Ecological Modelling 323 (2016) 61–76

Contents lists available at ScienceDirect

Ecological Modelling

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

Estimating absence locations of marine species from data of scientific surveys in OBIS

Gianpaolo Coro[a,*], Chiara Magliozzi[a], Edward Vanden Berghe[b], Nicolas Bailly[c,d], Anton Ellenbroek[e], Pasquale Pagano[a]

[a] Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, via Moruzzi 1, 56124 Pisa, Italy
[b] Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Elsene, Belgium
[c] LifeWatchGreece, Hellenic Centre for Marine Research (HCMR), Gouves, 71500 Heraklion, Greece
[d] FishBase Information and Research Group (FIN), 4031 Los Baños, Laguna, Philippines
[e] Food and Agriculture Organization of the United Nations (FAO), Viale delle Terme di Caracalla, 00153 Rome, Italy

ARTICLE INFO

Article history:
Received 16 July 2015
Received in revised form
14 December 2015
Accepted 16 December 2015
Available online 7 January 2016

Keywords:
Absence locations
Species distribution maps
Occurrence data
Ecological niche modelling
Marine biodiversity
Scientific surveys

ABSTRACT

Estimating absence locations of a species is important in conservation biology and conservation planning. For instance, using reliable absence as much as presence information, species distribution models can enhance their performance and produce more accurate predictions of the distribution of a species. Unfortunately, estimating reliable absence locations is difficult and often requires a deep knowledge of the species' distribution and of its abiotic and biotic environmental preferences and tolerance. In this paper, we propose a methodology to reconstruct reliable absence information from presence-only information, and the conditions that those presence-only data have to meet to make this possible.

Large species occurrence data collections (otherwise called occurrence datasets) contain high quality and expert-reviewed species observation records from scientific surveys. These surveys can be used to retrieve species presence locations, but they also record places where the species in their target list were not observed. Although these absences could be simply due to sampling variation, it is possible to intersect many of these reports to estimate true absence locations, i.e. those due to habitat unsuitability or geographical hindrances. In this paper, we present a method to generate reliable absence locations of this type for marine species, using scientific surveys reports contained in the Ocean Biogeographic Information System (OBIS), an authoritative species occurrence dataset. Our method spatially aggregates information from surveys focussing on the same target species. It detects absence locations for a given species as those locations in which repeated surveys (that included the species of interest in their target list) reported information only on other species. We qualitatively demonstrate the reliability of our method using distribution records of the Atlantic cod as a case study. Additionally, we quantitatively estimate its performance using another authoritative large species occurrence dataset, the Global Biodiversity Information Facility (GBIF). We also demonstrate that our approach has higher accuracy and presents complementary behaviour with respect to another method using environmental envelopes. Our process can support species distribution models (as well as other types of models, e.g. climate change models) by providing reliable data to presence/absence approaches. It can manage regional as well as global scale scenarios and runs within a collaborative e-Infrastructure (D4Science) that publishes it as-a-Service, allowing biologists to reproduce, repeat and share experimental results.

1. Introduction

Species distribution models (SDMs) estimate species distributions at global or local scale, by relating species occurrence records to a set of environmental parameters. SDMs have high potential in conservation biology and conservation planning, because they give hints to understand the relationship between a species and its abiotic and biotic environment, and to test ecological or biogeographical hypotheses about species distributions and ranges.

* Corresponding author. Tel.: +39 050 315 2978; fax: +39 050 621 3464.
E-mail addresses: coro@isti.cnr.it (G. Coro), chiara.magliozzi@isti.cnr.it (C. Magliozzi), evberghe@gmail.com (E. Vanden Berghe), nbailly@hcmr.gr (N. Bailly), Anton.Ellenbroek@fao.org (A. Ellenbroek), pagano@isti.cnr.it (P. Pagano).

http://dx.doi.org/10.1016/j.ecolmodel.2015.12.008
0304-3800/

They generalise the distribution that can be inferred from the observed locations, and possibly account for bias due to non-uniform observations sampling. Several technologies are used to build SDMs, ranging from explicit modelling of physiological limits and tolerances (Pearson, 2012), to the automatic correlation between species presence and environmental characteristics (Elith and Leathwick, 2009). Often, the output of an SDM is a probability distribution map reporting locations, at a certain resolution, where habitat is suitable for a species. SDMs usually use habitat information on recorded species observations and some models use also *habitat-related* absence locations, where habitat is unsuitable for species subsistence. Estimating these absence locations is a necessary step in these SDMs and requires separate modelling effort, e.g. envelope models based on species preferences to abiotic and biotic factors (Barbet-Massin et al., 2012). In this paper, we will distinguish these locations from the absences reported by scientific surveys (*sampling* absences). *Sampling* absences could refer either to complete absence of a species in a certain location, or to "undetected" presence, which could be due to intrinsic issues in species detectability and seasonality, or just to random sampling variation. Thus, the difference between *habitat-related* absences and *sampling* absences is in the fact that the former type is estimated from abiotic and biotic parameters, and the latter type is estimated from surveyed locations without presence data. Both the types are *pseudo*-absences because they use partial information about the species to estimate absence locations. Apart from *pseudo*-absences, in this work we will use the expression *absence locations* (or *true* absences) to indicate locations where the species is absent due to real habitat unsuitability or geographical hindrances. Based on this nomenclature, we define *reliable pseudo*-absences as those *pseudo*-absences that well approximate real absence locations.

