Biological Conservation 204 (2016) 16–22

Contents lists available at ScienceDirect

Biological Conservation

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

Overcoming data deficiency in reptiles

Lucie M. Bland [a,*], Monika Böhm [b]

[a] Centre of Excellence for Environmental Decisions, School of BioSciences, The University of Melbourne, Parkville 3010, VIC, Australia
[b] Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, United Kingdom

ARTICLE INFO

Article history:
Received 27 November 2015
Received in revised form 5 May 2016
Accepted 24 May 2016
Available online 3 June 2016

Keywords:
Data Deficient
Extinction risk
IUCN Red List
Random forests
Reptiles

ABSTRACT

We have no information on the risk of extinction of 21% of reptiles listed as Data Deficient on the Sampled Red List Index (SRLI), an indicator developed to track global change in species status. Data Deficient species are of high research priority, because they contribute to uncertainty in estimates of extinction risk and are neglected by conservation programmes. We review the causes of data deficiency in reptiles; the likely status of Data Deficient reptiles; and possible solutions for their re-assessment. We find that 52% of Data Deficient reptiles lack information on population status and trends, and that few species are only known from type specimens and old records. We build a random forest model for SRLI species of known extinction risk, based on life-history, environmental and threat information. The final model shows perfect classification accuracy (100%) in ten-fold cross validation. We use the model to predict that 56 of 292 Data Deficient reptiles (19%) are at risk of extinction, so the overall proportion of threatened reptiles in the SRLI (19%) remains unchanged. Regions predicted to contain large numbers of threatened Data Deficient reptiles overlap with known centres of threatened species richness. However, the model shows lower accuracy (79%) on 29 species recently re-assessed in the Global Reptile Assessment. Predictive models could be used to prioritize Data Deficient species and reptiles not included in the SRLI, and new reptile assessments could be used to improve model predictions through adaptive learning.

1. Introduction

Reptiles are one of the most diverse groups of terrestrial vertebrates, with 10,272 species described as of August 2015 (Uetz and Hošek, 2015). Hundreds of reptile species are still being discovered per year (Uetz and Hošek, 2015) on par with rapid species discoveries in amphibians (Köhler et al., 2005). Yet as a group, reptiles are poorly represented on The IUCN Red List of Threatened Species™, with only 44% of described species evaluated (IUCN, 2015). Within those, 1500 were assessed in a representative manner using the Sampled Red List Index (SRLI) method (Baillie et al., 2008), producing the first global analysis of the extinction risk of reptiles. This method selects species randomly from the species list, producing a broadly representative picture of extinction risk status for reptiles at the order and family level (Böhm et al., 2013). The analysis revealed that one in five reptile species is threatened with extinction, with risk highest among freshwater species and in tropical regions (Böhm et al., 2013).

Within the reptile SRLI assessment, 318 species were assessed as Data Deficient (Böhm et al., 2013) due to insufficient information on species' taxonomy, ecology, distribution, population trends, and/or threats (IUCN, 2001). The Data Deficient category does not correspond to a level of extinction risk, but indicates that further information should be collected to assign species to a risk category (IUCN, 2001). High levels of data deficiency within group assessments are problematic for a number of reasons. First, ignoring Data Deficient species (or treating them as threatened or non-threatened) contributes to considerable uncertainty in extinction risk patterns (Bland et al., 2012; Butchart and Bird, 2010; Hoffmann et al., 2010). This uncertainty not only affects the monitoring of progress towards global biodiversity targets (e.g. Convention on Biological Diversity (CBD) Aichi targets), but also conservation priorities which rely on threatened species lists –such as Key Biodiversity Areas, biodiversity hotspots, and many others (Brooks et al., 2006). Data Deficient species are rarely included in national recovery plans, conservation legislation, conservation planning (Sousa-Baena et al., 2014), and investment schemes. For example, only 2% of the awards from the Mohamed Bin Zayed Species Conservation Fund (MBZSC, 2014) and only one project of the World Association of Zoos and Aquaria (WAZA, 2013) exclusively focus on Data Deficient species. All in all, Data Deficient reptiles are offered very little protection and funding due to their uncertain extinction risk.

