, standard deviation of slope and aspect (north) and grid cells surveyed three times or more.

	Variables	AUC	P value	S.E.
PCA	All variables – no latitude, no aspect (north)	0.803	0	0.03
ENFA	All variables – no latitude	0.745	$3.82E^{-10}$	0.04
GARP	Distance from coast, slope, SD slope	0.793	$5\ E^{-15}$	0.04
GLM	Distance from coast, SD slope, aspect (north)	0.828	0	0.03

Table 7 Relative performance of each modelling technique in terms of AUC (calculated from ROC plots) and significance plus standard error values (S.E) of models predicting better than random

Model vs. Model	**P value**
GLM v GARP	0.313
GLM v ENFA	0.081
GARP v ENFA	0.137
GARP v PCA	0.783
ENFA v PCA	0.166
GLM v PCA	0.504

Table 8 Statistical comparison between AUC values of the models (P >0.05 = no significant difference)

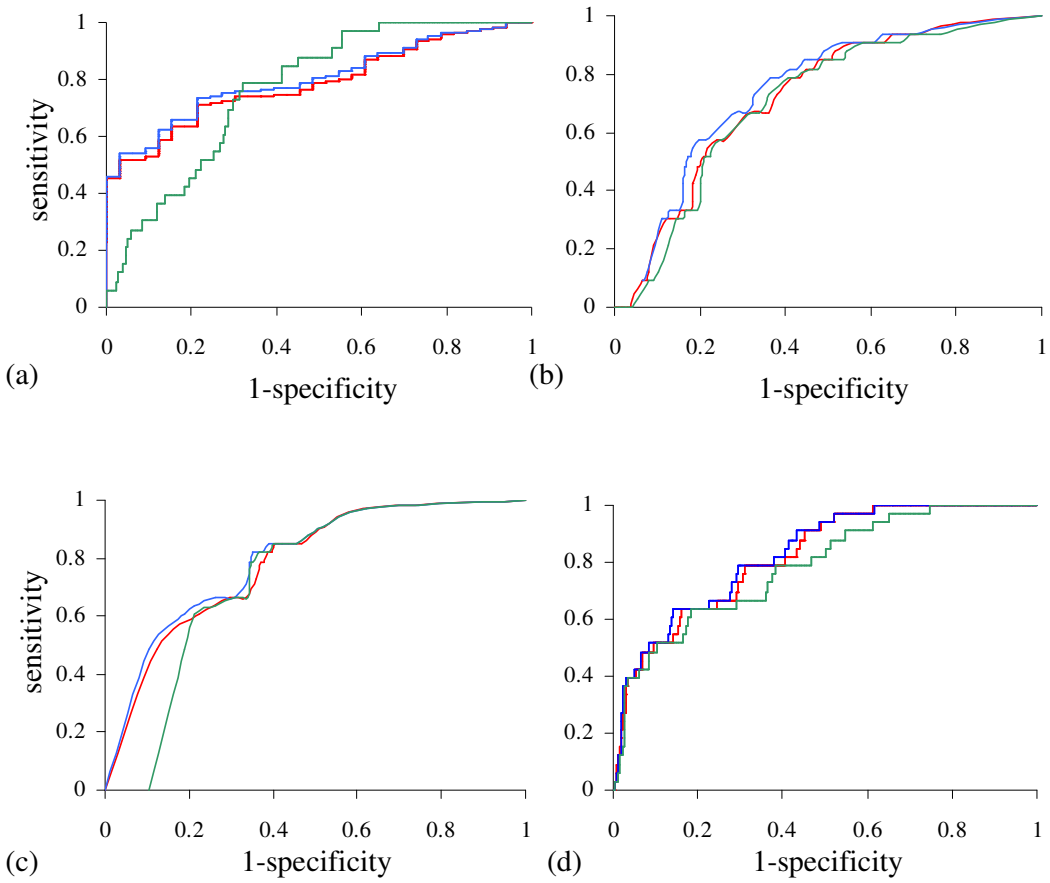


Figure 20 ROC plots for each modelling technique using *testing* dataset using all cells (**RED**), cells surveyed three times or more (**BLUE**) and five times or more (**GREEN**) a) PCA b) ENFA c) GARP d) GLM. In all cases the AUC value *increased* when testing dataset is restricted to cells surveyed three or more times.

