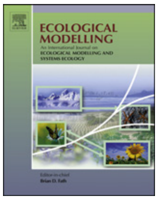
Ecological Modelling 221 (2010) 467–478

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



Predicting the distributions of marine organisms at the global scale

Jonathan Ready[a,*], Kristin Kaschner[b], Andy B. South[c], Paul D. Eastwood[c], Tony Rees[d], Josephine Rius[e], Eli Agbayani[e], Sven Kullander[f], Rainer Froese[g]

[a] Instituto de Estudos Costeiros, Universidade Federal do Pará – Campus de Bragança, Aldeia, Bragança 68600-000, Pará, Brazil
[b] Evolutionary Biology & Ecology Lab, Institute of Biology I (Zoology), Albert-Ludwigs-University Freiburg, Germany
[c] Centre for Environment, Fisheries, and Aquaculture Science (Cefas), Lowestoft Laboratory, Lowestoft, Suffolk, NR33 0HT, UK
[d] CSIRO Marine and Atmospheric Research, GPO Box 1538 Hobart TAS 7001, Australia
[e] WorldFish Center – Philippines Office, MCPO Box 2631, 0718 Makati City, Philippines
[f] Department of Vertebrate Zoology, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden
[g] Leibniz-Institute of Marine Sciences, Düsternbrooker Weg 20, D-24105 Kiel, Germany

ARTICLE INFO

Article history:
Received 20 March 2008
Received in revised form
15 September 2009
Accepted 16 October 2009
Available online 24 November 2009

Keywords:
Species distribution modelling
Range maps
Global marine biodiversity
Trawl surveys
Expert review
Model comparison

ABSTRACT

We present and evaluate AquaMaps, a presence-only species distribution modelling system that allows the incorporation of expert knowledge about habitat usage and was designed for maximum output of standardized species range maps at the global scale. In the marine environment there is a significant challenge to the production of range maps due to large biases in the amount and location of occurrence data for most species. AquaMaps is compared with traditional presence-only species distribution modelling methods to determine the quality of outputs under equivalently automated conditions. The effect of the inclusion of expert knowledge to AquaMaps is also investigated. Model outputs were tested internally, through data partitioning, and externally against independent survey data to determine the ability of models to predict presence versus absence. Models were also tested externally by assessing correlation with independent survey estimates of relative species abundance. AquaMaps outputs compare well to the existing methods tested, and inclusion of expert knowledge results in a general improvement in model outputs. The transparency, speed and adaptability of the AquaMaps system, as well as the existing online framework which allows expert review to compensate for sampling biases and thus improve model predictions are proposed as additional benefits for public and research use alike.

1. Introduction

Concerns over changing patterns of marine biodiversity resulting from climate change and human impacts have generated considerable interest in the use of models designed to generate spatial predictions (i.e. maps) of species' distributions from point occurrence data (Guisan and Thuiller, 2005). Ideally, prediction models would be generated from comprehensive species occurrence and absence data from targeted surveys. Unfortunately this level of data is only available for a relatively limited number of species and geographic locations, creating problems for assessments of changes in patterns of marine species distributions and biodiversity at regional and global scales.

As an alternative, modellers are making use of increasing volumes of presence-only data (Pearce and Boyce, 2006). These are published online through global databases such as FishBase (Froese and Pauly, 2007) and the Ocean Biogeographic Information System (OBIS, 2007), both of which feed data directly into the Global Biodiversity Information Facility (GBIF, 2007). These data frameworks compile species occurrence data from museum records and other sources. They therefore represent a highly patchy and biased view of patterns of species' distributions as a result of regional and local variations in sampling effort. The bias inherent to the data creates problems when data-driven modelling techniques are used to generate predictions of species' distributions. This is because an absence of occurrence records may not necessarily indicate a true absence in the distribution of the species, but rather a lack of adequate sampling. This is especially true for marine organisms, as inshore areas are more often sampled at a higher rate compared to offshore areas, causing a bias in the species–habitat relationship described by the data (Kaschner et al., 2006; MacLeod et al., 2008). In this scenario, an offshore species might well be predicted to have an inshore distribution if sampling had only occurred over a limited proportion of its overall depth range. Similarly, misidentification of species is a common weakness of all existing large online occurrence record deposi-

---

[*] Corresponding author. Tel.: +55 0 91 3425 1745; fax: +55 0 91 3425 1209.
E-mail address: jonathan.ready@gmail.com (J. Ready).

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.10.025

taries (Meier and Dikow, 2004), which in turn can lead to false predicted presences and unrealistic species distribution if these data sets are used as input for standard species distribution modelling.