Reptiles included in the SRLI show intermediate levels of data deficiency (21%; Böhm et al., 2013). Data Deficiency is higher than in birds (0.6%; Butchart and Bird, 2010) and mammals (15%; Schipper et al., 2008), but lower levels than in amphibians (25%; Stuart et al., 2004), Odonata (35%; Clausnitzer et al., 2009), and freshwater crabs (49%; Cumberlidge et al., 2009). Data Deficient reptiles are also better-known than Data Deficient species in other groups (Bland et al.,

---

* Corresponding author.
E-mail address: l.bland@unimelb.edu.au (L.M. Bland).

http://dx.doi.org/10.1016/j.biocon.2016.05.018

submitted). The reasons for Data Deficient status in reptiles often quote uncertain population status (33%) and uncertain threats (23%). Comparatively few listings are based on severe sources of uncertainty, namely type records (10%), few records (12%), old records (6%), or records of uncertain provenance (4%). These severe sources of uncertainty collectively represent a higher proportion of Data Deficient listings in freshwater crabs (93%), amphibians (43%), and mammals (42%) than in reptiles (32%; Bland et al., submitted). Some reptiles are listed as Data Deficient due to taxonomic uncertainty (6%) and recent discovery within 10 years of the SRLI assessment (6%). Lack of knowledge of reptiles' threat status is not only exemplified by Data Deficient species; indeed more than half of all known reptile species remain Not Evaluated (IUCN, 2015). Addressing data deficiency and broader data gaps are therefore key challenges for reptile assessments.

Data Deficient species have received increased interest from the conservation literature in recent years, with studies investigating the rationale for the use of the category (Butchart and Bird, 2010; Sousa-Baena et al., 2014), the effect of Data Deficient species on conservation priorities (Trindade-Filho et al., 2012), and their potential for informing future biodiversity inventories (Brito, 2010). Considerable progress has been made in predicting the likely status of Data Deficient species based on contextual data (e.g. biology, ecology, phylogeny, environment, and threats). For example, distribution maps are available for most Data Deficient species and can be used to quantify risk factors, such as geographical range size and exposure to anthropogenic threats. These contextual data alone are insufficient for making a decision on formal Red List status, but are available for a wide range of groups and have been used to predict risk in plants (Good et al., 2006; Walker, 2014), mammals (Bland et al., 2015a; Davidson et al., 2009; Jones and Safi, 2011), amphibians (Bland et al., 2015b; Morais et al., 2013), and crayfish (Bland et al., 2015b). Simple rules of thumb based on one or two factors may not capture complex extinction risk dynamics, as strong correlations between factors influencing both knowledge availability and extinction risk may lead to unreliable estimates of risk. For example, Data Deficient reptiles typically possess small geographic ranges (Bland, 2014) and may be more likely to be threatened (Böhm et al., 2015), but Data Deficient reptiles also tend to be small-bodied (Bland, 2014; Vilela et al., 2014), a factor usually associated with low extinction risk (Cardillo and Meijaard, 2012). Predicting risk in Data Deficient must therefore rely on a wide variety of contextual data and robust modelling approaches.

Predictive studies of risk in Data Deficient species span a diversity of methods, such as spatial-phylogenetic frameworks, rule-based methods, and machine learning (Bland et al., 2015b; Jones and Safi, 2011; Morais et al., 2013). Machine learning methods are powerful tools for finding patterns in large datasets and rely on few assumptions. This had made machine learning increasingly popular in ecology (Cutler et al., 2007; Prasad et al., 2006) and extinction risk analyses (Bland et al., 2015a; Davidson et al., 2009). A previous study highlighted the usefulness of random forests for predicting extinction risk in different taxonomic groups, including reptiles (Bland et al., 2015b). Random forests are an ensemble method related to classification trees, where many classification trees are constructed and predictions obtained by a majority vote (Breiman, 2001). For each tree, only a randomly chosen subset of the explanatory variables is used at each node, which reduces correlation between trees. Random forests are widely used in ecology due to their high predictive power and their robustness to overfitting and noise (Cutler et al., 2007). In addition, the outputs of random forests are probability estimates of a given outcome, which allow easy interpretation of uncertainty when predicting complex processes such as extinction risk.

