Software notes

Ecography 33: 637–642, 2010
doi: 10.1111/j.1600-0587.2010.06416.x
© 2010 The Authors. Journal compilation © 2010 Ecography
Subject Editor: Thiago Rangel. Accepted 7 March 2010

ModEco: an integrated software package for ecological niche modeling

Qinghua Guo and Yu Liu

Q. Guo (qguo@ucmerced.edu), Sierra Nevada Research Inst., School of Engineering, Univ. of California at Merced, CA 95343, USA. – Y. Liu, Inst. of Remote Sensing and Geographical Information Systems, Peking Univ., Beijing, P.R. China.

ModEco is a software package for ecological niche modeling. It integrates a range of niche modeling methods within a geographical information system. ModEco provides a user friendly platform that enables users to explore, analyze, and model species distribution data with relative ease. ModEco has several unique features: 1) it deals with different types of ecological observation data, such as presence and absence data, presence-only data, and abundance data; 2) it provides a range of models when dealing with presence-only data, such as presence-only models, pseudo-absence models, background vs presence data models, and ensemble models; and 3) it includes relatively comprehensive tools for data visualization, feature selection, and accuracy assessment.

With the increasing availability of ecological data (Graham et al. 2004, Wieczorek et al. 2004), environmental niche modeling has gained much attention for a wide variety of ecological applications (Feria and Peterson 2002, Chefaoui et al. 2005, Guo et al. 2005, Thuiller et al. 2005a, Pearce and Boyce 2006). Conventional GIS packages are useful for data management, collection, visualization, and spatial analysis, but they lack advanced statistical approaches, particularly methods that are relevant for modeling the distribution of species. While statistical packages are capable of analyzing and modeling species data with a variety of modeling techniques, the visualization and GIS data support are often poor and require a steep learning curve for users (Wielanda et al. 2006). There are many environmental niche modeling packages available; for example, MaxEnt (Phillips et al. 2006), and GARP (Stockwell and Peters 1999). Existing comparisons between different niche models do not show consistent conclusions (Lek et al. 1996, Mastrorillo et al. 1997, Stockwell and Peterson 2002, Elith et al. 2006, Graham et al. 2006, Stockman et al. 2006) in part due to the fact that the comparisons were primarily conducted on different platforms, which could implement the training and testing differently. Therefore, there is a need to develop an integrated platform to model species distribution data. In this software note, we present software for species data analysis and modeling (referred to as ModEco). The unique features of ModEco are: 1) it includes relatively comprehensive tools for dealing with different types of species data. ModEco contains models for dealing with presence-only data, presence and absence data, and abundance data (continuous values). Specifically, for dealing with presence-only data that are very common in ecological observation data, ModEco includes four types of models, namely, presence-only models, pseudo-absence data models, background vs presence data models, and ensemble models. 2) ModEco provides a user friendly interface that allows users to explore species data with ease. Functions that ModEco provides include: a) environmental and species occurrence data management and visualization; b) feature analysis and selection, such as factor importance analysis, comparison of selected environmental features vs background distribution, principal component analysis; and c) model performance evaluation and accuracy assessment to report associated uncertainty of the results, such as maximum kappa, error matrix, Receiver Operating Characteristic (ROC), and Area under a ROC curve (AUC), true positive rate vs fractional predicted area, which can be used to evaluate model performance based on the characteristics of the species data.

It should be noted that there are several other existing software packages. A recent excellent example is BIOMOD (Thuiller et al. 2009), which implements a range of ecological niche models in R. Openmodeller is another niche model platform that includes multiple niche models (Muñoz et al. 2009). Although extensive evaluation of the features in BIOMOD, Openmodeller, and ModEco is beyond the scope of this note, several distinguished features in ModEco are described as follows. 1) Compared to BIOMOD and Openmodeller, ModEco provides better support for different types of ecological observation data.

