CMEE MiniProject:

You are what you eat; Model optimisation for predicting predator size from prey size and associated abiotic factors, in global marine systems.

Katherine Bickerton, *Imperial College, London*

Word Count = [2591]

You are what you eat; Model optimisation for predicting predator size from prey size and associated abiotic factors, in global marine systems.

Abstract

Marine systems are one of the most difficult to monitor on the planet and one of the most at risk. Those at higher trophic levels are often the most at risk due to dependence on the stability of the rest of the food web. In this study, I aim to build several different types of model for predicting predator size, based upon prey size, feeding interaction, habitat, temperature and depth. The optimal model found was a generalised additive model (GAM), with prey mass, temperature and depth as explanatory variables. Feeding interaction was also found to have an effect, but violated the assumptions of GAM, therefore could not be included in the optimal model. These findings implies that predator size could be predicted from prey size and abiotic factors, which are less labour intensive to collect and could improve efficiency of conservation research.

Introduction

With ever increasing risks to biodiversity (**?**), and the close proximity of a sixth mass extinction extinction event (**?**), understanding the dynamics of ecological systems is increasingly important. One of the main threats to biodiversity is overexploitation, where many populations have been reduced to an unsustainable level. This is especially true of marine systems where, despite regulations, approximately 63% of fish stocks are below a the threshold population size required to persist (**?**).

Population data is particularly difficult to collect in marine systems, especially on pelagic organisms in deep oceans, often restricting data collection to shallow coastal areas. One means of supplementing information in data poor systems is to construct predictive models from existing data and extrapolate with a variety of parameters (**?**) This method can also aid in targeting studies to test likely models and aid in mitigation against overexploitation.

One classical model within ecosystem dynamics is that between predators and prey (**?**). Information regarding predator and prey physiology and consumption is some of the most easily available in marine systems. The most common method is stomach contents analysis of organisms caught as by-catch or stranded (**?**). This allows understanding of the position of each species within the food chain, and which prey sources are most important to or most threatened by specific predators.

This study uses an existing dataset of marine predators, both in coastal and pelagic zones to construct a model that could be used to predict predator size based upon available prey. The main aim is to investigate the relationship between predator mass and prey mass, on a global scale, across a range of habitats, temperatures and depths. As opposed to defining a typical ”null hypothesis”, this study is considering several different models for predator mass, based on the above factors, and will select and interpret the optimal model for the system.

Methods

Data Compilation

The data used for this was complied by Barnes et al, 2008, and comprises 19,625 records of 57 different marine predator species, from 27 locations globally, and 18 studies (**?**). The dataset included predator species, masses and lengths of predators and prey, habitat, location, type of feeding interaction, depth, mean annual air temperature and mean annual precipitation, for each predation record.

As this study aimed to examine relationship between predator and prey sizes, I chose one dimension, mass, as the measure of size, and excluded length as the two variables would violate the assumption of independence. The mean average of predator and prey masses, temperature and depth for each predator species were calculated. For the categorical variables, habitat and feeding interaction, the most common for each predator species was selected. Habitats were sorted into coastal and pelagic, dependent on the definitions of the habitats mentioned in each study. Feeding interactions were classified into either piscivorous (feeding only on fish), predacious (generalist predators, though diet may include fish), and planktivorous (feeding only on plankton), as defined by Barnes et al, 2008. A new data frame was compiled containing the averages calculated for each species, then used in model fitting.

Model Building and Fitting

I chose to compare three different type of model: a linear regression model, a generalised additive model (GAM) and a generalised linear mixed effects model (GLMM). As the response variable, predator mass, did not have a linear distribution, it was log transformed for use in the linear models, and as the response variable must be consistent to make the models comparable during selection, predator mass was also log transformed for the GAM.

When building the models, I started with the assumption that predator mass would be a function of the average mass of prey species. Additional explanatory variables were also considered and selected for in each model, dependent on whether they improved model fit, described below. The additional variables considered were: feeding interaction, habitat, depth and temperature. For each model, combinations of the explanatory variable were tested and the Akaike Information Criterion (AIC) calculated for each model, then compared to give the best fit to the data. The optimal models from each type of model were then compared, again using AIC, to find the model that overall best described the data. The linear regression model comparison was carried out using the ”step” function in R, however no equivalent function was available for GAMs and GLMMs, so comparisons were carried out manually using anova tests and AIC to systematically reject variables which decreased the fit of the model.

Model 1: Linear Regression Model - The linear regression model for predator mass was defined as below, where m = mass in grams and ϵ is the error not explained by the explanatory variable.

$$m_{pred} \sim m_{prey} + \epsilon$$

Linear regression models assume that the data is linear, therefore predator mass, prey mass and mean depth were log transformed. It also assumes collinearity and independence (**?**), therefore any explanatory variables that could interact must be removed. Due to this, habitat and feeding interaction were removed from the model as they correlated with depth and prey mass respectively. Finally, normally distributed variables are required, which occurred when the data was log transformed.