Overcoming data deficiency will be a costly exercise (Bland et al., 2015b): research needs to be carried out on the key aspects for which species information is missing (e.g. population status, threats, taxonomy), and this new evidence will need to be synthesized in assessments. Prioritizing Data Deficient species for re-assessment will require information on both their likely risk status and an understanding of the underlying causes of this data deficiency. We predict extinction risk of Data Deficient species in the random representative sample of 1500 reptiles by Böhm et al. (2013) (hereafter termed 'SRLI assessment') as a first step to reducing uncertainty in extinction risk patterns. We ask the following questions: i) Can a random forest model predict risk in species of known extinction risk (non-Data Deficient)?; ii) What is the predicted level of extinction risk faced by Data Deficient species?; and iii) How can our findings direct re-assessments?

2. Methods

2.1. Dataset

We used a published extinction risk trait database on 1416 terrestrial reptile species (Appendix A; Bland et al., 2015b) included in the SRLI assessment (Böhm et al., 2013). We did not collect data for non-SRLI species. No species of crocodiles were assessed as Data Deficient so predictions for this taxon cannot be made. We therefore excluded four crocodile species, resulting in 1412 species in the dataset. Data were compiled from species descriptions, field guides, museum specimens and published life-history studies, and supplemented with data obtained from expert herpetologists during the IUCN Red List assessment process. References and raw data are available in Appendices A and B, respectively. The dataset contained the following life-history and ecological variables: maximum body size (snout-vent length), reproductive mode, habitat mode, trophic level, island presence, and number of IUCN-listed habitats. Information on species' niche and threat exposure was available as mean values from within species' geographic ranges for: mean annual temperature, temperature seasonality (coefficient of variation), annual precipitation, precipitation seasonality, minimum elevation (Hijmans et al., 2005), Human Footprint (CIESIN, 2005a), and mean and minimum human population density for the year 2000 (CIESIN, 2005b). Geographical range size and latitude of range centroid were computed from the IUCN distribution maps of each species. We did not undertake variable selection as uninformative variables are unlikely to affect predictive performance in analyses with fewer variables than species (low dimension problems; Kuhn, 2008). We defined non-Data Deficient species on the IUCN Red List as threatened (IUCN categories Critically Endangered, Endangered or Vulnerable) or non-threatened (IUCN categories Near Threatened or Least Concern).

2.2. Modelling

We used random forests to predict the extinction risk (non-threatened or threatened) of data-sufficient species. Random forests cannot currently take into account phylogenetic relatedness among species, so we included taxonomic order, family, and genus as predictor variables to partially account for shared evolutionary history. All numeric variables were centred to a mean of zero and scaled to a standard deviation of one (Kuhn, 2008). All categorical variables were transformed to orthogonal dummy variables. We removed highly correlated ($r > 0.9$) and low variance (frequency ratio > 999 and unique value percentage < 0.0001) variables, which can lead to collinearity and zero variance in cross-validation partitions (Kuhn, 2008). We used ten-fold cross validation to optimize the number of variables chosen randomly at each node for each tree, and grew 500 trees for each random forest iteration. We maximized the area under the receiver operating characteristic curve (AUC) to measure classification performance as AUC is insensitive to unequal numbers of threatened and non-threatened species in training data, and does not require the specification of misclassification costs (Fawcett, 2006). We identified the probability threshold above which a species was identified as threatened by maximizing the Youden index ($Y = \text{sensitivity} + \text{specificity} - 1$; Youden, 1950), which lends equal weight to detecting threatened and non-threatened species (Bland et al., 2015a). All analyses were conducted in R version