00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

Ecography 39: 368–375, 2016
doi: 10.1111/ecog.01881

Subject Editor: Edmund Hart. Editor-in-Chief: Michael Borregaard. Accepted 27 January 2016

sdm: a reproducible and extensible R platform for species distribution modelling

Babak Naimi and Miguel B. Araújo

B. Naimi (naimi.b@gmail.com) and M. B. Araújo, Imperial College London, Silwood Park, Buckhurst Road, Ascot, Berkshire, SL5 7PY, UK, and Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, Univ. of Copenhagen, Denmark. MBA also at: Dept of Biogeography and Global Change, National Museum of Natural Sciences, CSIC, c/Jose Gutierrez Abascal, ES-28006 Madrid, Spain, and InBio-CIBIO, Univ. of Évora, Largo dos Colegiais, PT-7000 Évora, Portugal.

sdm is an object-oriented, reproducible and extensible, platform for species distribution modelling. It uses individual species and community-based approaches, enabling ensembles of models to be fitted and evaluated, to project species potential distributions in space and time. It provides a standardized and unified structure for handling species distributions data and modelling techniques, and supports markedly different modelling approaches, including correlative, process-based (mechanistic), agent-based, and cellular automata. The object-oriented design of software is such that scientists can modify existing methods, extend the framework by developing new methods or modelling procedures, and share them to be reproduced by other scientists. sdm can handle spatial and temporal data for single or multiple species and uses high performance computing solutions to speed up modelling and simulations. The framework is implemented in R, providing a flexible and easy-to-use GUI interface.

Species distributions models (SDMs), also known as bioclimatic envelope models, ecological niche models and habitat suitability models, explore the relationship between geographical occurrences of species and corresponding environmental variables (Guisan and Zimmermann 2000, Peterson et al. 2011). SDMs are widely used in a range of fields and applications including regional biodiversity assessments, spatial conservation prioritization, evolutionary biology, epidemiology, global change biology, and wildlife management (Araújo and Peterson 2012). There are several SDM techniques available. They differ in their ability to summarize the relationships between response and predictor variables (Segurado and Araújo 2004, Elith et al. 2006), and when used for transferring the distributions of species into different geographical (Randin et al. 2006) or temporal contexts (Thuiller et al. 2004, Araújo et al. 2005b, Pearson et al. 2006) projections can vary startlingly among techniques.

SDMs also vary with regards to the type of response variables used (e.g. presence and absence versus presence only), the types of predictor variables handled (e.g. continuous versus categorical), the type of output provided (e.g. probabilities, continuous indices of suitability, or binary predictions of presence and absence), the type of species–environment relationship assumed (e.g. simple linear to complex nonlinear), the approach used to estimate species distributions (e.g. parametric versus nonparametric approaches), and the approach to select relevant predictor variables (e.g. whether predictor contributions are weighted, and whether they allow for interactions among variables) (Segurado and Araújo 2004, Elith et al. 2006, Austin 2007, Naimi et al. 2011, Peterson et al. 2011).

The outputs of SDMs are sensitive to the specific rules used to parameterize them. When models are implemented in different platforms, rules used to fit them may not be comparable. For example, Domain (Carpenter et al. 1993), DesktopGARP (Stockwell and Peters 1999), and Maxent (Phillips et al. 2006) are typically implemented with different off-the-shelf software making cross-model comparisons challenging. Models are also generally implemented following different protocols for pre-processing of data and post-processing of the results, even when they are implemented within the same computer platform. Given the difficulties in comparing the results of different models, conclusions from model comparison studies are difficult to generalise beyond the specific case studies (Segurado and Araújo 2004, Elith et al. 2006).

An integrated framework enabling multiple SDMs to be fitted and compared simultaneously is required to move the field of species distribution modeling forward. Three off-the-shelf software including openModeller (de Souza Muñoz et al. 2009), BIOENSEMBLES (Diniz-Filho et al. 2009), and ModeEco (Guo and Liu 2010) have been independently developed to provide such frameworks. They enable several modelling algorithms to be fitted simultaneously and they perform the most common tasks related to species

distribution modelling (e.g. data evaluation, prediction). They also provide graphical user interface (GUI) making them click-and-run software and particularly friendly to users with less computational expertise. Simultaneously, they provide limited flexibility as users can only use algorithms, model comparison and evaluation procedures that are implemented therein. Moreover, insufficient understanding of what such click-and-run software is doing and how they were implemented makes users fret over whether they are doing what is expected (Joppa et al. 2013).