Both Openmodeller and BIOMOD can take presence-only and presence-absence data as input data, but not abundance data. 2) In terms of available niche models, both BIOMOD and ModEco include a relatively comprehensive set of advanced machine learning algorithms. ModEco contains more model types in dealing with species occurrence data such as presence-only model, pseudo-absence model, background-based model, and ensemble models. 3) ModEco also provides more utilities in feature selection, generating pseudo-absence data and threshold selection for binary predictions. 4) In terms of ease to use, ModEco and Openmodeller are more user friendly than BIOMOD. However, it should be noted that each software package has its own advantages and limits. For example, BIOMOD can easily incorporate more state-of-the-art machine learning algorithms in R. Considering the rapid development of species distribution modeling techniques and their applications, as well as the relatively early stage of these software packages (OpenModeller ver. 1.0, BIOMOD ver. 0), we believe ModEco could provide an important addition to the existing effort of integrating a range of niche models in the same software platform, which could result in better knowledge about the modeling technique (Santana et al. 2008).

ModEco is implemented using Visual C++; the overall interface of ModEco is shown in Fig. 1. A simplified diagram of the major components of the object-oriented architecture of ModEco is illustrated in Fig. 2. Specific functions are discussed below.

1. Data management

ModEco uses an XML (Extensible Markup Language) as the project file to store and manage the data layers and models' parameters used in ModEco. The project file has three main components: environmental data groups, species data points, and result maps. Descriptions of each component of the project file are given in the following subsections.

a) Environmental data group

An environmental data group is a set of raster environmental data with the same projection, while spatial extent and spatial resolution of the environmental layers do not need to be the same. In ModEco, when dealing with layers with different spatial extents, the minimum extent will be used for the final output; when dealing with layers with different resolutions, users can choose between: 1) a customized resolution, 2) the minimum resolution of the input layers, or 3) the maximum resolution of the input layers. The environmental data group will support environmental data storage for a certain species in different periods. For example, assume one is interested in evaluating the current species distribution and predicting its future distribution under climate change (Thuiller et al. 2005b). ModEco provides a way to store multiple environmental data groups such as current environmental layers and projected future environmental layers so that the prediction of future species distribution can be easily implemented using the same niche