Presence/Absence thresholds

The presence-absence thresholds derived from the ROC plots (Figure 21) produced four directly comparable maps of predicted harbour porpoise presence and absence within the study area (Figure 22 *a-d*). The PCA model predicted the widest area of harbour porpoise occurrence, whereas the other three models predicted presence over smaller areas. All models successfully predicted harbour porpoises to occur in very similar areas and identified key areas where harbour porpoises are usually seen (Shrimpton & Parsons 2000; Evans, 1997b; Pollock et al. 2000). The only area where predictions of all four models did not agree was around the coastal waters of the Outer Hebrides; all models except PCA predicted harbour porpoise presence in this area

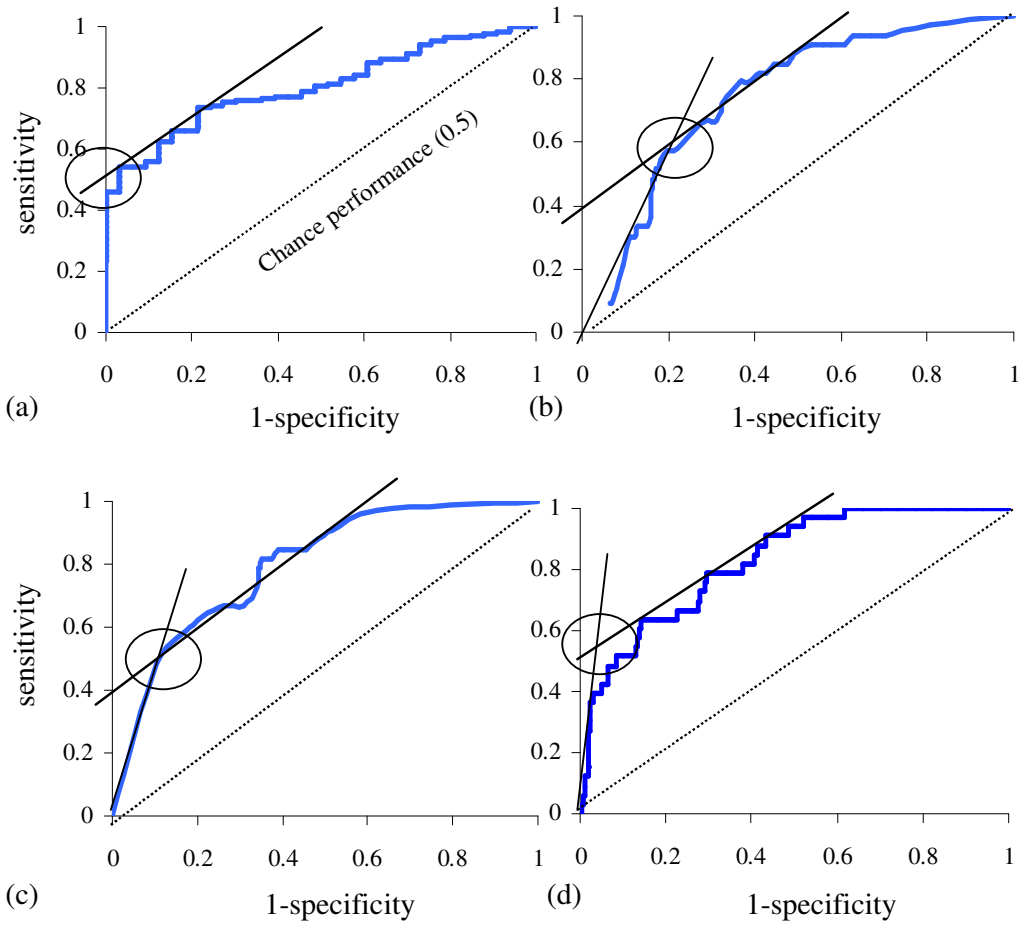


Figure 21 Resulting presence-absence thresholds set for each technique. Circled areas indicate the *1-specificity/sensitivity* threshold pairs derived from ROC curves a) PCA b) ENFA c) GARP d) GLM

Model	**Sensitivity**	*1-Specificity*	Presence/absence threshold
PCA	0.458	0	0.047
ENFA	0.545	0.182	43
GARP	0.485	0.104	19
GLM	0.485	0.066	0.201

Table 9 *Sensitivity* and *1-specificity* pairs defining presence/absence thresholds for each modelling technique

