Environmental Modelling & Software 100 (2018) 136–145

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Ensemble species distribution modelling with transformed suitability values

R. Kindt [1]

World Agroforestry Centre (ICRAF), 30677-00100 Nairobi, Kenya

ARTICLE INFO

Article history:
Received 12 May 2017
Received in revised form
13 October 2017
Accepted 8 November 2017

Keywords:
Species distribution model
Ensemble model
Spatial sorting bias
R statistical language and environment
Ecological software
BiodiversityR package

ABSTRACT

Species distribution modelling (SDM) was integrated in version 2.0 of the *BiodiversityR* package released in 2012. Ensemble habitat suitability is calculated as the weighted average of suitabilities predicted by different algorithms. Advanced options for SDM in the current version (2.8–4) of the package include tuning the best combination of the number and weights of models contributing to the ensemble suitability and calculating the absence-presence threshold as the average or minimum of recommended threshold values. Algorithm-specific suitability values can be transformed via generalized linear models with *probit* link so that they become more similar in range. Other options include reducing spatial sorting bias by selecting background locations in circular neighbourhoods and generating suitability maps that show the number of contributing models that predict species presence. The approaches are illustrated for two species with open-access point location data sets, *Bradypus variegatus* and *Thryothorus ludovicianus*.

1. Introduction

Species Distribution Models (SDMs; other names for these models include Ecological Niche Models and Habitat Suitability Models) are widely used in ecological studies (for an overview of the SDM framework, see Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005; Elith and Leathwick, 2009; Miller, 2010; Thuiller and Munkemuller, 2010; Hijmans and Elith, 2016). Recent developments in the calibration of SDMs such as the application of machine-learning algorithms (Elith et al., 2006; Wisz et al., 2008) and the use of ensemble (consensus) procedures (Araújo and New, 2007; Marmion et al., 2009; Thuiller et al., 2009; Buisson et al., 2010; Luedeling et al., 2014; Trolle et al., 2014) allow for the creation of more reliable habitat suitability maps. When new potentially superior algorithms for SDM become available, such as using zero-inflated random variables in hybrid Bayesian networks (Maldonado et al., 2016) or maximizing the likelihood of species occurrence probability (Royle et al., 2012), these can be easily integrated in the ensemble modelling framework. Since these powerful SDM methods have become available, challenges in SDM have shifted to areas such the availability of point location data (Boakes et al., 2010; Feeley and Silman, 2011; Duputié et al., 2014), dealing with errors and bias in point location data sets (Hortal et al., 2008; Loiselle et al., 2008; Platts et al., 2008; Phillips et al., 2009; Lobo and Tognelli, 2011; Syfert et al., 2013; Beck et al., 2014; Varela et al., 2014; Robertson et al., 2016) and projecting across space and time (Dormann, 2007; Elith et al., 2010; Stanton et al., 2012; Braunisch et al., 2013; Baker et al., 2016; Werkowska et al., 2016).

The *BiodiversityR* package (Kindt, 2017) was initially developed to accompany a manual on the statistical analysis of biodiversity and community ecology data (Kindt and Coe, 2005). The package currently provides a Graphical User Interface for ordination, cluster and diversity analysis using the *vegan* package (Oksanen et al., 2017). Ensemble approaches for SDM have been incorporated into *BiodiversityR* version 2.0 that was released in December 2012, initially to make model calibration procedures more explicit than in the *BIOMOD* package (model formulae are available as arguments in the main function that calibrates the ensemble model and the contributing models, whereas default formulae can be generated with function *BiodiversityR::ensemble.formulae*). A second reason was to build on functions offered by the *dismo* package (Hijmans et al., 2015) such as the *dismo::threshold* (allowing a wider range of methods of transforming suitability into absence-presence than were available in *BIOMOD*), *dismo::maxent*, *dismo::domain* and *dismo::mahal* functions (these functions fit suitability based on the

E-mail address: r.kindt@cgiar.org.
[1] http://orcid.org/0000-0002-7672-0712.

https://doi.org/10.1016/j.envsoft.2017.11.009
1364-8152/

maximum entropy (Phillips et al., 2006), DOMAIN (Carpenter et al., 1993) and Mahalanobis (1936) algorithms that were not available in *BIOMOD*). A third motivation was to address the challenge of developing *R* functions that create habitat suitability maps with limited user input during short training exercises or to familiarize users with model outputs. A final reason was to provide support to doctoral and post-doctoral SDM research (Ranjitkar et al., 2014a, 2016a, 2014b, van Breugel et al., 2015b, 2015a, 2016b).

Since the introduction of SDM methods in *BiodiversityR*, the development of SDM methods has developed independently in *BIOMOD* and *BiodiversityR*, whereas recently a new *R* package for SDM (*sdm*) was released (Naimi and Araújo, 2016). *BiodiversityR* offers various unique methods of SDM, whereas application of these methods is straightforward through a Graphical User Interface or by using functions with default argument settings.