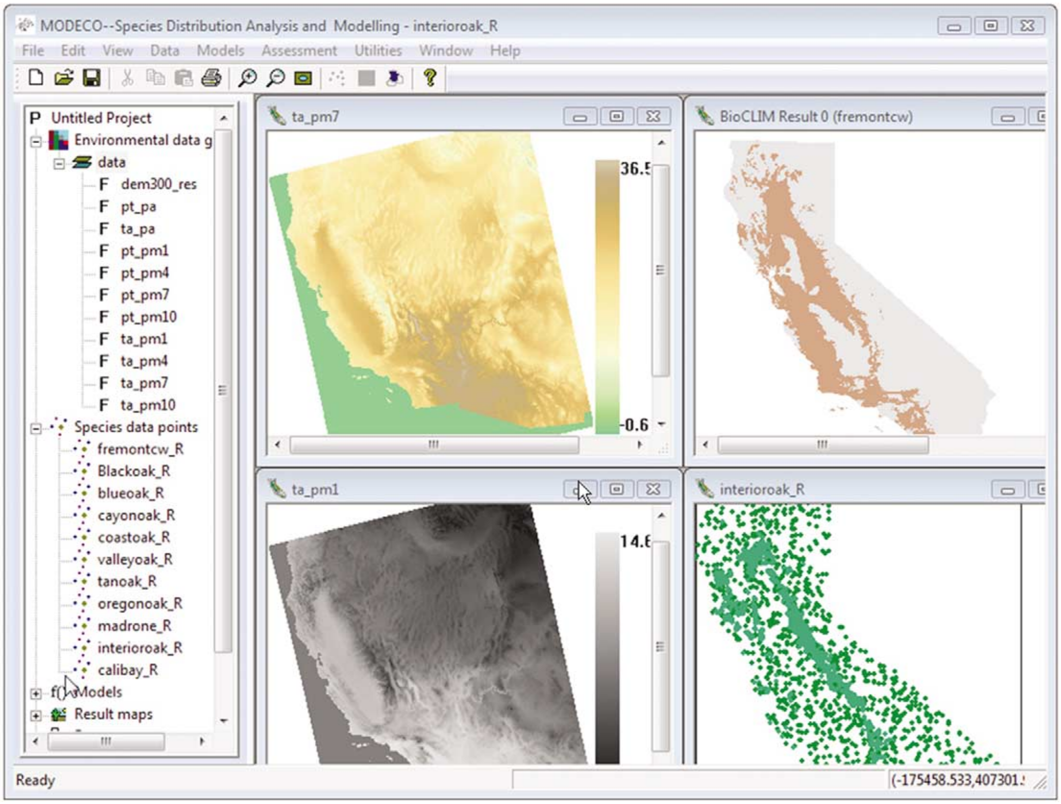


Figure 1. Examples of the graphical user interface of ModEco. The left panel shows the data and model components in a project; the right panel visualizes the different data, such as environmental data, species distribution points, and prediction maps.

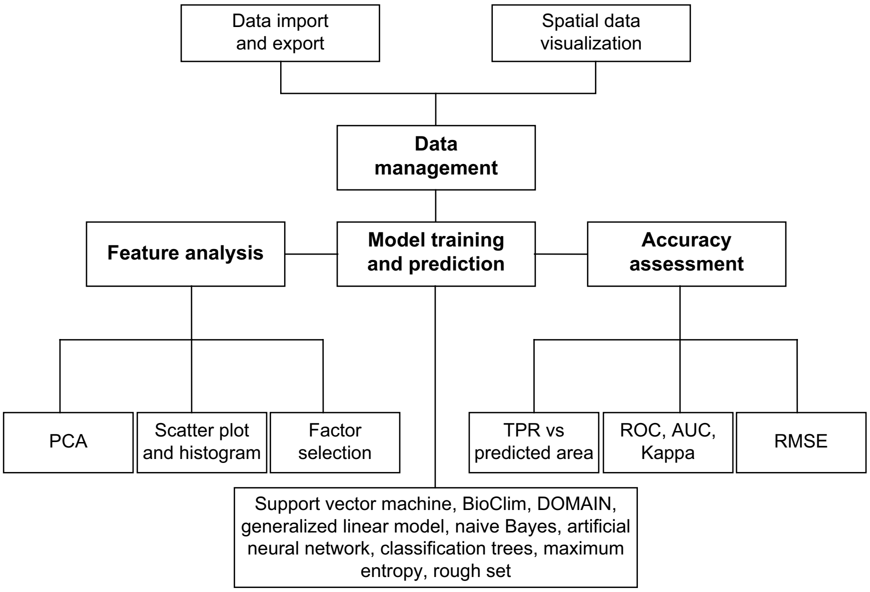

Data import and export
Spatial data visualization
Data management
Feature analysis
Model training and prediction
Accuracy assessment
PCA
Scatter plot and histogram
Factor selection
TPR vs predicted area
ROC, AUC, Kappa
RMSE
Support vector machine, BioClim, DOMAIN, generalized linear model, naive Bayes, artificial neural network, classification trees, maximum entropy, rough set

Figure 2. Component diagram showing the overall structure of ModEco.

model trained by the current species distribution and current environmental data. In ModEco, the metadata for each environmental layer are required to check the consistency between two environmental data groups, which ensures the units of the same variable in the two time periods are the same. Another important feature of the environmental data group is that categorical data are also supported. When users import the data, they can specify the data as categorical data, and ModEco has built-in functions to process those data to meet the needs of different niche models. Some models, such as maximum entropy and classification trees, can naturally take categorical data as input parameters, while others, such as Bioclim and the regression approach, need additional processes to convert the categorical data into a form that those models can use. For example, in regression analysis, categorical data could be converted into multiple dummy variables (0, and 1).

b) Species data

In order to deal with different types of observed species data, ModEco can support presence-only data, presence/absence data, and abundance data. ModEco provides functions to import different types of data in both the text and ESRI shape file formats. Because the model selection and many other implementations of niche modeling depend highly on the type of input species data (such as niche models and accuracy assessment), ModEco tracks the species data types to help users select the niche models suitable for their data. Data sets from multiple species are also supported by ModEco, and can be modeled sequentially in a batch mode. This function is particularly useful for modeling many species with the same set of environmental layers.