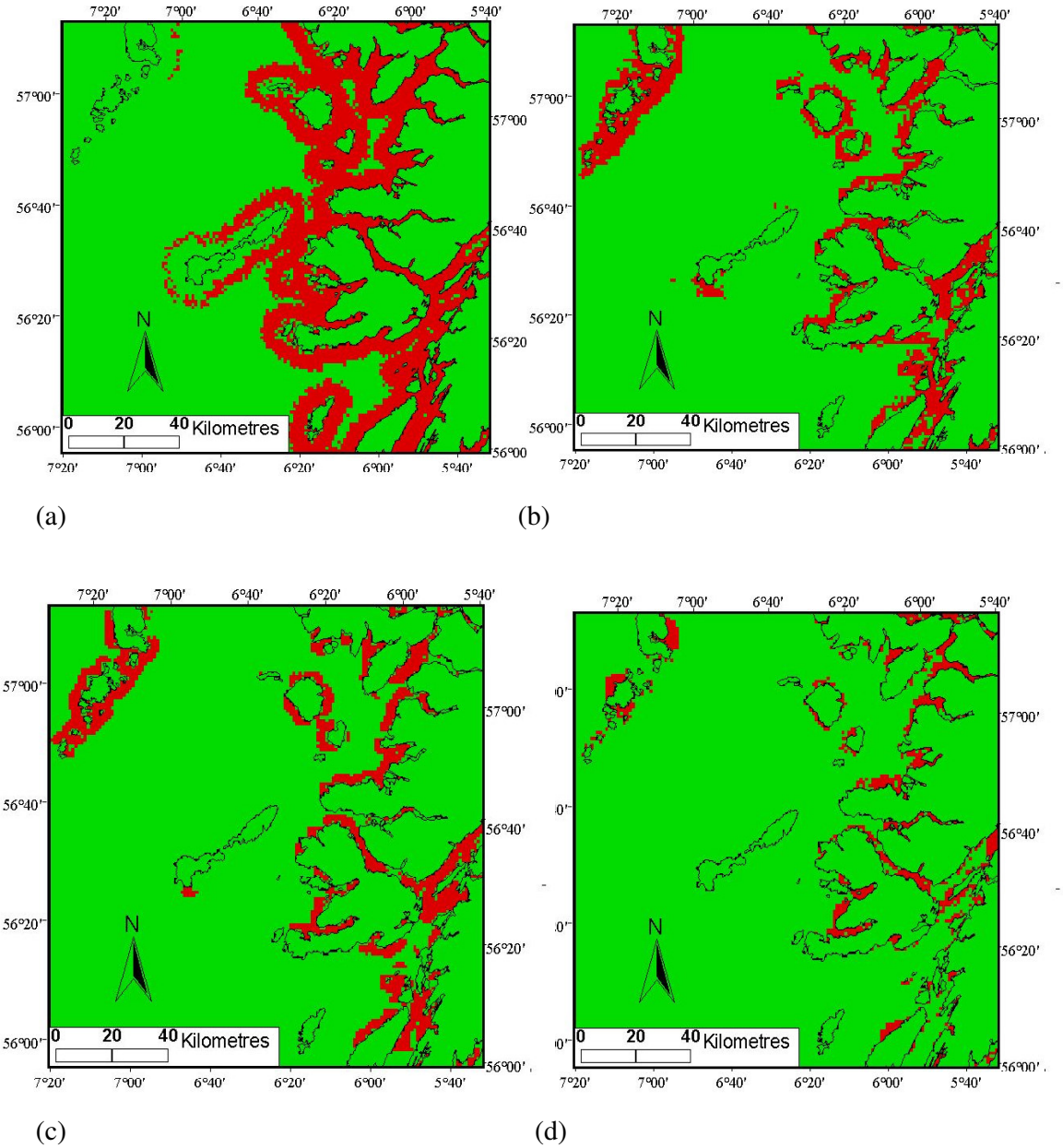


Figure 22 Presence-absence maps for a) PCA b) ENFA c) GARP d) GLM (Red areas indicate predicted occurrence)

Composite map

All models predicted occurrence in the same general areas. Values for this map ranged from 0 to 4. Zero indicating areas where none of the models predicted occurrence and four indicating areas where all models predicted occurrence (Figure 23). Values were highest off the north-west coast of Jura, along the West Coast of Mull, in the Sound of Mull, Loch Linnhe, along the coast from the Point of Ardnamuchan to the Sound of Arisaig, around the Small Isles (Rum and Eigg), in the Sound of Sleat and off the south-west coast of Skye. These areas

correspond to areas where harbour porpoises are known to occur from previous studies in the study area (Shrimpton & Parsons 2000; Evans, 1997b; Pollock et al. 2000).

