Research Paper

Assessing uncertainty and performance of ensemble conservation planning strategies

Yu-Pin Lin[a], Wei-Chih Lin[a,*], Johnathen Anthony[a], Tzung-Su Ding[b], Jean-Baptiste Mihoub[c], Klaus Henle[c], Dirk S. Schmeller[c,d,e]

[a] Department of Bioenvironmental Systems Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan
[b] School of Forestry and Resource Conservation, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan
[c] UFZ – Helmholtz Centre for Environmental Research, Department of Conservation Biology, Permoserstr. 15, 04318 Leipzig, Germany
[d] Université de Toulouse, UPS, INPT, EcoLab (Laboratoire Ecologie Fonctionnelle et Environnement), 118 route de Narbonne, 31062 Toulouse, France
[e] CNRS, EcoLab, 31062 Toulouse, France

ARTICLE INFO

ABSTRACT

Systematic conservation initiatives attempt to cater to the needs of many species via the integration of multiple species distribution models (SDMs), or via the integration of Systematic Conservation Planning (SCP) software, such as Zonation. Unfortunately, due to limited data and knowledge, it is often difficult to select the most suitable model for specific species, let alone an appropriate ensemble modeling method for multiple species. In general, model selection criteria are based on either model performance or consensus. The former integrates the highest-performing SDM for all focal species, whereas the latter integrates multiple SDM outputs based on consensus. While higher-performing ensemble models presumably identify high-quality habitats better, many have argued that high consensus ensemble models have less uncertainty originating from sporadic model variability. This study develops and validates seven ensemble-modeling strategies for integrating outputs of the systematic conservation tool Zonation. First, we considered the distributions of 11 bird species via 100 runs of five SDMs across Taiwan. Second, we evaluated the local and global uncertainty of all five models. Third, we used Zonation to obtain conservation priorities. We then used Principal Component Analysis (PCA) to quantify different sources of uncertainty. Finally, we used independent third-party habitat data to validate each strategy. On average, the 'best model' strategy (based on the highest AUC value) performed best. Based on our modeling exercise we present a comprehensive framework for conservation prioritization, validation and the quantification of uncertainty intrinsic to SDMs according to different conservation scenarios and goals.

1. Introduction

Species distribution models (SDMs) are widely used to support conservation decision-making (Guisan et al., 2013; Guillera-Arroita et al., 2015). Researchers and decision makers have used various types of SDMs along with systematic conservation planning (SCP) (Buisson, Thuiller, Casajus, Lek, & Grenouillet, 2010) software packages to assess or delimit conservation areas (Guisan et al., 2013). SDM outputs generally represent the probability of species presence (Liu, Berry, Dawson, & Pearson, 2005), and/or environmental suitability (Guisan et al., 2013; Robertson, Peter, Villet, & Ripley, 2003). The primary objective of SCP is to prioritize areas that maximize biodiversity. To accomplish this, most initiatives reference the outputs of SDMs (Guisan et al., 2013; Moilanen, Wintle, Elith, & Burgman, 2006). Since all SDMs

suffer from uncertainty, e.g. the overestimation of species distributions (Beale & Lennon, 2012), conservation initiatives should include uncertainty analysis (Diniz-Filho et al., 2009; Guisan et al., 2013; Lentini & Wintle, 2015; Meller et al., 2014; Moilanen, Runge et al., 2006; Moilanen, Wintle et al., 2006; Watling et al., 2015). There are currently a number of studies which attempt to examine the effects of uncertainty and performance within and between SDMs on conservation area selection (Carvalho, Brito, Crespo, Watts, & Possingham, 2011; Gallo & Goodchild, 2012; Guisan et al., 2013; Hermoso & Kennard, 2012; Langford, Gordon, & Bastin, 2009; Lentini & Wintle, 2015; Meller et al., 2014; Moilanen, Wintle et al., 2006; Rocchini et al., 2011; Watling et al., 2015). Given the spatial variability of SDM outputs and the central role they play in conservation planning, it is essential to continue examining both the sources of SDM uncertainty and the effects

they have on conservation area selection (Watling et al., 2015).

When discussing uncertainty, SDM structure is of utmost importance since it determines both how real world phenomena are simplified and the complexity of input data (Diniz-Filho et al., 2009; Huang, Lin, & Chiang, 2014; Moilanen, Wintle et al., 2006; Watling et al., 2015). Dormann, Purschke, Márquez, Lautenbach, and Schröder (2008) classified SDMs into three groups: i) "traditional" methods, e.g. Generalized Linear Models (GLM), ii) "machine-learning" methods, e.g. Support Vector Machines (SVM) and random forest (RF), and iii) "presence-only" techniques, e.g. maximum entropy. Since different SDM methods are more realistic for different species (Segurado & Araújo, 2004), with different data, and or under different requirements, there is no consensus on an optimal SDM for all situations. Furthermore, due to limitations e.g. logistical, financial (MacKenzie, 2005), it is often exceedingly difficult or impractical to identify an optimal SDM for all focal species considered in a specific conservation initiative. Therefore, the integration of model outputs via modeling, and subsequent uncertainty analysis has attracted attention in conservation area selection (Meller et al., 2014).

Ensemble modeling is attractive since it allows for the simultaneous consideration of multiple species, model types, and sampling regimes when undertaking modeling initiatives. Additionally, researchers may be able to identify optimal model types and sampling regimes, i.e. sets of samples, by comparing output performances. Conversely, and possibly more importantly, researchers can perform uncertainty analysis, thereby identifying and quantifying the sources of uncertainty. Although uncertainty is inherent to most, if not all, conservation decisions (Guisan et al., 2013) the use and analysis of ensemble models can quantify uncertainty and should lessen its effects.