c) Models

A range of niche models, such as BioClim, Domain, generalized linear model, classification tree and regression (CART), artificial neural networks (ANN), maximum entropy (MAXENT), support vector machine (SVM), naive Bayes (Duda et al. 2001), and rough sets (Pawlak 1991), have been implemented in ModEco. One useful feature is that users can save the particular model that contains the parameters used in the model training. Consequently, the same model could be used to predict the species distribution in different geographic spaces or under future climate scenarios. Detailed model implementation will be discussed in the modeling and training section.

d) Prediction results

ModEco stores two types of prediction result data in raster format. The first type of data is Boolean (e.g. absence or presence); the second type is continuous raster data including the suitability index for presence-only data, the probability of species occurrence ranging from 0 to 1 for presence and absence data, and the abundance value as a real number for abundance data. All the resulting maps can be exported to generic binary and ASCII raster formats, which can then be easily imported into other GIS software packages (e.g. ESRI ArcGIS) for further data analysis and visualization if needed.

2. Feature analysis

Before users start to predict species distribution, it is important to examine input environmental and species data. ModEco provides functions that allow users to visualize the relationship between the observed species localities and environmental features. Functions include the factor histogram and scatter plot, and factor importance analysis.

Factor histogram analysis is designed to compare the frequency distributions of environmental variables between the observed species localities and the whole study area. If the environmental factor histograms of the observed species follow a pattern similar to the background distribution, it could indicate that this environmental variable may not be relevant to determine the species distribution at the scale of interest (the size of the study area). The scatter plot is another graphical tool to evaluate the ability of two selected environmental factors to discriminate the species distribution. If the presence and absence data can be easily separated in the scatter plot, then these two environmental factors have the potential to discern the presence data from the absence data, and consequently they could be included in the niche model to improve model performance.

Factor selection (also referred to as variable and feature selection) is used to select environmental layers that are most predictive of species distribution. In ModEco, one variable at a time is used to evaluate its performance according to a particular selection metric (e.g. Kappa values) (Forman 2003). Note that results from factors importance analysis should be interpreted with caution if there is strong correlation among the variables. In order to reduce the multi-collinearity issue, ModEco also implements principal component analysis and functions that enable users to group certain environmental layers together before factor selection.