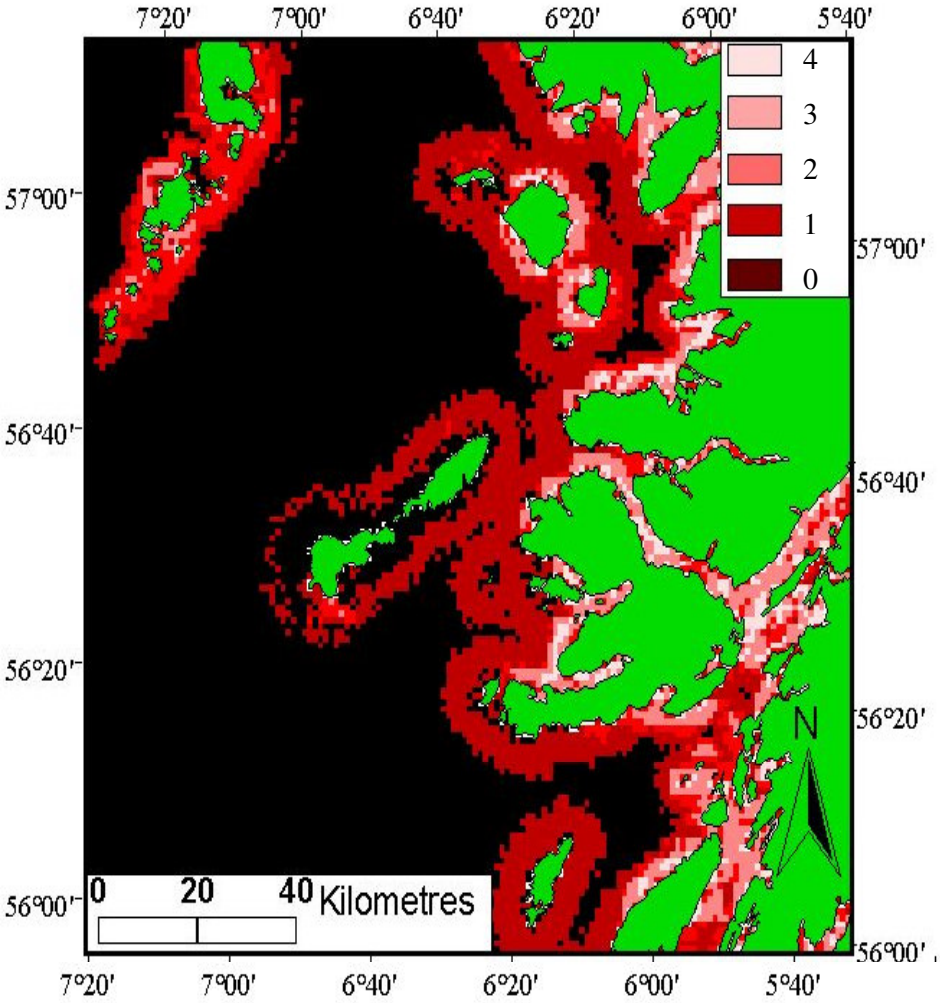


Figure 23 Composite map combining predictions of harbour porpoise occurrence for all four modelling techniques. A value of zero indicates areas where *none* of the models predicted harbour porpoise occurrence and a four indicates areas where *all* models predicted occurrence.

Discussion

Few studies have investigated the performance of two or more statistical methods when applied to the *same* dataset (Guisan and Zimmerman, 2000). Even fewer studies have compared the predictive performance of the newer presence-only techniques to the more traditional approaches, such as GLM, which rely on presence-absence data for constructing models to predict species distributions in a given area (Hirzel *et al*. 2001; Zaniewski *et al.* 2002; Williams, 2003; Brotons *et al*. 2004). Comparative studies of this kind are essential if we are to fully understand how the different techniques compare in their predictive abilities and this is the first study to do so for any marine species.

This study compared the ability of three presence-only techniques (Principal Component Analysis, Ecological Niche-Factor Analysis and Genetic Algorithm for Rule-set Prediction) and one presence-absence technique (Generalised Linear Modelling) to model and predict the distribution of harbour porpoises on the West Coast of Scotland. Principal Component Analysis (PCA), Ecological Niche Factor Analysis (ENFA), Genetic Algorithm for Rule-set Prediction (GARP) and Generalised Linear Modelling (GLM) were all successful in identifying key areas corresponding to current knowledge of harbour porpoise distribution in the study area. In addition, they were consistent in predicting the same areas of high occurrence (e.g. Sound of Mull and the Small Isles) and low occurrence (e.g. Sea of Hebrides). Differences were found between the modelling techniques in their predictive abilities, as indicated by the AUC values. However when the AUC values were compared statistically, no significant difference was found between them. This is in contrast to previous studies; Hirzel *et al.* (2001) used a modelling approach based on a virtual species with predetermined habitat preferences and found GLM predictions to be generally more accurate than those obtained with ENFA. Brotons *et al.* (2004) supported these results and found GLM predictions of the distribution of a forest bird species to be more accurate than those obtained using ENFA.

Absence Data

GLM relies on accurate absence data; if absence data are accurate and reflect low habitat suitability, inclusion of absence data can improve model quality (Hirzel *et al*. 2001). In contrast, if detectability of a species is low, the likelihood that the absence data contain 'false' absences is greater, which can cause a decrease in predictive power. The reliability of absence data is dependent on the probability of detecting the species. Cetaceans spend long periods of time underwater and therefore remain undetectable to visual observers until they surface for

air. Harbour porpoises are especially difficult to detect due to their small body size, small group sizes and often 'cryptic behaviour when at the surface. Moreover, detectability decreases further still in higher sea states. For example, it has been shown that including data collected in Beaufort sea states greater than one causes negative bias abundance estimates (Palka, 1996). Williams (2004) recently investigated the sensitivity of habitat models to detectability. Presence-absence models were found to be more sensitive to those species with a low probability of detection thus presence-only techniques were recommended for the modelling of species that are difficult to detect during surveys. Presence-only techniques in the present study performed as well as the presence-absence technique, GLM. Presence-only approaches for the modelling of cetacean species therefore provide valuable alternatives to the more traditional presence-absence techniques, particularly if the quality of absence data is poor, or if an area has not been surveyed on multiple occasions.

Low detectability of cetaceans and subsequent 'false' absences in the dataset may have limited the predictive ability of GLM in the present study explaining why it did not perform significantly better than the presence-only techniques as expected from previous studies. The GLM constructed using tougher absence criteria produced a better model (lower AIC value) than the model constructed using all surveyed grid cells, suggesting the occurrence of 'false absences'. Inclusion of 'false' absences in model construction may explain why the probability map for GLM predicted occurrence in fewer areas.

