Copula-based species distribution modeling of highly migratory species☆

Craig Marsh[a,b], Nokuthaba Sibanda[b,*], Matthew Dunn[b], Alistair Dunn[a]

[a] *National Institute of Water and Atmospheric Research, Private Bag 14901, Wellington 6021, New Zealand*
[b] *Victoria University of Wellington, P O Box 600, Wellington, New Zealand*

Abstract

This template helps you to create a properly formatted LaTeX manuscript.

Keywords: `elsarticle.cls`, LaTeX, Elsevier, template

2010 MSC: 00-01, 99-00

1. Introduction

A key question for ecologistis is understanding why certain species occur in particular geographical areas [1]. For the fisheries industry, this understanding is particularly crucial when considering highly migratory species. A basic premise in biogeography is that climate contributes significantly to species distributions in space and time [2]. In particular, environmental and biological factors such as sea surface temperature or amount of dissolved oxygen are known to influence fish migration. Understanding how these factors affect the spatial distribution of fish populations will lead to better understanding of the drivers of fish population dynamics and fishery performance.

Traditionally spatial variation has usually been ignored in tactical fish population models. Rather, fish populations have been modeled with an implicit assumption of being spatially homogeneous within a defined area. More recently, the importance of explicit inclusion of spatial and temporal structure in

*Corresponding author
Email address: `nokuthaba.sibanda@vuw.ac.nz` (Nokuthaba Sibanda)

population models has been recognised [3], for reasons such as; preventing local stock depletion [4], investigating spatial protected areas [5], investigating bias in model outcomes [6] and understanding spatial ecological processes [**?**]. This means spatially explicit theoretical frameworks for answering these queries need to be investigated and developed.

In fisheries research, abundance is typically measured using an abundance index: a ratio of catch amount (C) to effort expended (E). This index is referred to as Catch Per Unit Effort (CPUE). These data are sourced either from commercial fishery records or from records of government mandated observers who, by regulation, are present to observe and record information of commercial fishing activities. An issue that arises with species abundance data in fisheries research and in ecology in general, is the occurrence of zero catches or zero counts. For a given study, careful thought must be given to the source of zeros as that determines the structure of the model used for abundance indices, counts or presence/absence records. Zeros can be 'true zeros' (species absent) or 'false zeros' (species present but not detected due to sampling or observer errors in the course of data collection) [7]. An appropriate model must be used that accounts for either source of zeros to avoid bias in parameter estimates and their associated measures of uncertainty [8, 9]. We assumed that any zero catches found in our data were 'false' zeros. In other words, the record of a zero catch was not taken to mean that the species was absent at a given location. Rather that either some fish were caught but missed in the process of counting or weighing (observer error); or the fish were present at a location but were not caught (sampling error). Given the temporal and spatial scale used to define a cell (1° by 3 months approx) and the limited number of sampling events (2-100) this assumption seemed plausible. This meant the interpretation species status is of detection rather than presence. We also make the assumption that the species is present at all locations where effort was expended. This is a reasonable assumption to make since, once a fishery is established, fishing vessels target locations where species are known or expected to be present, and zero catches were included for analysis only if $E > 0$.

In this article we model the spatio-temporal distribution of highly migratory fish species on the premise that these species migrate in response to environmental and climatic conditions. The core problem is to use information about the presence and abundance of a species in an area and about relevant environmental factors to build a model for predicting how likely the species is to be present or absent, and if present, what the predicted abundance would be in unsampled locations and also for predicted climate changes in the sampled locations. We propose the use of copulas to create a joint species response index to multiple to environmental and climatic attributes. The response index expresses the relative preference for the environmental and climatic attributes at a given location, relative to other locations. This is a novel development in the development of a preference index. A method that is currently used to create a preference index uses a geometric mean to calculate a combined index over multiple attributes [10]. The preference index is calculated for each area from some function, $f()$, of a number of variables that describe the current state of the location, such that at location k, the preference index is given by:

$$p_k = f(x_{1k}, x_{2k}, \ldots, x_{mk}),$$

for a set of m attributes. The index p_k represents the combined response of the species to the m attributes. The current approach of combining the individual responses calculates a geometric mean:

$$p_k = \prod_{l=1}^{m} (p_{lk})^{\frac{1}{m}} ,$$

which is equivalent to an arithmetic mean on the log-scale. The implicit assumption in the current approach is that on the log scale, each attribute has an additive effect on the the species' response to environmental and climatic conditions. This ignores any correlation that may be present in how species respond to multiple attributes. Our proposed approach of using copulas to construct a joint response index will take into account any existing correlations

in species' response to multiple attributes []. NEED REFERENCE FOR JOINT RESPONSE INDEX. We expect the use of copulas to result in improved model fit and precision in predictions (???NEED DIAGNOSTICS THAT REFLECT THIS). Our developments are implemented and tested in Spatial Population Modelling (SPM)[**?**][21], a software tool that currently implements the existing approach that uses the geometric mean.