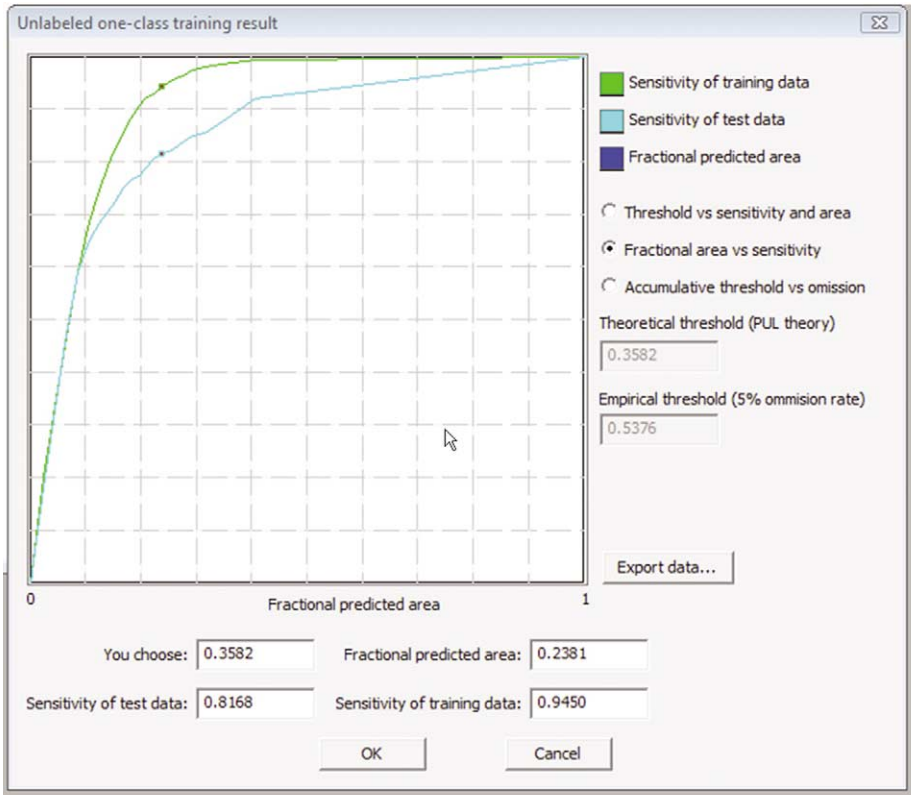
3. Model training and prediction

ModEco incorporates a range of environmental niche models in dealing with presence-only, presence and absence, and abundance data. Specifically, as the observation data of many species contain presence-only data, ModEco provides three types of model solution: a) presence-only model, such as Bioclim, Domain, and one-class SVMs; b) pseudo-absence model, such as two-class SVMs, maximum entropy, generalized linear model (GLM), artificial neural networks (ANN), classification tree, rough sets, and naive Bayes classification. In addition to commonly used pseudo-absence models, an iterative approach was also implemented in ModEco, i.e. after model prediction based on pseudo-absence data, ModEco generates the pseudo-absence data from the absences area of the prediction map and re-runs the model until the final prediction results are stabilized (Yu 2005). The third model solution is the implementation of background vs presence data models. Recent studies have demonstrated that background-based models are promising in dealing with presence-only data (Elith et al. 2006, Phillips et al. 2009). In ModEco, we implemented maximum entropy, support vector regression, GLM, and ANN as background-based models. Instead of using conventional pseudo-absence data generated from regions outside the presence data, the background-based models sample the ''pseudo-absence data'' from the whole study area, which results in certain types of conditional probability depending on the models used (Phillips and Dudík 2008). Moreover, ModEco also implements sample selection bias that allows users to provide the biased background to improve model performance (Phillips et al. 2009). Thresholds need to be applied in order to generate the binary output (i.e. presence and absence). Two methods are implemented in ModEco with respect to threshold selection (Fig. 3): one method is based on the empirical accumulative probability (e.g. 95%) and the other is derived from a statistically proven theory (Elkan and Noto 2008). Both methods need a validation dataset, which normally is the subset of the training data (e.g. 25%). The empirical accumulative probability method seeks to find the threshold that corresponds to a certain percentage of accumulative probability (e.g. 95%), while the latter method seeks to derive the threshold based on the theory of positive and unlabeled learning algorithms, which has shown promise in one-class classification when background information (i.e. unlabeled data) is available (Noto et al. 2008). In addition, recent research has found that ensemble models (i.e. combining several model outputs) are promising alternatives to overcome the variability of model selection on prediction results (Araújo and New 2006). In ModEco, methods that combine different niche modeling outputs include: 1) unweighted simple average of model predictions; and 2) weighted average based on user specified accuracy (e.g. AUC). It is worth mentioning that, for presence-only data, the ensemble models should only be used to combine binary outputs (presence and absence) instead of continuous outputs since the continuous outputs may have different meanings from different models, so they cannot be simply added together without appropriate adjustments. Note that the set of models proposed for inclusion in ModEco is not comprehensive. Nevertheless, the object-oriented design of the software facilitates the incorporation of additional models with relative ease once those models are available. The implementation of those models is based on either our own programs or open source codes with extensive testing and validation. For example, the SVM implementation is based on LIBSVM (Chang and Lin 2001). A summary of the available models in ModEco is given in Table 1.

4. Accuracy assessment

ModEco includes a range of accuracy assessment methods, which are important for evaluating model performance and model selection. In addition, comparisons between different methods may suffer from accuracy assessment methods that are implemented in different software platforms, which may introduce unnecessary bias to the comparisons. Accuracy assessment methods available in ModEco are cross-validation accuracy, ROC, AUC, error matrix, and maximum Kappa values. These assessments are commonly used as standard measures to evaluate the performance of environmental niche models (Wiley et al. 2003, Elith et al. 2006). In addition, for presence-only data, the above-mentioned measures are not applicable since they all require true absence data. One possible solution is to plot the true positive rate (TPR) vs the factional prediction area (FPA) as a proxy for true positive rate vs false positive rate and the area under TPR vs FPA (Guo et al. 2005, Phillips et al. 2006). For abundance data, the accuracy is measured by standard root mean square errors for the predicted values against the testing data. Below are brief descriptions of the accuracy assessment methods.

