A practical method to speed up the discovery of unknown populations using Species Distribution Models

Mauro Fois [a], Giuseppe Fenu [b], Alba Cuena Lombraña [a,*], Donatella Cogoni [a], Gianluigi Bacchetta [a]

[a] Centro Conservazione Biodiversità, Dipartimento di Scienze della Vita e dell'Ambiente, Università degli Studi di Cagliari, Viale S. Ignazio da Laconi, 11-13, Cagliari 09123, Italy
[b] Dipartimento di Biologia Ambientale, 'Sapienza' Università di Roma, P.le A. Moro 5, 00185 Roma, Italy

ARTICLE INFO

ABSTRACT

Species Distribution Models (SDMs) could be an important tool to limit search efforts by selecting the areas where field surveys are to be carried out; due to the constant decrease of financial funds, this challenging purpose is particularly necessary. In particular, these methods are useful when applied to endangered and/or rare species with a poor known distribution area, especially due to difficulties in plant detection and in reaching the study areas.

We hereby describe the development of maximum-entropy (Maxent) models for the endangered yellow gentian *Gentiana lutea* L. in Sardinia with the aims of (i) guiding survey efforts; (ii) estimating SDMs utility by post-test species current/extinct localities through the Observed Positive Predictive Power (OPPP) values; and (iii) evaluating the influence of sample data addition. Besides the Area Under Curve (AUC) values, we used the OPPP (observed/modelled positive localities ratio) to compare results from eight, 24 and 58 presence-only data points. Even with the initial small and biased sample data, we found that surveys could be effectively guided using such methods, whereby the focus of our research was on 48% of our initial 721 km^2 study area. The high OPPPs values additionally proved the reliability of our results in discovering 16 new localities of *G. lutea*. Nevertheless, the predictive models should be considered as a complementary tool rather than a replacement for expert knowledge.

© 2015 Elsevier GmbH. All rights reserved.

Introduction

Although the constantly increasing number of threatened species (plants and animals) call for prioritization (Pimm et al., 1995), even funds and experts availability for discovering biodiversity are decreasing (Grieneisen et al., 2014) due to the current financial crisis; this trend has also been detected for Sardinian threatened flora (Fenu et al., 2015). Thus, a critical issue for research is nowadays the optimization of the efforts (Stroud et al., 2014). In this framework, several software packages implementing Species Distribution Models (SDMs) might be of great help since they are often open source and they can be used to carry out statistical analyses without excessive costs.

Consequently, SDMs have become commonplace in biological studies, as a tool for exploring basic ecological questions (e.g. Ashcroft et al., 2011; Bucklin et al., 2015; Guisan & Zimmermann, 2000), paleo-ecological scenarios (e.g. Patsiou et al. 2014; Russo et al., 2014; Varela et al., 2011), future ecological changes (e.g. Araújo et al., 2011; Randin et al., 2009) and providing support to species conservation or reserve planning (e.g. Araújo & Williams, 2000; Bosso et al., 2013; Guisan et al., 2013). SDMs also have practical applications to environmental management, such as detecting unknown populations of endangered species (e.g. De Siqueira et al., 2009; Jarvis et al., 2005; Jiménez-Valverde et al., 2008; Rebelo & Jones, 2010; Williams et al., 2009).

The utility of such models depends on many factors; e.g. in case of rare and\or difficult-to-detect species, most of these models are mainly connected to the quantity and quality of initial distributional data. Indeed, they are often limited to small samples of observed localities due, for example, to scarce recent field survey efforts and to the lack of precise localities and bio-ecological data associated with some museum specimens (Graham et al.,

* Corresponding author. Tel.: +39 0706753508; fax: +39 0706753509.
E-mail addresses: mau.fois1@studenti.unica.it (M. Fois), giuseppe.fenu@uniroma1.it (G. Fenu), albacuena@gmail.com (A. Cuena Lombraña), donatella.cogoni@unica.it (D. Cogoni), bacchet@unica.it (G. Bacchetta).

2004; Soberón & Peterson, 2004; Tessarolo et al., 2014). These problems are particularly frequent when data come from poorly known ecosystems (e.g. tropical) where distributional data are scarce (Bosso et al., 2013; Lomba et al., 2010; Pearson et al., 2007). Some of the methods employed in ecological modelling require absence data to generate SDMs, e.g. General Linear Model (GLM) and Random Forest (RF), whereas others are exclusively based on presence data, e.g. Maximum Entropy (Maxent) and Genetic Algorithm for Rule-set Prediction (GARP). Modelling species with presence-only data has been particularly used for such species with small distributional range and where knowledge is scarce (Pearson et al., 2007; Shcheglovitova & Anderson, 2013). However, the lack of a surveyed locality still affect the model performance and validation can be problematic (Chen & Lei, 2012; Pearson et al., 2007; Wisz et al., 2008).

Such analyses have been used for population discoveries of either cryptic, rare or endangered animals (e.g. Raxworthy et al., 2003; Rebelo & Jones, 2010; Verovnik et al., 2014) or wild plants (e.g. Bourg et al., 2005; De Siqueira et al., 2009; Jarvis et al., 2005). This is the case with the species addressed in our study: the yellow gentian (*Gentiana lutea* L. subsp. *lutea*, hereafter *G. lutea*) in Sardinia.