Similarly for model evaluation, the *testing* dataset was improved. When the *testing* dataset was restricted only to grid cells that had been surveyed on multiple occasions, the apparent predictive ability of the models was improved (AUC values of the ROC plots increased). Therefore, this indicates that a *testing* data with stricter absence criteria is more accurate for a 'hard to detect' species such as the harbour porpoise. In this case, a more accurate dataset was achieved by not assigning a grid cell to the absence category unless it had been surveyed at least three times. However, as the survey threshold was further increased to five surveys per grid cell, models showed a decrease in accuracy. This is attributable to a reduction in the number of grid cells in the *testing* dataset. These results are consistent with those of Schweder (2003) who found PCA models of harbour porpoise distribution peaked in accuracy when cells had been surveyed at least three times. A larger dataset would be needed to investigate if only including cells surveyed five and ten times or more still further increased predictive accuracy.

Impact of input of variables

The choice of variables is very important for presence-only models. Variables to be included in presence-only ecological modelling of a species must be representative of the niche occupied by the species to reduce any possible bias in model predictions. In terms of this study, the vast majority of surveyed grid cells fell within a narrow range of latitudes in comparison to the rest of the study area (Figure 10) despite the fact that harbour porpoises are known to occur at all latitudes within the study area (Shrimpton & Parsons 2000; Evans, 1997b; Pollock et al. 2000). Subsequent inclusion of latitude as one of the variables resulted in the production of biologically erroneous, but statistically valid models. For example, the PCA model, despite having the highest AUC value, produced a biologically insensible model, which only predicted harbour porpoises to occur in the central region of the study area. This distribution pattern does not correspond with previous studies in the area, which have found harbour porpoise distribution to stretch over a wider proportion of the study area. In particular, harbour porpoises are frequently sighted around the Small Isles (Jeewoonarian et al. 2000), an area where the PCA model including latitude failed to predict any occurrence. This example not only illustrates the importance of including only well-sampled variables in predictive models, but also emphasises that a model with a high predictive accuracy does not necessarily produce a biologically sensible prediction of occurrence.

In a comparative study of this kind, ideally models should be produced for all possible combinations of variables, enabling the best model to be constructed from the best combination of variables for each technique. Each model differs in how it selects and incorporates the ecogeographic variables to fit the model. For example, GLM selects variables that are significant in a stepwise procedure, eliminating less important variables from the analysis, whereas ENFA weights all input variables, so that less important variables are given a lower weighting and contribute less to the final model. In PCA however, the user must select different combinations of variables and the choice of input variables can dramatically influence the predictions. This was a limitation of this study as producing models for all possible combinations of variables would have been time consuming and in the time allowed it was only possible to construct nine models using different combinations of variables.

Time constraints also limited the full implementation of the 'environmental layer jack-knifing' function in GARP, which automatically selects the variables that influence occurrence the most from all possible combinations. Although this would have been an optimal approach for this study, it was only feasible to instruct GARP to run models for all possible combinations of three variables. Nevertheless, this allowed the best possible

combination of three variables to be determined which resulted in subsequent production of a model higher in predictive accuracy than the model that was constructed using all of the variables (except latitude).