Until better data sets are available, these biases in sampling effort can be best countered if model algorithms are able to incorporate expert information on species–habitat preferences. These represent a rich but currently underutilized resource. Here, we define expert information as habitat use information that is not directly available as raw data, i.e. published information about habitat use/preference that is based on quantitative investigations of species occurrence in relation to environmental knowledge. Examples include: evidence of a pelagic lifestyle, known depth ranges, latitudinal and longitudinal limits to ranges or physiological tolerances of species. Additionally, experts working on the taxa could include personal knowledge either about occurrence records not yet accessible through online data depositaries, or maximum range extents not described in the literature. This information could also be included should such experts review a map. However, as research into species distribution modelling has progressed, so has the complexity of model algorithms to the extent that users have little or no opportunity to influence the model outcome through the use of expert information. While the goal of species distribution modelling is to increase prediction accuracy (which might be expected to increase with model complexity), the use of increasingly sophisticated methods may also be a barrier to non-expert modellers such as biodiversity managers, decision makers, and planners. All of these people have a vested interest in the reliability of model outputs and therefore need to understand how the models were constructed. Simple and transparent numerical approaches combined with expert guidance on the form of the species–habitat relationships may therefore help circumvent some of the inherent problems in predicting regional and global distributions from patchy, heavily biased occurrence data from global biodiversity databases. If these same algorithms are transparent and produce reliable and verifiable results, the likelihood that predictions will have practical use and feed into decision making and planning will be further increased.

We describe such an approach, called AquaMaps (available for use via the webpage http://www.aquamaps.org, and based on the global distribution tool for marine mammals developed by Kaschner et al., 2006). It was developed for the mass-production of predicted distributional ranges of marine organisms from global occurrence databases, using simple and pre-defined numerical descriptions of species–habitat relationships that can be modified where needed. Predictions from AquaMaps for 12 selected marine fish and mammal species are compared alongside those generated from a range of other methods (GARP, GLM, GAM, MAXENT) that are commonly used to construct species distribution models but which are limited in the extent to which experts can influence model parameterisation. Model comparisons were made using independent data from fisheries trawl surveys conducted in UK and Australian waters and dedicated marine mammal surveys in Antarctic waters and in the North Sea.

The objective of the assessment was to compare the performance in terms of predictive accuracy of AquaMaps, a system that can be automated to a great extent and allows the speedy processing of large number of species, with a range of popular and generally more sophisticated routines. If, at the scale of entire species ranges, AquaMaps can produce similarly reliable and verifiable results as commonly used high-end methods, then its greater transparency, ability to incorporate expert knowledge and its online accessibility would facilitate the broad application of such an approach, increasing practical use in the context of decision making and planning processes.

2. Materials and methods

2.1. Marine species occurrence data

Global occurrence data for model building were obtained from two sources. For marine fish, occurrence records were extracted from FishBase, the most comprehensive, online database on fish occurrence records from museum collections and selected, regional trawl surveys (Froese and Pauly, 2007). Marine mammal occurrence records were obtained from OBIS (OBIS, 2007). Similar to FishBase, OBIS is a comprehensive, online database of occurrence data from national museum collections and other sources.

For the marine fish, the species selected represented a broad range of taxa and life histories and were species which were also relatively well represented in the two regions used for model testing, i.e. UK and Australian waters (Table 1). Nine fish species were selected: four that were adequately represented in fisheries surveys conducted in UK waters by the Centre for Environment, Fisheries, and Aquaculture Science (Cefas); four that were adequately represented in fisheries surveys conducted in Australian waters by the Commonwealth Scientific and Industrial Research Organisation (CSIRO); and one (John dory, *Zeus faber*) that was represented in both regions. Raw occurrence data (all accumulated occurrence data per species) from FishBase were extracted for these species. Records deriving from CSIRO surveys were removed, as this data would form the test data for validating the models (Cefas survey data, also used for testing, is not yet represented in FishBase or OBIS and so did not need removing). Occurrence records were spatially aggregated at a resolution of 0.5° latitude × 0.5° longitude and assigned a unique *c*-squares code (Rees, 2003). These could then be converted to a binary format that distinguishes between presence and absence in each cell as input for most subsequent analyses. The exception to this is the testing of predicted gradients of species occurrence with independent survey data where proportional data is used. *c*-squares is a global, spatial indexing system that allows geographic features to be referenced at multiple spatial resolutions, and provides the framework for the database structure behind AquaMaps. Using a fixed spatial resolution and indexing system facilitated the process of constructing and testing the models as data could easily be passed between the various programs containing the modelling routines (see below). Having assigned raw occurrence records to 0.5° c-squares cell, potentially erroneous cells were removed if they were: (i) located entirely over land; or (ii) located outside of UN Food and Agriculture Organisation (FAO) fisheries reporting areas where the species is known to occur; or (iii) located outside of expert defined geographic range extents (bounding boxes). FAO areas and bounding boxes were assigned to species using information on species distributions from the many references listed in FishBase (for fish) and those provided in Kaschner et al. (2006), Appendix 2 (for marine mammals). This process is automated in AquaMaps. Further cleaning of data to check for other errors in digitisation, misidentification or data corruption requires significant human input. As the ability of different modelling methods is to be assessed based on their capacity to deal with publicly available data with maximal automation to produce reasonable predictions, such further cleaning was not performed for training data. Test data from surveys are assumed to have minimal error as they came direct from the data source, though basic tests for error in digitalisation were performed. Certain types of error, such as misidentification, will remain in almost any dataset not prepared entirely by a taxonomic expert from original samples.

Three marine mammal species were selected for model comparison (Table 1). These species were chosen due to the contrasting geographic ranges they are known to occupy and the availability of sufficient occurrence data needed for model constructing and testing. Records were treated similarly to those for marine fish.