1) Cross-validation accuracy assessment is achieved by first randomly splitting the training data into *n* subsets of equal size, and then each subset is used in turn for accuracy assessment and the remaining $n-1$ subsets are used for training. Finally, the total accuracy is estimated by averaging the accuracies of each subset.



Unlabeled one-class training result
🟩 Sensitivity of training data
🟦 Sensitivity of test data
🟪 Fractional predicted area
0
Fractional predicted area
1
Export data...
OK
Cancel

○ Threshold vs sensitivity and area
● Fractional area vs sensitivity
○ Accumulative threshold vs omission

Theoretical threshold (PUL theory): 0.3582

Empirical threshold (5% ommision rate): 0.5376

You choose: 0.3582

Fractional predicted area: 0.2381

Sensitivity of test data: 0.8168

Sensitivity of training data: 0.9450

Figure 3. Accuracy assessment for presence-only data using the background vs presence data model. ROC curves for both training data and testing data are generated based on true positive rate vs fractional predicted area. Thresholds are recommended for the users to convert the probabilistic output from the model into a binary result (i.e. presence and absence).

2) The ROC curve is a plot of the sensitivity (true positives rate) vs 1-specificity (true negative rate) by varying discrimination thresholds. One advantage of the ROC curve is that it does not depend on a specific threshold. Comparisons between different ROC curves often need to calculate the AUC values.

3) The error matrix is a common measure for classification accuracy (Congalton and Green 1999). An error matrix can then be computed by comparing the prediction map with respect to the observed point layer, and from this total classification accuracy and Kappa values can be computed. The Kappa value takes into consideration the effects of random change on accuracy assessment, and it is often used in evaluating the performance of niche models (Loiselle et al. 2003, Elith et al. 2006). The maximum Kappa value is calculated by iteratively selecting the parameters until the model reaches its maximum Kappa value. Therefore, the maximum Kappa value can be considered as the best possible accuracy achieved by the model with a specific set of parameters.

4) For real presence-only data, which are very common in ecological observation data, the aforementioned accuracy measures are not applicable. Engler et al. (2004) proposed that a good model prediction with presence-only data should predict a potential area as small as possible while still covering a maximum number of the species occurrences. Guo et al. (2005) demonstrated this concept for selecting parameters of one-class SVM in modeling the potential distribution of a tree disease in California. ModEco allows users to plot true positive rate vs fractional prediction area, aiding users to select appropriate parameters for the model or to select thresholds to convert continuous outputs into dichotomous classifications of presence and absence. In addition, ModEco also reports an AUC value based on presence and the factional prediction area curve. The AUC value here is interpreted as a measure of the

Table 1. Models implemented in ModEco.

Model types	Models	Accuracy assessment
Presence-only	BioCLIM, DOMAIN, One-class SVM	True positive rate (TPR) vs factional predicted area (FPA), Area under TPR and FPA
Pseudo-absence	SVM, naive Bayes, ANN, GLM, MaxEnt, rough set	TPR vs FPA, Area under TPR and FPA
Presence/absence	SVM, naive Bayes, ANN, GLM, MaxEnt, rough set	AUC, ROC, error matrix, and Kappa
Background-based	SVM, MaxEnt, GLM	TPR vs FPA, Area under TPR and FPA
Ensemble	Two weighting methods: average weighting, and weighted by user specified accuracy (e.g. AUC); and two combination methods: additive, and productive	AUC, ROC, error matrix, Kappa or TPR vs FPA, area under TPR and FPA
Abundance	SVM, linear regression, and GLM	Root mean square errors (RMSE)

ability of the classifier to discriminate between presence and background as opposite to discriminating between presence and absence from traditional AUC (Phillips et al. 2006). Note that several matrices for accuracy assessments are also used to tune the model parameters during the model training process. The differences are that for tuning model parameters, these metrics are used on the training dataset, while for assessing model performance they are used on the validation dataset.

The ModEco package and a detailed user's guide are available at the website <gis.ucmerced.edu/ModEco>.

To cite ModEco or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for "Version 0":

Guo, Q. and Liu, Y. 2010. ModEco: an integrated software package for ecological niche modeling. – Ecography 33: 637–642 (Version 0).