Model 2: GAM - The general additive model uses a smoothing function on each explanatory variable, to map each individuals fit to the response variable, and uses a Gaussian distribution as the variables used are continuous. The equation takes the following form, where m = mass, f denotes the smoothing function, x$_n$ denote the explanatory variable and ϵ is error not explained by the variables.

$$m_{pred} \sim f(x_1) + f(x_2) + ... + f(x_n) + \epsilon$$

The assumptions for GAM also require independence therefore habitat and feeding interaction were also excluded from the GAM models. Log transforms were not required for the model, as GAMs do not require linear variables (**?**), however log transformed models were compared as they were used in the previous model.

Model 3: GLMM - Mixed models give the ability to account for non-independence of variables by adding them as random effects, meaning they effect the data but not predictably (**?**). The following equation gives the general model for the GLMM, where m = mass, subscripts indicate fixed and random variables, d = mean depth, h = habitat, i = feeding interaction and ϵ accounts for any other error.

$$m_{pred} \sim m_{prey-fixed} + d_{fixed} + h_{random} + i_{random} + \epsilon$$

As a type of linear model, prey mass and depth were both log transformed to give a linear distribution. Additionally collinearity and independence are also assumptions of GLMMs unless the non-independent factors are accounted for as random effects, allowing the inclusion of habitat and feeding interactions in this model.

Computing Languages

Python version 3.5.2 (**?**) was used to manipulate the raw data into the data frame used for model building, as the pandas package (**?**) is fast and efficient at manipulating and building data frames and csv files.

R version 3.4.4 (**?**) was used for model fitting, selection and plotting models. R was most appropriate for this due to the wealth of packages available for model fitting and plotting. The package mgcv (**?**) fits GAM models and allows plots equivalent to the residual plots that can be produced for linear models. The lme4 package (**?**) fits linear mixed effects models and gives information about the significance of each different factor. R also has an inbuilt function to calculate the Akaike Information Criterion which was used to select the optimal model. Finally, the ggplot2 (**?**) package was used to generate plots used in this report (specified in figure legend when used).

Shell scripts in bash were used to compile the LaTeX document into a pdf format with references from the associated BibTeX file and to run the final project, as bash has inbuilt commands to run R and Python script files.

Results

Model 1: Linear Regression Model

Initially the linear regression model tested contained the three independent explanatory variables: prey mass, depth and temperature. To fulfil the assumption of linearity, prey mass and depth were both log transformed (note that predator mass is log transformed for all models), and a comparison using AIC was made between the fit of the model with and without log transforms (AIC$_{notlogged}$ = 343.8, AIC$_{logged}$ = 236.1). The smaller AIC value of the log transformed model indicated that it better fitted the data and the large difference between the values shows a significantly different fit.

Each variable in the model was then tested using the step function in R, which removed the least significant variable in the model then refitted it to the data until the optimal model was found. This model only contained one significant explanatory variable, prey mass (linear regression: R^2 = 0.90, F$_{1,55}$ = 496.3, p $<$ 0.001). The R^2 value for this model was high and positive meaning 90% of the variation in predator mass could be explained by prey mass. Figure 1a indicates the fit of the regression model (shown by a black line) to the data. The coefficients of the regression line were extracted from the model to give an overall equation of:

$$log(m_{pred}) = 5.41 + 0.761 log(m_{prey})$$

Where m = mass in grams, therefore when predator mass increases with prey mass.

../Results/lin_mod_feeding.pdf

../Results/density_feeding.pdf

(a)

(b)

Figure 1: (a) Mean average mass of 57 predator species as a function of mean average mass of prey, both log transformed. The black line represents the linear regression model: $log(m_{pred}) = 5.41 + 0.761 log(m_{prey})$, colours represent the three feeding interactions, piscivorous (red,n=19), predacious (blue, n=31) and planktivorous (green, n=7). (b) Density plot of log transformed mean average predator masses by feeding interaction, colours same as (a).

Although feeding interaction was not accounted for in the linear regression model, due to possible correlation with prey size, I split Figure 1a by feeding interaction and there appears to be strong grouping, with larger predators tending to be piscivorous and smaller to be planktivorous, with generalised predation more broadly spread. This relationship is further explored in Figure 1b, a density plot of each feeding interaction against predator mass, which reiterates this relationship. Planktivorous species also appear to be confined to a small range of predator masses whereas piscivorous cover a broader range of predator masses. This is explained further in the linear mixed model section.

