--- Chapter One ---

Introduction

The aim of this workbook is to help marine biologists familiarise themselves with using GIS in their research. To do this, it uses a Task Oriented Learning (TOL) approach, first introduced in *An Introduction To Using GIS In Marine Biology*, to provide five exercises based around species distribution modelling (SDM), an increasingly important field in marine biology. As such, it does not represent a stand alone GIS book and is meant to act as a companion to the original book rather than to replace it in any way. It does not provide any background information on using GIS as this has already been covered within *An Introduction To Using GIS In Marine Biology* itself. Instead, it simply provides instructions on how to do the exercises themselves.

Thus, this workbook is primarily aimed at those who have read some or all of *An Introduction To Using GIS In Marine Biology*. If you have not already done so, it is recommended that at a minimum you read chapters seven (*Translating biological tasks into the language of GIS*), eleven (*How to use the 'How To...' sections of this book*) and twenty (*How to combine instruction sets for basic tasks to create instruction sets for more complex tasks*) of *An Introduction To Using GIS In Marine Biology* before working through any of these exercises. It will also help if you are familiar with the basics of GIS (chapter two), common concepts and terms in GIS (chapter three), the importance of projections, coordinate systems and datums (chapter four), types of GIS data layers (chapter five), starting a GIS project (chapter six) and how to set up a GIS project (chapter thirteen). Finally, it is worth at least flicking through chapters thirteen to nineteen to familiarise yourself with how instruction sets are laid out using the TOL approach introduced in *An Introduction To Using GIS In Marine Biology*.

This supplementary workbook uses ArcGIS® 10.2 GIS software and R statistical software to illustrate how these tasks should be done. However, similar processes are likely to be used to achieve similar outcomes in other GIS and statistical software packages.

The exercises provided in this book are designed to be worked through in a sequential manner. This is because the same data set is used throughout most of these exercises and you will need to use some of the data layers generated in earlier exercises for later ones. In addition, each exercise represents a discrete, but important, step in the process of implementing a species distribution modelling (SDM) project within a GIS framework, and they are provided in the same order which you would need to undertake them if you were conducting your own SDM project. For example, the first exercise covers issues associated with processing your raw survey data in such a way that it creates the appropriate information on species distribution required to conduct an SDM, while later exercises cover the processing of environmental variables, creating your model in a statistical software package, visualising the predicted spatial distribution from your model and validating its predictive ability using an independent data set. The data sets used in each exercise can be downloaded from www.gisinecology.com/books/marinebiologysupplementaryworkbook.

The exercises are provided using the same flow diagram based format introduced in the '*How To...*' reference guide section of *An Introduction To Using GIS In Marine Biology*, and specifically in chapter twenty which outlines how to combine individual instruction sets to work out how to do more complex tasks. This means that for each exercise, you will first find an outline of what will be achieved by the end of it, why it is useful for marine biologists to be able to do this and what data layers you will need to start with. You will then find a summary flow diagram which will detail the order which individual instruction sets for basic tasks must be done. Finally, you will find a set of numbered instruction sets based on those provided in *An Introduction To Using GIS In Marine Biology*. These have been customised to make them specific to the data set used for each example. To complete a specific exercise, you will need to work through each of these instruction sets in the order given in the summary flow diagram. In order to allow you to know whether you are progressing correctly, figures will be provided at regular intervals which will show you what the contents of the MAP window, LAYOUT window, TABLE OF CONTENTS window and/or TABLE window should look like at that specific stage.

Before starting these exercises, it is worth taking the time to understand the background behind species distribution modelling. This will not be covered in any sort of detail here, but, briefly, a SDM aims to predict where a species is likely to occur based on the relationship between environmental variables and its spatio-temporal distribution. In this respect, SDM is closely related to the ecological niche concept. However, while the theory behind SDM is straight-forward and easy to understand, applying an SDM to a real world situation in a biologically meaningful way can be exceedingly difficult. This is because it can be difficult to choose which environmental variables should be used to model the distribution of a particular species, what resolution these variables should be sampled at, what aspect of a species distribution should be modelled (such as presence-absence or abundance), what modelling techniques should be used and how the predictive ability of the model should be validated.

This supplementary workbook does not aim to provide advice about how to deal with any of these issues. Instead, it aims to provide exercises which will allow you to integrate a species distribution modelling project into a GIS environment once these issues have been successfully resolved. It will do this by taking you through a specific SDM project, starting with the raw survey data and ending with validating the predictive ability of the SDM using an independent dataset. Thus, it provides a framework for the practical implementation of a SDM project rather than providing advice on how to deal with all the tricky issues associated with selecting the appropriate environmental data, identifying the resolution they should be sampled at, selecting an appropriate dependent variable to answer a specific research question or choosing an appropriate statistical technique.

While it is relatively easy to do from a technical point of view, especially following the framework provided in this supplementary workbook, implementing an SDM project in a biological meaningful way is not necessarily straight forward and it relies on a number of important assumptions which are easy to overlook. The first of these is that the data which are used to build the model are representative of the true relationship between a species distribution and the environmental variables used within the model. This is known as representativeness. If your data are not representative, it means that the results of your modelling process cannot be applied beyond the locations which you sampled when collecting it. This is not in itself a bad thing and does not mean that you cannot improve your understanding of the mechanisms which drive species distribution from such data. It does, however, mean that you cannot make any extrapolations beyond these locations (i.e. you should not undertake the type of visualisation outlined in exercise four). This means that if the aim of a specific study is to make any sort of prediction of the likelihood of occurrence of a species in areas beyond your sampled locations, you need to ensure that your data are representative of the wider area to which you wish to apply your model before you start your SDM project. In fact, in a ideal world, this would be done as part of your survey design process to ensure you collect sufficiently representative data. The trouble with this is that objectively assessing representativeness for a specific data set is not necessarily straight forward. However, one approach is to use a Habitat Representativeness Score (HRS) as outlined in MacLeod (2010). A PDF of this paper can be found in the compressed folder containing the data for the exercises in this book which you will download at the start of exercise one.