Some SDMs rely on presence information only, for example Genetic Algorithm for Rule-set Production (GARP) and Ecological Niche Factor Analysis (ENFA) are based on simulated *pseudo*-absences (Stockwell, 1999; Engler et al., 2004), but models relying on both presence and absence information may reach higher accuracy and are especially better when modelling rare species distributions (Guisan and Thuiller, 2005; Ferrier, 2002; Gibson et al., 2007). Unfortunately, this requires estimating reliable pseudo-absence information (Guisan and Zimmermann, 2000; Coro et al., 2013c), which is not always possible. Today, large species occurrence data collections, sometimes referred to as species occurrence datasets (Jones et al., 2012; Casal et al., 2013) (SODs), expose high quality, expert-reviewed species observation records. For each record, these datasets usually provide information about (i) the recording time, (ii) the scientist who recorded the occurrence, (iii) the revision of the database record and (iv) the scientific survey this record belongs to. These surveys are the main source of information of large SODs, but they record species occurrences only along their routes (OBIS, 2015a; Vanden Berghe et al., 2010b,a; Tsontos and Kiefer, 2002; Ricard et al., 2010; Zeller et al., 2005; Halpin, 2009). These datasets usually store only observation records, and few examples of SODs storing also routes trajectories are available (Halpin et al., 2006), which would be useful when assessing absence locations.

Surveys usually focus on a limited taxonomic scope, for which the research vessel's scientific crew has expertise in identification. Large SODs usually do not report sampling absence information from surveys for a given species, but it is possible to reconstruct this information from locations where only other species' presence was reported. This reconstructed pseudo-absence information could be just due to the random geo-temporal sampling variation, possibly causing missed observations (e.g. individual undetected due to its behaviour or poor survey conditions), and could not reliably indicate habitat unsuitability or geographical absence in general. Further processing, in fact, is required to separate true absence from absence due to random sampling variation.

This paper presents a method to estimate absence locations for marine species, based on sampling-absences. In particular, we present a process to generate reliable pseudo-absence locations, i.e. absences that well approximate true absences. This process uses scientific survey data from an authoritative SOD containing a large amount of marine species observation records, i.e. the Ocean Biogeographic Information System (OBIS, Grassle, 2000; Vanden Berghe et al., 2010b; OBIS, 2015c). For each analysed species, our method (i) collects information from surveys that had the species in their target list, (ii) intersects and processes surveys' report locations to produce presence locations and sampling-absences, (iii) selects sampling-absence locations as those that are well separated from presence locations, i.e. not overlapping with presence locations according to a user-defined distance threshold.

In the paper, we take the Atlantic cod (*Gadus morhua* Linnaeus, 1758; Gadiformes: Gadidae) as a case study to demonstrate the reliability of our method. Additionally, we compare the performance of our process with another approach based on environmental envelopes. We use benchmark data from another authoritative SOD, the Global Biodiversity Information Facility (GBIF, Edwards et al., 2000; GBIF, 2014) for this comparison. Our process runs within a collaborative e-Infrastructure that publishes it as-a-Service (D4Science, 2015; Candela et al., 2015a; Coro et al., 2013a). The D4Science e-Infrastructure hosts this algorithm within a free-to-use platform using Cloud computing to execute processes (Coro et al., 2014a). This platform allows for (i) producing reliable pseudo-absence records, (ii) enriching them with environmental information, (iii) filtering on environmental values and (iv) using them in ecological niche models.

This paper is organized as follows: Section 2 reports about modelling methods to estimate species pseudo-absences. Section 3 gives the details of our approach along with its limitations. Section 4 reports a case study for the Atlantic cod and a statistical analysis on 550 aquatic species to quantitative estimate of the performance of our process. It also evaluates the sensitivity of our method to the values of two crucial input parameters. Finally, Section 5 contains summary considerations, including possible usages of our method in other models.

2. Overview

In this section, we briefly introduce species distribution models and describe their dependency on species presence and absence information. Then, we report methods to generate pseudo-absences. Finally, we discuss about the dependency of presence-only methods on the quality of data.

A variety of methods are currently used to build predictive species distribution models, which can be classified as those relying on presence-only versus presence/absence data (Pearson, 2012). Presence-only methods usually search for correlations between environmental parameters and observation records, whereas presence/absence approaches also use information about locations where the species of interest was not found. Presence/absence models have proven to improve their performance when reliable pseudo-absence information is available (Brotons et al., 2004; Guisan and Zimmermann, 2000).

Several methods are available to automatically estimate *pseudo*-absence locations, e.g. randomly taking locations (named "background points") in the area under analysis (Stockwell, 1999) to maximize relative differences with respect to known presence points (e.g. in the MaxEnt model, Elith et al., 2011), or using weighting criteria based on environmental information (Engler et al., 2004; Zaniewski et al., 2002). However, producing realistic