Model 2: Generalised Additive Model (GAM)

The initial GAM, as with the linear regression model, included all independent explanatory variables, however there is no equivalent of the step function available for GAMs so models fits were tested manually. In each step, the least significant factor was removed, and model refitted, then compared to the previous model. Log transforms were also tested in the same way. The final model with the best fit to the data contained all three explanatory variables, but only prey mass was log transformed:

$$log(m_{pred}) \sim f(log(m_{prey})) + f(temp) + f(depth) + \epsilon$$

Where m = mass in gramms, ϵ indicates error not accounted for by the explanatory variables, and all variables are averaged. Figure 2 shows the relationship between each of the smoothed variables and predator mass. Figure2top shows predator mass increased with prey mass, as seen in the linear regression model. Figure 2middle shows a more complex relationship with the smallest predators found at lower temperatures, then medium to larger predators varying within a similar range of temperatures. Figure 2bottom shows another linear relationship where larger predators are found at higher depths, whereas small predators tend to stay in shallower areas.

../Results/gam_mod_plot.pdf

Figure 2: Smoothed functions of the explanatory variables used in the GAM model, where dotted lines give standard error and average predator mass is log transformed, n=57. Top: predator mass increases linearly with log transformed prey mass, middle: temperature fluctuates with predator mass change, indicating no clear relationship, bottom: predator mass increase linearly with depth.

Model 3: Linear Mixed Effects Model (GLMM)

As with the first two models, I started the GLMM model with the maximum explanatory variables, including habitat and feeding interaction as random effects which allows for their lack of independence. I then compared the models manually, excluding the least significant factor and comparing the fit of the new models using analysis of variance (ANOVA) as GLMMs are a type of linear model. The AIC values calculated gave the best fitting model as:

$$log(m_{pred}) \sim log(m_{prey}) + feeding_{random} + \epsilon$$

Where m = mass in grams, feeding interaction is included as a random effect and epsilon indicates variance not explained by the other variables. This model produced a very similar relationship to the linear regression model, however also explains the grouped feeding interactions observed in Figure 1.

Model Selection

Once optimal models were chosen for each type of model, the AIC values were compared, with the linear regression model giving an AIC = 232.5, GAM AIC = 200.4 and GLMM AIC = 239.0, therefore the GAM has the best fit. Additionally, Figure 3 shows the residual plots for each of the final models, and GAM has the narrowest range of residuals and most even spread compared to the other models, further supporting that it is the most appropriate model for the data.

../Results/lin_mod_resid.pdf

(a)

../Results/gam_mod_resid.pdf

(b)

../Results/glmm_mod_resid.pdf

Discussion

This study aimed to find an optimum model with which to predict predator mass, based upon a range of potential factors. Three types of model were tested: linear regression, generalised additive (GAM) and generalised linear mixed effects (GLMM), and the optimal model was found using GAM method. This model indicated that predator mass was a function of average prey mass, annual air temperature at the location and average depth at which the prey species was found. Predator mass (averaged and log tranformed) had a positive linear relationship with prey mass, therefore as prey mass increased, so did predator mass. Additionally, predator mass also increased with depth and varied with temperature fluctuations.

Many marine predators, especially those of larger mass, tend to have large home ranges and encounter a wide variety of prey (**?**). This makes them particularly difficult to monitor, whereas smaller species are often more localised, therefore easily monitored. The model generated could be useful in locations where potential prey species are present, as expected size of predator species could be predicted. This could also highlight those areas where predators are absent despite prey availability. Furthermore, species at higher trophic levels tend to be more vulnerable to disturbance as their persistence depends on an entire chain of species, as opposed to one producer (**?**), so having a means to predict their presence, or lack of, could aid monitoring groups globally.

In the optimal GLMM, the factor of feeding interaction was present, and relatively distinct clusters for piscivorous and planktivorous species were shown in 1. However, this variable was not included in the final model as it was unlikely to be independent of prey mass, and GAMs do not allow for random effects. Despite this, it is likely that feeding interaction does effect predator mass. This could be accounted for by generalised additive mixed effect models (**?**), which combine the inclusion of random effects with a GAM, and would be the next logical step from this study.

Another important assumption of this study, was that density of predators or prey had no affect on predator mass or was constant. This may oversimplify the system as, for example, many whale species feed on plankton however the density of plankton allow a much higher body mass and many piscivorous fish species have been observed increasing predation when only smaller prey species are available (**?**). This contradicts the findings in this study that indicated planktivorous predators had some of the lowest body masses. Therefore, another area for further study would be to include a factor for density into the model.

Conclusion

Overall, this study showed that the assumptions of models can have a marked effect on which explanatory variables are selected during optimisation. For marine predator prey systems, GAM were found to have the best fit to the data, and the function of mixed effects models to include random effects also showed interactions that may have otherwise been missed. To further this study, models combining these two types of model are advised, especially if used in predictions for conservation.