In Section 2 we present the general format of the population model used in SPM, and present development of a copula-based joint preference index. Section 3 presents the data used and associated summary statistics. In Section 4 we present the results of applying the copula-based index and results of model comparisons. The Discussion is in Section 5.

2. Model development and evaluation

2.1. Model for CPUE

A model was developed relating CPUE, Y, to multiple environmental and climatic attributes combined in the form of a preference index. A log-Normal distribution was assumed for $Y > 0$, such that at location k defined by location and time, given $Y_k > 0$ we assume:

$$Y_k = qN_k e^{\epsilon_k},$$

for $\epsilon_k \sim N(0, \sigma^2)$, where q is an unknown catchability constant and N_k is the absolute abundance. To incorporate the 'false' zero-catch observations, we used a delta-logNormal [11, 7] approach, in which the probability of a non-zero catch (an 'encounter') is modelled separately from the catch rate for each encounter.

The variable Y therefore has the following density function:

$$\begin{aligned} f(y|y > 0) &= p \cdot g(y) \quad \text{for } y > 0, \\ Pr(Y = 0) &= 1 - p \qquad \text{with } 0 \leq p < 1 \end{aligned}$$

where p represents the probability of encounter and $g(y)$ represents the log-Normal density function given by

$$g(y) = \frac{1}{y\sigma_k\sqrt{2\pi}} \exp\left(\frac{log(y_k) - \mu_k}{2\sigma_k^2}\right).$$

Given a sample of data over k location and time grid cells, we write the likelihood function as follows:

$$f(y_1, \ldots, y_K) \quad = \quad \prod_{k=1}^{K} \left[g(y_k)\right]^{z_k} (p_k)^{z_k} (1-p_k)^{1-z_k}, \tag{1}$$

where

$$z_k = \begin{cases} 1 & \text{if } y_k > 0 \\ 0 & \text{if } y_k = 0. \end{cases}$$

For a log-Normal likelihood for the non-zero observations, this gives the following objective function (based on the negative log-likelihood):

$$\begin{aligned} -\log L &= -\sum_{k=1}^{K} z_k \log p_k + (1-z_k)\log(1-p_k) + z_k \left(\log\sigma_k + \frac{1}{2}\log(2\pi) + \frac{1}{2}\left(\frac{\log(y_k) - \log(qN_k)}{\sigma_k} + \frac{1}{2}\sigma_k\right)^2\right) \\ &= \sum_{k=1}^{K} z_k \log\left(\frac{1-p_k}{p_k}\right) - \log(1-p_k) - z_k \left(\log\sigma_k + \frac{1}{2}\log(2\pi) + \frac{1}{2}\left(\frac{\log(y_k) - \log(qN_k)}{\sigma_k} + \frac{1}{2}\sigma_k\right)^2\right) \\ &= \sum_{k=1}^{K} z_k \log\left(\frac{1-p_k}{p_k}\right) - \log(1-p_k) - z_k \left(\log\sigma_k + \frac{1}{2}\log(2\pi) + \frac{1}{2}\left(\frac{\log(y_k) - \log(qN_k)}{\sigma_k} + \frac{1}{2}\sigma_k\right)^2\right) \end{aligned} \tag{2}$$

Here p_k represents the probability of encounter for a given species at location k and therefore represents $E(Z_k)$. For any given location, p_k is influenced by a number of factors. First is the relative preference for that location as determined by the levels of environmental factors ie how favourable is the location for that species relative to other species. This is expressed in terms of the preference function. The second factor is the population density at that location ie the number of fish present relative to the area.

The first two factors are incorporated into the model using:

$$E(Z_k) = q \cdot p_k^f \cdot D_k,$$

where p_k^f is the relative preference for the cell determined from the population preference response curves for the environmental factors; D_k is the population density at location k.

Two datasets were used in this study for model fitting and evaluation. The first dataset was Catch and Effort data from a longline observer program. The second dataset is satellite derived oceanic conditions (preference attributes) that are assumed to accurately reflect the state of the ocean. The spatial extent of both data sets is around New Zealand's Exclusive Economic Zone (26°S to 49°S and 164°E to 175°W), and the temporal window was from 2002-2013. Spatial resolution was defined by the coarsest preference attribute, which was 1° latitude and longitude bins. The temporal resolution was defined by seasons (Summer, Autumn, Winter and Spring).For preference attribute data available at a finer resolution, the arithmetic mean of all values in a cell created by the coarser scale was used as the final input to the model.