G. lutea deserves special attention because it has been included in the CITES convention and European Habitats Directive; furthermore, the species was assessed as Least Concern (LC) according to the IUCN methodology both at European (Bilz et al., 2011) and national levels (Gentili et al., 2013). The roots are traditionally used to prepare bitters and liqueurs (Pérez-García et al., 2012), as well as pharmaceuticals such as anti-inflammatory agents and diuretics (Nastasijević et al. 2012). Consequently, an excessive harvesting and a subsequent decrease in abundance of this species has been observed in several European territories (e.g. Gentili et al., 2013; Kery et al., 2000). In Sardinia, the plant distribution range is characterized by small groups or scattered individuals located at the edge of its distribution range, as a typical peripheral and isolated plant population (PIPP). In addition, due to a lack of knowledge, the current distributional information on *G. lutea* in Sardinia is incomplete and biased.

Owing to these limits, presence-only modelling appeared best suited to deal with its potential distribution in order to reduce survey efforts. We also tried to identify the historical extinct localities through a deep investigation guided by model results and ecological field investigations. In this study, we developed a Maximum Entropy (Maxent; Phillips et al., 2006) presence-only distribution model for *G. lutea* to reach the following goals: (1) to use the SDMs trained by small sample data for guiding discoveries of new localities; (2) to evaluate the influence of extant and extinct localities addition on model; and (3) to evaluate the models using the positive predictive power values calculated from a post-test observed data.

Methods

Study area and data collecting

According to bibliographic data (Chiappini & Angiolino, 1983; Gentili et al., 2013), herbarium specimens (CAG, CAT, FI, RO, SASSA, SS, TO *Herbaria*), and unpublished data by the authors, we selected the Gennargentu Massif as the whole distributional area of *G. lutea* in Sardinia (Western Mediterranean Basin). The Gennargentu Massif (Fig. 1), located in the Central-Eastern part of the island, has a surface of c. 721 km^2 and consists of a system of summits and windy ridges at 1400–1500 m a.s.l., with four peaks at more than 1800 m a.s.l. This area has been recently defined as an independent biogeographical sector based on the peculiar presence of its endemic flora and geomorphological units principally

constituted by metamorphic outcrops (Bacchetta et al., 2013; Fenu et al., 2014).

First eight localities used for the SDMs indicated by the scarce bibliographic sources (Chiappini & Angiolino, 1983; Gentili et al., 2013) were confirmed by field surveys and georeferred with a hand-held GPS receptor (Garmin e-Trex 20, Schaffhausen, SW). The same methodology was thus used for the subsequent discoveries.

Distribution modelling

The model comparison was very complicated due to the fact that our study case was based on a plant species with a poorly known distribution, and investigations started based only on occurrence data. We therefore chose *a priori* the most applied method for modelling species distributions with scarce presence-only data, i.e. the Maximum Entropy modelling (Maxent; Phillips et al. 2006) ver.3.3.3k (http://www.cs.princeton.edu/~schapire/maxent). This method usually results in good predictive models compared to other presence-only models (e.g. Elith et al., 2006; Ramírez-Villegas et al., 2014; Wisz et al., 2008). Besides the predictive qualities of the technique, it is a generative approach, rather than a discriminative, which can be an inherent advantage when the amount of training data is limited (Phillips et al., 2006). Furthermore, it has a good ability to predict new localities for poorly known species (Pearson et al., 2007; Rebelo & Jones, 2010; Verovnik et al., 2014).

Recommended default values were used for the convergence threshold (10^{-5}) and maximum number of iterations (500), and the analysis of variable importance was measured by jackknife, response curves and random seed. Suitable regularization multiplier (fixed at 1), included to reduce overfitting, were also selected automatically by the Maxent program (Phillips et al., 2006). The form of replication used was the cross-validation; as suggested by Pearson et al. (2007) for testing small samples, this run type makes it possible to replicate *n* sample sets removing each time one locality.

Eco-geographical variables

According to our research goals, the extension of the study area, the previous ecological knowledge and sample size, we chose to avoid the promiscuous use of variables (i.e. model complexity) in order to reduce model overfitting (Anderson & Gonzalez, 2011). In addition, we used the finer resolution as possible which usually provides better predictions, especially for fixed or very locally mobile organisms (Guisan & Thuiller, 2005). Environmental data at different spatial resolution (10, 250 and 1000 m) were obtained from the web: Digital Elevation Model (DEM) (http://www.sardegnageoportale.it) and Moderate Resolution Imaging Spectroradiometer (MODIS) data (http://modis.gsfc.nasa.gov). Topographic variables (altitude, slope and aspect) were derived from a 10-m resolution DEM. In addition, we computed four layers that represented estimates of vegetation cover and surface temperatures. Two variables were generated from the MODIS 16-day Enhanced Vegetation Index (EVI) and the 16-day Normalized Difference Vegetation Index (NDVI) at 250-m spatial resolution. Layers for the month of June (*G. lutea* flowering period; Gentili et al., 2013) for the last five years (2010-2014) were downloaded using USG MODIS Reprojection Tool (https://lpdaac.usgs.gov/tools/modis_reprojection_tool) and the ensemble of years were computed by Biodiversity R package (Kindt, 2008) in R (R Development Core Team, 2010). The same packages were also used to process further variables at a lower resolution (1000 m). These were obtained by the daytime MODIS 8-day land surface temperature: the surface temperature of June (Tjune) and February (Tfeb) were generated by the ensemble of two extreme dates of each month (since 2000 to 2006). All data