BIOMOD – a platform for ensemble forecasting of species distributions

Wilfried Thuiller, Bruno Lafourcade, Robin Engler and Miguel B. Araújo

BIOMOD is a computer platform for ensemble forecasting of species distributions, enabling the treatment of a range of methodological uncertainties in models and the examination of species-environment relationships. BIOMOD includes the ability to model species distributions with several techniques, test models with a wide range of approaches, project species distributions into different environmental conditions (e.g. climate or land use change scenarios) and dispersal functions. It allows assessing species temporal turnover, plot species response curves, and test the strength of species interactions with predictor variables. BIOMOD is implemented in R and is a freeware, open source, package.

Species distribution models (SDM, Guisan and Thuiller 2005) are being used in nearly all branches of life and environmental sciences. A quick search in ISI Web of Science (18/02/08) using “species distribution models” OR “niche models” OR “habitat models” OR “bioclimatic models” highlights 21 973 papers, 74% of which published in the past 10 yr, in fields as varied as environmental sciences (53% of the records), zoology (15%), marine and freshwater biology (15%), life sciences and biomedicine (9%), biodiversity and conservation (8%), evolutionary biology (8%), fisheries (6%), forestry (6%), oceanography (5%), genetics and heredity (5%), amongst others. Advancement of knowledge in these fields is now intertwined with technical innovation in species distribution modelling and dependent on the existence of suitable software for fitting models and examining results. One difficulty with the use of species distribution models is that the number of techniques available is large and is increasing steadily, making it difficult for “non-aficionados” to select the most appropriate methodology for their needs (Elith et al. 2006, Heikkinen et al. 2006). Recent analyses have also demonstrated that discrepancies between different techniques can be very large, making the choice of the appropriate model even more difficult. This is particularly true when models are used to project distributions of species into independent situations, which is the example of projections of species distributions under future climate change scenarios (Thuiller 2004, Pearson et al. 2006). A possible solution to account for this inter-model variability is to fit ensembles of forecasts by simulating across more than one set of initial conditions, model classes, model parameters, and boundary conditions (for a review see Araújo and New 2007) and analyse the resulting range of uncertainties with bounding box, consensus and probabilistic methodologies rather than lining up with a single modelling outcome (Araújo and New 2007, Thuiller 2007). BIOMOD offers such a platform for ensemble forecasting (Fig. 1) using freeware and open-source R software (R Development Core Team 2008). It overcomes some of the limitations of existing software (e.g. being able to fit and compare different models) and incorporates several features for testing models (e.g. k-fold cross validation) and for examining species-environment relationships (e.g. using randomization tests) (Fig. 2).

Earlier implementations of BIOMOD (Thuiller 2003, 2004) provided limited ensemble simulations across model classes (i.e. four modelling techniques) and boundary conditions (i.e. up to five climate scenarios). Currently, BIOMOD enables larger simulations across initial conditions (i.e. by randomly re-sampling species distribution data and fitting different models for each sample), nine model classes (generalised linear models (GLM, McCullagh and Nelder 1989), generalised additive models (GAM, Hastie and Tibshirani 1990), multivariate adaptive regression splines (MARS, Friedman 1991), classification tree analysis (CTA, Breiman et al. 1984), mixture discriminant analysis (MDA, Hastie et al. 1994), artificial neural networks (ANN, Ripley 1996), generalised boosted models (GBM, Ridgeway 1999), random forests (Breiman 2001), and one rectilinear envelope similar to BIOCLIM (SRE, Busby...
1991), a variable number of model parameterizations (e.g., polynomials and smoothing splines of different orders in general linear or additive models, nodes in classification trees, hidden layers in neural nets), and a virtually unlimited number of boundary conditions. Most modelling techniques implemented in BIOMOD require that species distribution data are presence and absence. When data are presence-only, a simple solution is to generate random pseudo-absences. This can be done in BIOMOD using strategies of increasing complexity.

Figure 1. Probabilistic approach for forecasting species potential distributions (adapted from Thuiller 2007).

Figure 2. Schematic representation of the modelling procedure in BIOMOD.
Evaluation of models in BIOMOD includes two sorts of analysis: assessments of the goodness-of-fit (= explanatory power) and of model accuracy (= predictive power). The former uses standard approaches associated with each technique; for example, ANOVA decomposition and AIC are available for both GLM and GAM, whereas rate of misclassification is used for CTA. The latter can be performed with three different procedures: the area under the relative operating characteristic curve (AUC, Hanley and McNeil 1982), Cohen’s K (Monserud and Leemans 1992), and the Cohen’s K statistic.
techniques are model-specific, so are the conclusions that one may extract from them. To overcome this limitation, BIOMOD uses a randomisation procedure to estimate the importance of each variable. The procedure is independent of the modelling technique, thus enabling direct comparison across models. This procedure uses Pearson correlation between the standard predictions (i.e. fitted values) and predictions where the variable under investigation has been randomly permuted. If the correlation is high, i.e. it is showing little difference between the two predictions, the variable permuted is considered not important for the model. This is repeated a user-defined number of times for each variable, and the mean correlation coefficient over the runs is kept. BIOMOD then gives a ranking of the variables for each of the model selected.

Finally, when projecting potential distributions of species into future environmental conditions, different dispersal assumptions can be made: no dispersal; unlimited dispersal; and user-defined species-specific dispersal. Measures of temporal turnover in potential species richness can then be calculated for each period (Thuiller et al. 2005), as well as species habitat change (Fig. 3a), and visualized according to the different models used (Fig. 3b) to emphasize the potential uncertainty coming from modelling technique or climate change scenarios’ choice.

The BIOMOD R-package and a detailed user’s guide of BIOMOD is available at the R-Forge website <biomod.r-forge.r-project.org>. To cite BIOMOD, or acknowledge its use, cite this Software Note as follows, substituting the version of the application that you used for “Version 0”:


Acknowledgements – BIOMOD is currently developed with funding from FP6 EU MACIS (No. 044399 SSPI) and ECOCHANGE (GOCE-CT-2003-506675) projects; MBA is also funded by FBBVA BIOIMPACTO project. We are indebted to the R community for their valuable contribution to the development and improvement of R code and libraries.

References

Hastie, T. J. and Tibshirani, R. 1990. Generalized additive models. – Chapman and Hall.