The second major assumption behind SDM which will be considered here is that the model is accurately capturing the niche that a particular species occupies in relation to the environmental variables which are included within the model. The main issue with this assumption is that it implicitly assumes that the niche occupied by a species is constant in time and space. That is, it assumes that data collected at one location and in one time period can be applied to all other locations and time periods, especially during any spatial visualisations of your model (see exercise four). However, in reality, the exact niche occupied by a species may vary from location to location, meaning that it may not be possible to create a single model that accurately captures the species niche across whole regions or oceans. This is because the niche occupied by a species can be influenced by a wide range of biotic and abiotic factors, such as variations in water temperature or the presence of competing species. For example, in the northwest Atlantic, white-beaked dolphin show a strong preference for water depths between about 10 and 150m, but only if common dolphin are not present. Where they are present, white-beaked dolphin appear to be excluded from the local area regardless of the availability of waters of suitable depths[1]. This means that an SDM for white-beaked dolphin created from data collected in areas where common dolphin are absent cannot be used to accurately predict the likelihood of white-beaked dolphin in areas where they are present because of the differences in the

[1] MacLeod et al. 2008. JMBA 88, 1193-1198.

relationships between occurrence and environmental variables, such as water depth, when this competing species is present.

Similarly, the niche occupied by a species may vary temporally. Within the marine environment, this may mean that species vary the exact niche they occupy depending on the time of day, the state of the tide, the phase of the moon and/or the time of the year. This is because all these variables can influence the distribution of productivity and food availability within the marine environment, and this, in turn, can influence the relationship between the distribution of a particular species and other environmental variables, such as water depth, seabed slope or the distance from the nearest coast. For example, in the Sea of Hebrides in western Scotland, harbour porpoises (the species which is the focus of the main exercises in this supplementary workbook) occur at higher densities as you move further from the coast in early summer, but nearest the coast in late summer.[2] This change in habitat preferences, and therefore the niche occupied, is probably driven by changes in availability of sandeels (a key prey species in early summer) and a switch to targeting different prey species as the summer progresses. This means that you cannot necessarily take an SDM of porpoise density for this region based on data collected from one time of year and apply it to another. This is quite obvious. However, somewhat less obviously, it also means than you cannot create an accurate model using data collected in both early and late summer as this will not accurately capture the niche occupied at either time of year. Worse, it can actually lead to the identification of preferences which do not represent the true niche occupied at either time of year.

These underlying assumptions mean that SDM projects should not be rushed into or undertaken lightly, especially if the aim is to produce visualisations or predictions beyond the locations sampled which are to be used to inform any conservation or management strategy or decision. Instead, before you create an SDM for such purposes, you need to ensure that you have both a detailed understanding of the ecology of your target species and a detailed knowledge of the environment to which you will apply your SDM so that you can identify where and when spatio-temporal variations in the niche occupied by the species may occur. In addition, it is also essential to understand the importance of the need

[2] Bannon S.M. 2013. The Influence of Environmental Variables on the Spatial and Temporal Occurrence of Cetaceans off the West Coast of Scotland. Ph.D. Thesis, University of Aberdeen, UK.

to conduct a proper and detailed validation of the predictive ability of your model (see exercise five) if it is to be used to make any visualisations or predictions of the species likely distribution beyond the locations which were sampled (see exercise four).
NOTE: While this supplementary workbook contains information on how to do one specific statistical test for one specific data set, it is not a statistical textbook. This means that the statistical test used in this book (a generalised additive model) should not be applied to any other data set without first ensuring that it is appropriate. In addition, even if it is appropriate, this does not mean that other statistical tests are not just as valid for use with a particular data set or for a particular project. If you wish to learn more about statistical tests and how to apply them in a correct and biological meaningful manner, you will need to consult one of the many good statistical textbooks which are available. In particular, I would recommend those written by Alain Zuur (see http://www.highstat.com/books.htm for details) .

NOTE: The instruction sets provided here are for training purposes only, and are only meant to be an aid to learning how to use GIS in marine biological research. While every effort has been made to ensure that these instructions are complete and error-free, they come with no guarantee of accuracy and, as with all technical books, some errors may have slipped through undetected. Whenever I become aware of any such issues, I will post corrections on www.GISinEcology.com/books/marinebiology/corrections rather than waiting to correct them in the next edition of this book. As a result, before you do any of the exercises in this book, you should check this page to see whether there are any corrections that should be applied. In addition, it is important to realise that there is no guarantee that these instructions will produce the desired outcome in every circumstance. As a result, if you are using the instruction sets provided here to learn how to do critical tasks, it is essential that you check (and then double check) that they work for your given circumstances rather than blindly following them without thinking. The author will not be responsible for any errors which occur because of the application of these instruction sets to real world situations.

NOTE: As with many things in GIS, there may be more than one way to do the exercises outlined in this book. The instructions presented here will work for the data set provided and for the exercises outlined in this book. They should also work in most other circumstances. However, if you find an alternative way to do them which works for your data, or if you have someone who can show you how to do them in another way, feel free to do them differently.