CPUE data was also amalgamated to fit in with the spatial and temporal scale used for the preference attributes. The geometric mean was used to combine taken for cells that contained more than one observation [**Is this correct - this would give a zero mean whenever a zero observation occurs. Shouldn't we take the median or arithmetic mean?**].

2.2. Copula-based preference index

A copula is used to model the dependence structure of a set of random variables independently of their marginal distribution functions. Sklar's theorem states that for any two continuous random variables, X_1 and X_2, there exists a unique bivariate copula C such that

$$F(x_1, x_2) = C(F_{X_1}(x_1), F_{X_2}(x_2)),$$

which we use to couple the marginal distribution functions $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ to give their bivariate distribution function $F(x_1, x_2)$. The copula family used is chosen to express an assumed dependence structure between X_1 and X_2. The preference index was constructed by assigning a marginal preference function to each attribute and then combining the marginal preference function into a combined preference function by applying an appropriate copula. The joint preference index was therefore determined from the resulting joint distribution.

copula families and dependence structures

2.3. Model evaluation

The models were evaluated using the Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{n_k}(\hat{y_k} - y_k)^2}$$

and the AIC

$$AIC = 2(-\log L + K).$$

3. Data

- CPUE summary for different species (CM)
- environmental attributes and resolution (CM)

4. Results

Analysis results (CM, NS)

5. Discussion

Appendix

5.1. Preference functions

This section considers marginal densities and corresponding cumulative distribution functions used as preference functions for the environmental variables.

The density functions represent the measure of preference for a given value of the environmental variable relative to all others. The shape of the density function is key to an accurate expression of the preferences. This means that in some cases, only a portion of the density curve is relevant for expressing the relative preference, and effectively, this is equivalent to using a truncated density function.

The following density functions are considered:

Exponential distribution

The Exponential density function is given by:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

for $\lambda > 0$, with truncated density function:

$$g(x;\lambda,a,b) = \begin{cases} \frac{\lambda e^{-\lambda x}}{e^{-\lambda a} - e^{-\lambda b}} & a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

The cdf is given by $G(X) = 1 - e^{-\lambda x}$ and the truncated cdf, assuming $a < x < b$, is given by:

$$G(x;\lambda,a,b) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{e^{-\lambda a} - e^{-\lambda x}}{e^{-\lambda a} - e^{-\lambda b}} & \text{for } a < x < b \\ 1 & \text{for } x \geq b \end{cases}$$

Normal distribution

The Normal distribution is given by:

$$f(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2} \exp -\frac{1}{2\sigma^2}(x-\mu)^2, \quad -\infty < x < \infty$$

with a truncated density given by:

$$g(x;\mu,\sigma^2,a,b) = \begin{cases} \frac{(2\pi\sigma^2)^{-1/2}\exp -\frac{1}{2\sigma^2}(x-\mu)^2}{\Phi(b;\mu,\sigma^2)-\Phi(a;\mu,\sigma^2)} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

The parameters μ and σ^2 are the mean and variance of the parent Normal distribution, respectively, and $\Phi(x;\mu,\sigma^2) = P(X \le x|\mu,\sigma^2)$ is the cumulative distribution function of the parent Normal distribution evaluated at x.

The truncated cdf, assuming $a < x < b$, is given by:

$$G(x;\mu,\sigma^2,a,b) = \begin{cases} 0 & \text{for } x \le a \\ \frac{\Phi(x;\mu,\sigma^2)-\Phi(a;\mu,\sigma^2)}{\Phi(b;\mu,\sigma^2)-\Phi(a;\mu,\sigma^2)} & \text{for } a < x < b \\ 1 & \text{for } x \ge b \end{cases}$$

Double-Normal distribution

The density function of a double normal random variable is given by:

$$f(x;\mu,\sigma_1^2,\sigma_2^2) = \begin{cases} A\exp\left[-\frac{(x-\mu)^2}{2\sigma_1^2}\right] & x \le \mu \\ A\exp\left[-\frac{(x-\mu)^2}{2\sigma_2^2}\right] & x \ge \mu \end{cases}$$

where $A = \left(\sqrt{2\pi}(\sigma_1+\sigma_2)/2\right)^{-1}$. The distribution is formed by taking the left half of a normal distribution with parameters (μ,σ_1^2) and the right half of a normal distribution with parameters (μ,σ_1^2), and scaling them to give the common value $f(\mu) = A$ at the mode μ.

References