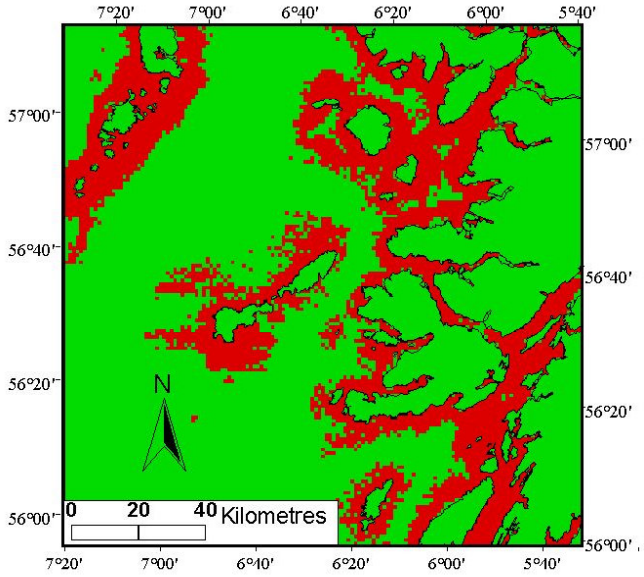
Interpreting the models

It is important to note that the different models do not report the same type or scale of predictions (e.g. probability of occurrence in PCA versus habitat suitability in ENFA) and therefore are not directly comparable with one another. For example, a GLM model predicting a probability of occurrence of 0.4 is not directly comparable with 0.4 probability of occurrence in a PCA model. In order to make direct comparisons between the predictions of the different techniques, it was necessary to set a threshold level to determine presence and absence using the same criteria in each case. By definition, the ROC plot does not provide an inherent rule for the classification threshold of separating predicted presence-absence, as it is independent of any threshold measure. However, there are strategies that may be used to develop decision rules for defining presence/absence (Zwieg & Campbell, 1993). In this study, the threshold for presence/absence was defined by finding the point on the ROC plot which maximised sensitivity and 1-specificity. Although a subjective measure, in this context it was justified as being the point at which the highest level of sensitivity (true positives fraction) could be attained without too much loss of specificity (true negative fraction). The resulting presence/absence maps (Figure 22) gave a clear comparison between the predictions of each modelling technique. All of the maps predicted harbour porpoise presence in the same key areas, however the PCA model was the only model that did not predict any occurrence in the Outer Hebrides. It is not clear why there were differences in predictions between the models in this particular area. However, survey effort was low in this area, and more survey effort would be needed to determine whether the models were accurate in their predictions in this area.

For effective implementation of a predictive model, the user must define at which point on the probability threshold to define species presence and absence i.e. a 'cut-off' point. The threshold level at which this limit is set is critical to the model's predictions of occurrence as it can dramatically change the resulting predictions (Figure 24). The defined threshold limit will also depend on what the predictive model is to be used for. For example, if the objective were to designate a special area of conservation to protect an endangered species, overestimating areas of potential occurrence might be preferable than underestimating their existence; in which case a lower threshold level of presence/absence would be defined (Figure

24a). However, if the objective were to outline areas of species occurrence for the purposes of ecotourism, then only the highest probabilities of presence might be desired, in which case a higher threshold limit of presence/absence would be applied to highlight only those areas where the animals are most likely to occur (Figure 24b).

a)



b)

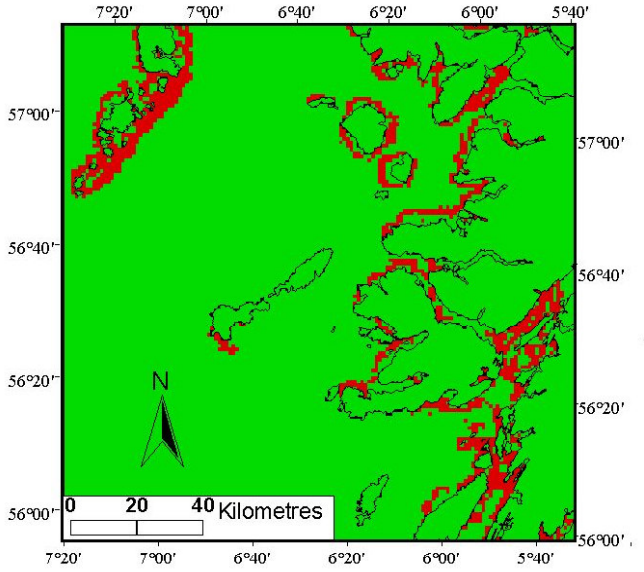


Figure 24 Habitat suitability maps with different presence-absence thresholds a) Habitat Suitability Index (HSI) < 20 indicates absence and > 20 indicates presence and b) HSI <60 indicates absence and >60 presence

Conclusions

The results of this study show that all of these techniques can be used to predict the likelihood of occurrence of cetaceans in relation to ecogeographic variables with similar levels of predictive ability. The ability of a model to accurately predict areas of high species occurrence has several beneficial implementations for conservation and species-habitat management. Not only can the models assist in identifying core areas that are important to a species, but they can also tell us more about why those particular areas are important.

For the predictive modelling of cetacean species, the choice between techniques very much depends on the data available to construct a model. Presence/absence models such as GLM can produce models of high predictive accuracy providing the absence data are reliable and bias due to 'false' absences has been minimised. However, the collection of presence-absence data for cetaceans can be expensive and involves the running of complex, effort-based surveys at sea, often requiring many dedicated vessels and observers. 'Platforms of opportunity' surveys such as the present study can allow presence-absence effort-based data to be reliably collected for relatively little cost. The PCA approach to modelling cetacean distribution requires absence data only to build a *testing* dataset, which may permit two different sampling methodologies to be combined. For example, an existing database containing presence-only data (i.e. public sightings database) can be used to construct the models, and only small-scale dedicated surveys need to be carried out to build the testing dataset (Schweder, 2003). To collect these data for cetaceans, ferries can be used as 'platforms of opportunity' at a relatively low cost. If no absence data are available, or if there is uncertainty associated with the absence data, ENFA and GARP can be successfully applied to produce models of a similar quality to the other two techniques.

The choice of technique also depends on the objectives of the model predictions. If the objective is to protect rare or endangered species, overestimating areas of potential occurrence might be more preferable then underestimating their existence (Fielding and Bell, 1997). Zaniewski et al. (2002) found that when comparing ENFA with presence-absence techniques, ENFA was more suitable for identifying areas of high conservation concern than the presence-absence techniques, which tended to 'under-predict' areas of potential biodiversity. In this study GLM may have underestimated harbour porpoise occurrence as it predicted the smallest area occurrence in comparison to the other techniques (Figure 22d). This 'under-prediction' may be a result of the inclusion of 'false' absences in model construction and in

this case, would not be the appropriate technique for an environmental risk assessment, where *all* areas of potential occurrence would be desired.