2. Methods and features

In a similar way to ensemble modelling approaches implemented in the *BIOMOD* and *sdm* packages, SDM functions in *BiodiversityR* calculate ensemble suitability (S_e) as a weighted average of suitabilities predicted by contributing models (S_i):

$$S_e = \frac{\sum_i w_i S_i}{\sum_i w_i}$$

Previous studies have shown that the consensus method based on weighted averages may significantly increase the accuracy of SDM (Marmion et al., 2009). The current version of *BiodiversityR* (2.8–4 released in 2017) allows the calibration of 23 candidate models that could contribute to the calculation of S_e, including maximum entropy models (available via argument MAXENT; main R calibration function of *dismo::maxent*; Phillips et al., 2006; Elith et al., 2011; Hijmans et al., 2015), maximum likelihood models (MAXLIKE; *maxlike::maxlike*; Chandler and Royle, 2013; Royle et al., 2012), two different implementations of boosted regression trees (GBM and GBMSTEP; *gbm::gbm* and *dismo::gbm.step*; Friedman, 2001; Friedman et al., 2001; Elith et al., 2008; Ridgeway, 2015), random forests (RF; *randomForest::randomForest*; Breiman, 2001; Liaw and Wiener, 2012), (stepwise) generalized linear regression models (GLM and GLMSTEP; *stats::glm* and *MASS::stepAIC*; McCullagh and Nelder, 1989; Venables et al., 2002; Venables and Ripley, 2013), (stepwise) generalized additive models (GAM and GAMSTEP; *gam::gam* and *gam::step.gam*; Hastie and Tibshirani, 1990; Hastie, 2013), generalized additive models with integrated smoothness estimation (MGCV and MGCVFIX; *mgcv::gam*; Wood, 2011, 2013), multivariate adaptive regression spline models (EARTH, *earth::earth*; Friedman, 1991; Leathwick et al., 2005; Milborrow, 2014), recursive partitioning and regression trees (RPART; *rpart::rpart*; Breiman et al., 1984; Therneau et al., 2014), artificial neural networks (NNET; *nnet::nnet*; Venables et al., 2002; Ripley and Venables, 2013), flexible discriminant analysis (FDA; *mda::fda*; Hastie et al., 1994; Leisch et al., 2013), two different implementations of support vector machine models (SVM and SVME; *kernlab::ksvm* and *e1071::svm*; Karatzoglou et al., 2013; Meyer et al., 2014), lasso or elastic-net regularized generalized linear models (GLMNET; *glmnet::glmnet*; Friedman et al., 2016, 2010), two different implementations of the BIOCLIM algorithm (BIOCLIM and BIOCLIM.O whereby BIOCLIM.O follows the original methodology more closely in predicting suitability as 0, 0.5 or 1.0; *dismo::bioclim* and *BiodiversityR::ensemble.bioclim*; Nix, 1986; Booth et al., 2014; Hijmans et al., 2015), the DOMAIN algorithm (DOMAIN; *dismo::domain*; Carpenter et al., 1993) and two different implementations of the Mahalanobis algorithm (MAHAL and MAHAL01; *dismo::mahal*; Mahalanobis, 1936).

With default settings of the *ensemble.batch* function (Table 1), the only inputs required from users to generate suitability maps are presence point locations and a *rasterStack* object with raster layers representing explanatory variables. In the first step of the SDM procedure, ensemble weights for the models are obtained by a 4-fold cross-validation procedure whereby each of the four models are calibrated and tested with data not used for calibration (Hijmans, 2012; van Breugel et al., 2015a; Ranjitkar et al., 2016b) and the ensemble weight is calculated as the average AUC (the Area Under the Receiver-operator curve, a statistic commonly used to evaluate SD models; Bradley, 1997; Hijmans, 2012; Wisz et al., 2008; Jiménez-Valverde, 2012; Varela et al., 2014) over the four cross-validations. With default settings, presence and absence locations are randomly assigned to the four cross-validation bins. An alternative procedure (argument *get.block*) is available whereby presence and absence locations are divided in four blocks created by lines of latitude and longitude that divide the locations as equally as possible, a procedure that is expected to reduce spatial correlation between training and testing locations important for evaluating models that will transfer suitability across space or time (Muscarella et al., 2014).

Although the AUC is the most commonly used statistic to evaluate SDMs (Hijmans, 2012), various authors have criticised its use. Jiménez-Valverde (2012) documented strong correlation between the AUC and the absence-presence threshold that makes sensitivity (the proportion of correctly predicted presence locations) equal to specificity (the proportion of correctly predicted absence locations). As a consequence, the AUC may be equivalent to using a threshold that discriminates between predicted absence (unsuitable habitat) and predicted presence (suitable habitat), whereas the feature of avoiding the use of such absence-presence threshold is the main argument in favour of the AUC. When the realized distribution of the species does not represent the full potential distribution of a species, models with lower AUC can produce habitat suitability maps that better represent the potential distribution. Therefore, the AUC is appropriate mainly for models of realized distributions (Jiménez-Valverde, 2012). The AUC will be larger for species with more restricted distributions. Therefore, the statistic is expected to inflate when background locations are sampled from larger areas (Lobo et al., 2008; Hijmans, 2012). Despite these shortcomings, the AUC remains a valid measure of relative model performance for the same species (assuming that presence locations are representative of suitable habitat) and the same study area (Wisz et al., 2008). As such, AUC values can be used to compare different models that are candidates to contribute to S_e.

In the second step of the procedure, models with AUC values larger than 0.7 (an AUC threshold that is often used to identify "good" models; Hijmans, 2012) are calibrated with the full set of presence and background point locations instead of the 75% subsets used for the 4-fold cross-validations. In the final step, suitability maps are generated. It is possible to modify default settings such as the number of cross-validation steps (argument *k-splits*) or how models with lower AUC values are down-weighed (setting argument *ENSEMBLE.exponent* results in weights calculated as $AUC^{ENSEMBLE.exponent}$, a procedure suggested by Hijmans and Elith, 2016). Users can also substitute objects for contributing models, opt not to use default formulae to calibrate contributing models or use other weights for the final step.

Suitabilities predicted by the contributing models can be transformed by generalized linear models with *probit* link (argument *PROBIT*), thus ensuring that all suitabilities are probability values (the BIOCLIM, DOMAIN, MAHAL and MAHAL01 algorithms do not provide probability values). Since species presence (1) or absence (0) are used as binary response variables for the transformations, the transformation is expected to result in the ranges of