Acknowledgements – ModEco is partially supported by the National Science Foundation (BDI-0742986) and BioGeomancer project from the Gordon and Betty Moore Foundation. The development of ModEco benefited from discussion and comments from Craig Moritz and John Wieczorek at the Museum of Vertebrate Zoology at UC Berkeley. We are grateful to Charles Elkan for the insights on his positive and unlabeled algorithm, and Wenkai Li, Robert Klinger, Michelle Koo, Adam Smith, and Erin Conlisk for their helpful comments on the early version of the software. We also thank Otto Alvarez, Hong Yu, Miguel Fernandez, Andrew Zumkehr, and Jacob Flanagan for help on software debugging and website maintenance.

References

Araújo, M. and New, M. 2006. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Chang, C. and Lin, C. 2001. LIBSVM: a library for support vector machines. – <www.csie.ntu.edu.tw/~cjlin/libsvm>, accessed 10 Feb 2010.

Chefaoui, R. M. et al. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian copris species. – Biol. Conserv. 122: 327–338.

Congalton, R. and Green, K. 1999. Assessing the accuracy of remotely sensed data: principles and practices. – CRC Press.

Duda, R. O. et al. 2001. Pattern classification. – Wiley.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Elkan, C. and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. – Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 213–220.

Engler, R. et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. – J. Appl. Ecol. 41: 263–274.

Feria, T. P. and Peterson, A. T. 2002. Prediction of bird community composition based on point-occurrence data and inferential algorithms: a valuable tool in biodiversity assessments. – Divers. Distrib. 8: 49–56.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. – J. Mach. Learn. Res. 3: 1289–1305.

Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.

Graham, C. H. et al. 2006. Habitat history improves prediction of biodiversity in rainforest fauna. – Proc. Natl Acad. Sci. USA 103: 632–636.

Guo, Q. et al. 2005. Support vector machines for predicting distribution of sudden oak death in California. – Ecol. Model. 182: 75–90.

Lek, S. et al. 1996. Application of neural networks to modelling nonlinear relationships in ecology. – Ecol. Model. 90: 39–52.

Loiselle, B. A. et al. 2003. Avoiding pitfalls of using species distribution models in conservation planning. – Conserv. Biol. 17: 1591–1600.

Mastrorillo, S. et al. 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. – Freshwater Biol. 38: 237–246.

Muñoz, M. E. d. S. et al. 2009. OpenModeller: a generic approach to species' potential distribution modelling. – Geoinformatica, doi:10.1007/s10707-009-0090-7.

Noto, K. et al. 2008. Learning to find relevant biological articles without negative training examples. – Lect. Not. Comput. Sci. 5360: 202–213.

Pawlak, Z. 1991. Rough sets: theoretical aspects of reasoning about data. – Kluwer.

Pearce, J. L. and Boyce, M. S. 2006. Modelling distribution and abundance with presence-only data. – J. Appl. Ecol. 43: 405–412.

Phillips, S. and Dudík, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. – Ecography 31: 161–175.

Phillips, S. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.

Phillips, S. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.

Santana, F. S. et al. 2008. A reference business process for ecological niche modelling. – Ecol. Inform. 3: 75–86.

Stockman, A. K. et al. 2006. An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. – Divers. Distrib. 12: 81–89.

Stockwell, D. and Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. – Int. J. Geogr. Inform. Sci. 13: 143–158.

Stockwell, D. and Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. – Ecol. Model. 148: 1–13.

Thuiller, W. et al. 2005a. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. – Global Change Biol. 11: 2234–2250.

Thuiller, W. et al. 2005b. Climate change threats to plant diversity in Europe. – Proc. Nat. Acad. Sci. USA 102: 8245–8250.

Thuiller, W. et al. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373 (Version 0).

Wieczorek, J. et al. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. – Int. J. Geogr. Inform. Sci. 18: 745–767.

Wielanda, R. et al. 2006. Spatial analysis and modeling tool (SAMT): 1. Structure and possibilities. – Ecol. Inform. 1: 67–76.

Wiley, E. O. et al. 2003. Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. – Oceanography 16: 120–127.

Yu, H. 2005. Single-class classification with mapping convergence. – Mach. Learn. 61: 49–69.