Different modelling techniques have different limitations and no model is perfect. Interestingly, a composite map combining the predictions of all of the modelling techniques provided a predictive distribution map that intuitively appears to be the best representation of actual harbour porpoise distribution based on previous studies in the area (Shrimpton & Parsons 2000; Evans, 1997b; Pollock *et al*. 2000) (Figure 23). Combining the predictions of different techniques may provide a clear understanding of actual distribution as the limitations of any one technique can be compensated for by the strengths of another model.

Therefore, perhaps the best and non-discriminate way to model species occurrence is not to use a single technique but instead a combination of techniques to maximise predictive accuracy.

Glossary

Training dataset		Data points used to construct the model
Testing dataset		Data points used to evaluate the model's predictions
False absence		Uncertainty associated with a recorded absence
Confusion matrix (Fielding & Bell, 1997)		A summary of the relative proportions of prediction errors
True positive (correct classification)	*a*	Areas of known occurrence correctly predicted as present by model
True negative (correct classification)	*d*	Areas where the species has not been found that are classified by the model as absent
False positive	*b*	Areas of known occurrence predicted absent by the model
False negative	*c*	Areas where the species has not been found that are classified present by the model
Commission or 'overprediction' (presence-only models)	*(b)*	A measure of areas of absence incorrectly predicted present
Omission or 'underprediction'(presence-only models)	*(c)*	Areas of known distribution predicted absent by the model
Sensitivity	*a/(a+c)*	True positive fraction
Specificity	*d/(d+b)*	True negative fraction
1-sensitivity	*b/(b+d)*	False positive fraction
ROC plot		Threshold independent method of assessing model accuracy
AUC (Area Under Curve)		Index derived from ROC curve measuring overall model accuracy

Acknowledgements

I would like to thank my supervisor Colin MacLeod for all his precious time, and advice and also for the supply of technical resources. I would also like to thank Dr. Graham Pierce for his expert advice.

I express my gratitude to Philip Preston, the Deputy Technical Director of Caledonian Macbrayne who so kindly gave me permission carry out surveys from the bridge and to the captains and crew of the MV Lord of the Isles, MV Clansman and MV Loch Nevis, who welcomed me on board.

Funding for this Masters course was provided by the Natural Environment Research Council (NERC) and this research project was part-funded by DSTL.

Finally thank you to my family for their endless support and encouragement in life.

References

Altman, D.G., Lausen, B., Sauerbrei, W. & Schumacher,M. 1994. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* **86**: 829-35

Anderson, R.P., Law, D., 2003. Peterson, A.T. Evaluating the predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling* 162: 211-232.

Bannon, S.M., Factors influencing the occurrence of cetaceans in the Minch, Scotland. May to July (2002). MRes Thesis. University of Aberdeen, U.K

Baumgartner, M.F., 1997. The distribution of Risso's dolphins (*Grampus griseus*) in relation to the physiography of the northern Gulf of Mexico. *Marine Mammal Science* **13**(4) 614-638

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

Chambers, J.M. and Hastie, T.J. 1997. Statistical models in Science. – Chapman and Hall.

Davis, R.W., Fargion, G.S., May, N., Leming, T.D., Baumgartner, M., Evans, W.E., Hansen, L.J., Mullin, K., 1998. Physical habitat of cetacean along the the continental slope in the north-central and western Gulf of Mexico. *Marine Mammal Science* **14** (3) 490-507

Deleo, J.M. 1993. Receiver Operating Characteristic Laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In: *Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis*, pp.318-25. College Park, MD: IEEE Computer society Press

Ellet DJ. 1979. Some oceanographic features of Hebridean waters. *Proceedings of the Royal Society Edinburgh B* 77:61-74

Evans, P.G.H. (1997a). Whales, dolphins and porpoises. In: *Coasts and Seas of the United Kingdom. Region 14 South-west Scotland:Ballantrae to Mull.* (Ed. by J.H. Barnes, C.F.

Robson, S.S. Kaznwska, J.P. Doody, N.C.Davidson & A.L. Buck), pp. 140-143. Joint Nature Conservation Committee, Peterborough,UK.

Evans, P.G.H. (1997b). Whales, dolphins and porpoises. In: *Coasts and Seas of the United Kingdom. Region 15 & 16 North-west Scotland: The Western Isles and West Scotland.* (Ed. By J.H. Barnes, C.F. Robson, S.S. Kaznwska, J.P. Doody, N.C. Davidson and A.L. Buck), pp.162-166. Joint Nature Conservation Committee, Peterborough.