R (R Development Core Team) is a general-purpose high-level programming language and a free (under the GNU general public license) open source environment. It is widely used for statistical analysis and graphical visualization and, recently, its suitability for mathematical computing (Soetaert et al. 2010), manipulation and analysis of complex spatial data sets and modelling (Bivand et al. 2008) has increased. R can be extended through user-created packages, which allow developing new and specialized analytical techniques, graphical devices, import/export capabilities, reporting tools, etc. Growing collections of tools are explicitly being developed to bridge R and the known modelling software (Naimi and Voinov 2012). All of these capabilities make R very powerful. Despite the advantages of R, there are also some disadvantages. R involves a steep learning curve preventing beginners and script-avert scientists from taking advantage of its capabilities. Moreover, different packages are not equivalent regarding their computational efficiency (García-Callejas and Araújo unpubl.) or capability for handling errors. Sometimes they simply do not work under given circumstances and users have to struggle with errors and bugs. When users apply and compare alternative models, it becomes difficult to keep track of the syntactical nuances implemented in different packages (Kuhn 2008).

R provides an increasing number of packages for modelling (e.g. gbm, gam, maxlike, deSolve, simecol). At least two R platforms have been developed for fitting (e.g. BIOMOD and dismo; Thuiller et al. 2009, Hijmans and Elith 2013) and processing of species distributions modelling outputs (e.g. SDMTools; VanDerWal et al. 2011). BIOMOD (Thuiller et al. 2009), including its recent version biomod2, offers several functions for ensemble modelling of species distributions (Araújo and New 2007). The other package, dismo (Hijmans and Elith 2013), can be used to fit several SDMs including maxent (Phillips et al. 2006) in R, and facilitates using common spatial data in the procedure of modelling and predicting species distributions. However, it does not support fitting and comparison of multiple SDMs as in BIOMOD.

BIOMOD and dismo combine a limited number of packages and modelling techniques. Even if technically feasible to add more techniques into these platforms, the task is beyond reach by most users. The platforms also lack convenient GUI interfaces, thereby being unpalatable to users with very basic knowledge of R. More importantly, because implementation of the different techniques is not standardized, lessons learned from comparing outputs of different SDMs are impaired. Are results of a particular modelling technique better because the technique is superior to other, or because of particular default implementations in the software? Developing model-independent methods (i.e. procedures that can be applied with any SDM) for common tasks in species distribution modelling (e.g. variable selection, variable importance) followed with a good software design would override such shortcomings for comparing modeling outputs in existing SDM platforms.

We introduce a new R package, sdm, that solves the limitations of existing platforms for species distributions modelling. sdm an extendable framework that enables fitting of individual and community-based SDM approaches, while supporting markedly different modelling approaches, including correlative, process-based (mechanistic), agent-based, and cellular automata. It generates ensembles of models, and several options for evaluation of model results and and projection of species potential distributions in space and time. The generic design of sdm is object-oriented making it flexible and amenable to efficient handling of errors. The object-oriented design also makes it easily extended by users wanting to support additional models and/or procedures for any of the main steps in species distribution modelling. Finally, sdm provides a graphical user interface (GUI) making it easy-to-use even for users who are not familiar with R.

Design of the sdm package

The sdm package is designed to create a comprehensive modelling and simulation framework that: 1) provides a standardised and unified structure for handling species distributions data and modelling techniques (e.g. a unified interface is used to fit different models offered by different packages); 2) is able to support markedly different modelling approaches, including correlative, process-based (mechanistic), agent-based, cellular automata, etc.; 3) enables scientists to modify the existing methods, extend the framework by developing new methods or procedures, and share them to be reproduced by the other scientists; 4) handles spatial as well as temporal data for single or multiple species; 5) employs high performance computing solutions to speed up modelling and simulations, and finally; 6) uses flexible and easy-to-use GUI interface.

sdm was built following a fully object-oriented design. The object-oriented approach enables formulation of problems using interacting objects rather than sets of functions (Alfons et al. 2010). The properties of these objects are defined by general and extensible class description, suitable for species distributions models and their corresponding data. Their behavior and interactions are modeled with generic functions and methods. One of the most important concepts of object-oriented programming is class inheritance, i.e., subclasses inherit properties and behavior from their superclasses. Thus, code can be shared for related classes, which is the main advantage of inheritance (Alfons et al. 2010). In addition, subclasses may have additional properties and behavior, so in this sense they extend their super-classes.

In the sdm framework, we used S4 and reference class systems (Chambers 2014), which provide mechanisms for object-oriented programming in R. The reference class system allows the use of encapsulated object-oriented programming, and their objects behave more like objects in the other object-oriented programming language such as Java and C++. We defined several classes to handle species

data, different methods, and settings for modelling and simulation. There are some container classes whose instances are collections of the methods for a specific purpose (e.g. model fitting, evaluation). These classes are extensible by users (i.e. a new method can be included to the collection by a user). Furthermore, the specific container classes were designed to handle the chain of processes (workflows). They are followed by some methods to facilitate their reproducibility on a new machine (i.e. they can be shared and reproduced by a new user on a new workstation). Reproducibility of an experiment refers to not only its' exact repetition (repeatability), but also using the general idea and settings of the experiment in a new experiment. There is also a class to manage the metadata can be used for both methods and data in the framework. An object of the metadata class keeps some information (e.g. authors, date of creation, citation, and website) about the corresponding data or method. A user can find, for example, how to cite a new data, method, or process that has been created and shared by another user. An example of a data class and a container class in the sdm framework is provided in Fig. 1. A class may contain several subclasses and itself being a subclass of a superclass. A set of methods is defined for each class and can be used to handle the class during the simulation.