Here, we develop a five-step conservation area selection procedure, capable of considering numerous species and able to evaluate the uncertainty and performance of multiple SDM types, sampling regimes, conservation area sizes and conservation area selection strategies (see workflow chart in Fig. 1). Our five-step procedure consists of (1) multiple runs of two traditional SDM methods, two machine-learning and one presence-only method for each species considered. Although all raw SDM outputs represent the probability of species presence or the relative scores, which indirectly represent the habitat suitability, for simplicity we consider species use directly proportional to habitat suitability; therefore we use the two terms interchangeably throughout this paper and define all SDM output values as Habitat Suitability Indices (HSI). (2) After comparing model AUC, true skill statistics (TSS) values (Allouche, Tsoar, & Kadmon, 2006) and levels of global uncertainty, we (3) use the software Zonation to create a hierarchy of conservation priorities for each SDM output. (4) We then use a principal component analysis (PCA) and Spearman rank correlation to assess the consistency among the conservation priority hierarchies. In the fifth step, we evaluate the effectiveness of seven conservation area selection strategies under varying conservation scenarios (percentage of land designated to conservation) with independent third-party data. Here, we applied our approach to data from Taiwan and show its ability to deal with SDM uncertainty. Following and developing the "post-selection" method, as presented in Meller et al. (2014), each conservation area selection strategy represents a different Zonation ensemble modeling technique. In addition, we compared the performance of each selection strategy, in terms of habitat coverage, with a null model. A large number of randomly generated conservation areas represented the null model.

2. Materials and methods

2.1. Environment data

Taiwan is a mountainous country located off the eastern coast of China (Fig. 2). The Tropic of Cancer bisects the country and divides it into two climatic zones, the tropical monsoon in the south and the sub-

tropical monsoon in the north. Approximately 60% of the island is covered by mountains, with over 200 summits exceeding 3000 m a.s.l. The complex geology and landscape has led to highly diverse vegetation. Here, we used 803 2 km × 2 km resolution presence/absence samples of 11 protected bird species to project species across Taiwan (Fig. 2(a)): *Erithacus johnstoniae, Regulus goodfellowi, Parus ater-aegithalosconcinnus, Actinodura morrisoniana, Brachypteryx montana, Liocichlasteeri garrulusglanderius, Parusmonticolus heterophasiaauricularis, Pericrocotussolaris niltavavivida, Arborophila crudigularis, Ficedula hyperythra* and *Myomela leucura* (Appendix Table S1).

We considered the following environmental variables in our species distribution models: elevation, temperature, precipitation, Normalized Difference Vegetation Index (NDVI), relative humidity, distance to major roads of national and provincial levels, sea or river (see Appendix Table S3 for more information) all at a resolutions of 2 km × 2 km. We used data from Lee, Ding, Hsu, and Geng (2004) for modeling. Finally, we used the 2013 data of Taiwan Breeding Bird Survey (BBS Taiwan) as independent data to validate our conservation selection strategies. BBS Taiwan, organized by Taiwan Endemic Species Research Institute, is a citizen science project, in which volunteers survey bird community composition in more than 300 sampling sites across Taiwan since 2009 (http://bbstaiwan.tw). The BBS Taiwan data were collected using consistent sampling methods and effort within the 2 km × 2 km sampling sites and are therefore suitable for independent validation.

2.2. Species distribution modeling and evaluation

Five species distribution models (GLM logistic, Random forest – RF, GAM; Support Vector Machine (SVM); maximum entropy) estimated the probability of species presence, i.e. HSI, for each planning unit. All models were implemented using our own code and different packages of the R software, including random Forest (Liaw & Wiener, 2002), M, gam (Hastie & Hastie, 2013), SDM Tools (VanDerWal, Falconi, Januchowski, Shoo, & Storlie, 2014) and e1071 (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2012). In GLM, we only included linear terms of environmental variable and a binomial link function. In RF, we used the default number of 500 trees. For GAM, we applied the default smooth terms, thin plate regression spline, for environmental variables. In SVM, we used the default setting of each parameter including radial basis kernel function, degree of 3 for the polynomial in kernel function, constraints violation cost of 1. The settings in maximum entropy were: convergence threshold = 10^{-5}, maximum iterations = 1000, regularization value $\beta = 10^{-4}$, with the use of linear features. The data for the 11 species were randomly divided into a training set, consisting of 80% of the data, and a 20% model validation set (Ko, Schmitz, Barbet-Massin, & Jetz, 2014; Parida, Hoffmann, & Hill, 2015; Strubbe, Jackson, Groombridge, & Matthysen, 2015). We chose an 80% training sample split in order to maximize the representativeness of each species. We repeated this split-sample procedure 100 times per SDM approach, creating 500 different outputs for each species. We verified the adequacy of using no more than 100 outputs for each model by running the same procedure but with double the outputs, i.e. 200 times per SDM for a total of 1000 outputs, which yielded nearly identical results (see Appendix Tables S4 and S5 in Supplementary material).

We evaluated the SDM performance for each species via the average of three different indices, AUC, Kappa, and TSS values (see Appendix A in Supplementary material). A multivariate analysis of variance (MANOVA) was applied to evaluate the difference of model performances between 5 SDMs in terms of the AUC values of the eleven species (Michel & Knouft, 2012). In order to confirm the pairwise difference between any two model performances, the post-hoc test of MANOVA was conducted via Hoteling's *t*-test. Hoteling's *t*-test is a multivariate analogue and a generalization of the univariate *t*-test that is designed to evaluate the differences between means of different distributions (Turner, Collyer, & Krabbenhoft, 2010). Due to the nonnormality of AUC, Kappa and TSS distributions, we used a permutation