Fielding, A.H., & Bell, J.F.1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**(1): 38-49

Gil de Sola, L., 1993. Las pesquerias demersales del mar de Alboran (submediterraneo iberico), evolucion en los ultimos decenios. *Informe Tecnico del Instituto Espagnol de Oceanografia* **142**, 179pp

Guisan, A. & Zimmerman, N.E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* **135** 147-186

Hirzel, A.H., Helfer, V., Metral, F. 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling* **145** 111-121

Hirzel, A.H. *et al*. 2002b. Ecological Niche-factor analysis: How to compute habitat suitability maps without absence data? *Ecology* **83**: 2027-2036

Hutchinson, G.E. 1957. Concluding remarks. Cold spring Harbour Symposium on Quantitative Biology **22**: 415-427.

Jeewoonarain T, Parsons ECM, Evans PGH. 2000. Operation sightings: sightings of cetaceans in the southern Hebrides, Scotland. In *European Research on Cetaceans.* 13:237-41.

Lerczak, J.A. and Hobbs, R.C. 1998. Calculating sightings distances from angular readings during shipboard, aerial and shore-based marine mammal surveys. *Marine Mammal Science* **14** 590-599.

Mendes, S., Swift, R., Hastie, G., MacLeod, C. 2004. In: *18th Conference of the European Cetacean society, 28th-31st August, Kolmarden, Sweden, 2004.*

Ortega-Huerta, M. and Peterson, A.T. 2004. Modelling spatial patterns of biodiversity for conservation prioritisation in North-eastern Mexico. *Diversity and Distributions* **10** 39-54

Palka D. 1996. Effects of Beaufort sea state on the sightability of harbour porpoises in the Gulf of Maine. *Report of the International Whaling Commission* 46:575-82

Payne PM, Nicolas JR, O'Brien L, Powers KD. 1986. The distribution of the humpback whale, *Megaptera novaeangliae,* on Georges Bank and in the Gulf of Maine in relation to densities of the sand eel, *Ammodytes americanus. Fishery Bulletin* 82(2):271-7

Pollock CM, Mavor R, Weir CR, Reid A, White RW, Tasker ML, Webb A, Reid JB. 2000. *The distribution of seabirds and marine mammals in the Atlantic Frontier, north and west of Scotland.* Seabirds and Cetaceans Branch, Joint Nature Conservation Committee, Aberdeen, Scotland. 92 pp.

Robertson, M.P., Caithness, N. & Villet, M.H. 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions* **7** 15-27

Rubin, J.P., 1997. La influencia de los procesos fiscio-quimicos y biologicos en la composicion y distribucion del ictioplacton estival en el mar de Alboran y estrecho de Gibraltar. *Informe Tecnico del Instituto Espagnol de Oceanografia* **24**, 88pp

Schweder, C. 2003. The use of Principle Component Analysis (PCA)-based modelling for constructing predictive models of occurrence for large mobile marine mammals. MSc. Thesis. University of Aberdeen, U.K

Shrimpton JH, Parsons ECM. 2000. *Cetacean conservation in the Hebrides.* Hebridean Whale and Dolphin Trust, Isle of Mull, Scotland, UK. 99 pp.

Stockwell. D., and Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science.* 13, 143-158

Townsend CH. 1935. The distribution of certain whales as shown by logbook records of American whaleships. *Zoologica* XIX(1):1-50

Uda M. 1954. Studies of the relation between the whaling grounds and the hydrographical conditions (I). *Scientific Report of the Whales Research Institute* 9:179-87

Williams, A.K. (2003). The influence of probability of detection when modelling species occurrence using GIS and survey data. PhD thesis, Blacksburg University, Blacksburg (USA).

Zaniewski, A.E., Lehman, A. & Overton, J.M. 2002. Predicting species spatial distributions using presence-only data: a case study of the New Zealand ferns. *Ecological Modelling* **157** 261-280.

Zwieg, M.H. & Campbell, G. 1993. Receiver Operating characteristic (ROC) plots: a fundamentl evaluation tool in clinical medicine. *Clinical Chemistry* **39** 561-77

Appendix I

The methodology for constructing the PCA models follows that of Robertson et al. (2001) A summary of the steps involved is described below:

1) All ***presence*** cells in the 'training' set were selected for constructing the model. Values for the environmental variables of the ***presence*** cells were standardised by subtracting the mean and dividing by the standard deviations of each variable. This removes the effects of differing units (**Matrix U**).
2) PCA was then performed using different combinations of variables.
3) Means and standard deviations used in step 1 were used again, but this time to standardise the values of all remaining grid cells in the dataset i.e. the ***absence*** cells (**Matrix W**).
4) The principal components explaining a minimum of 90% of total variance were used for model construction (Schweder, 2003).
5) These values were then multiplied by the component loadings obtained from the PCA to produce a matrix of component scores for all map localities in the model (**V x W)**.
6) In order to standardise the variance of each component axis, scores of each component were divided by their respective eigenvalues to produce a matrix of standardised component scores (**Matrix Z**).
7) The probability associated with each observation was obtained by summing the squares of the standardised component scores.
8) This value was substituted into the chi-square probability distribution function
9) The probability values for each grid cell were mapped back into Arcview® to their associated original geographical co-ordinates in IDRISI format.