How does sdm work?

The sdm framework helps constructing and executing a chain of procedures that constitute the backbone of species distributions modelling. These procedures can be grouped into three steps: pre-processing; processing; and post-processing. Pre-processing includes all procedures by which data becomes available for processing, when SDMs are fitted. After being processed, the model results are post-processed given user-specified settings (Fig. 2). An extensible set of functions (methods) is available for each step, which can be included into the chain by a user.

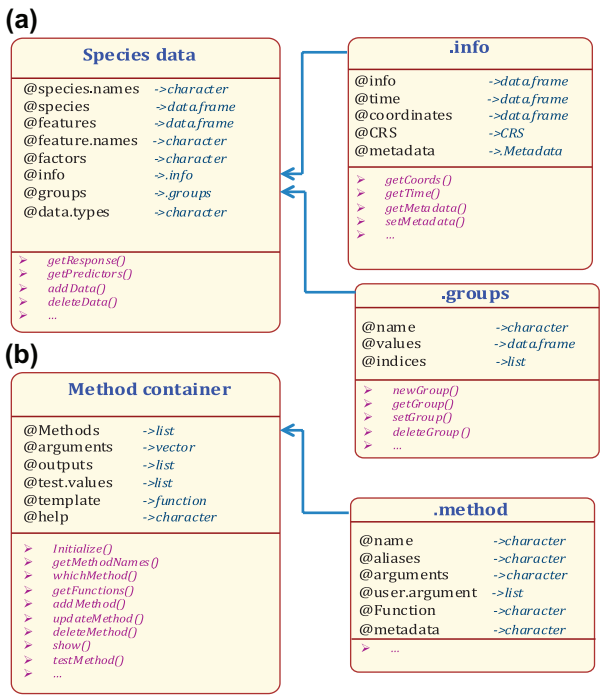

Figure 1. Class diagrams of a species data object (a), and a method container (b); each class contains several data, known as attributes or fields, kept in different slots (@slot-name), and several methods defined as a list of functions to access the data objects in the class.

Data management and pre-processing

A set of utility functions is available in the sdm framework to read and handle species and environmental data in a flexible and automated way. Species data are usually available as a list of coordinates, or as a spatial point dataset. Environmental variables are mostly available as spatial data in the form of spatial vectors (e.g. lines, points, polygons) or rasters (i.e. spatial grids). GIS (Geographic Information Systems) operations are typically required to convert these kinds of data into a structure that is suitable for species distribution modelling. Such process of data manipulation is usually a challenge for non-GIS experts, especially when the data vary in their extent or their coordinate systems. sdm can read species and environmental data with different common structures (spatial or non-spatial), and is not sensitive to these problematic issues as they are automatically handled and fixed through the pre-processing step. For instance, sdm uses several procedures to manipulate data when spatial datasets are introduced as the input data (e.g. species data as spatial points, and environmental predictors as a set of raster datasets) including: checking whether all the data use the same coordinate system, and if not, a project transformation is called to convert them into a unique coordinate system; checking whether they are spatially match and whether they use the same spatial extent, and if not, the extent will be matched and also the records outside of the main extent are recognized.

Data used in species distributions modeling typically carry a number of statistical problems (e.g. lack of absence data, multicollinearity among predictors, spatial autocorrelation in both response and predictor variables, positional uncertainty). Whilst solutions have been proposed to deal with these problems (Dormann 2011), current platforms for SDM tend to ignore them. The pre-processing phase includes all procedures through which data are controlled for problematic issues and prepared for the processing (modeling) phase. These procedures are implemented as functions according to state-of-the-art methods for the corresponding issues. We briefly describe some important procedures.

Pseudo-absence – some models required absences as well as presences to be fitted. Yet all too often presence data alone are available. One option to deal with this problem is to generate pseudo absences. Pseudo absences tend either to be randomly drawn from a studied region, or environmentally or spatially stratified (Barbet-Massin et al. 2012). These procedures for pseudo-absence generation are available in sdm and can be used separately or within the modelling procedure. Furthermore, one can generate several replications of pseudo absences to explore the variability of the process through a simulation.

Collinearity – correlation between two or more predictor variables in a statistical model can cause problems of collinearity (also called multicollinearity). Many statistical models (especially regression-type models) are sensitive to

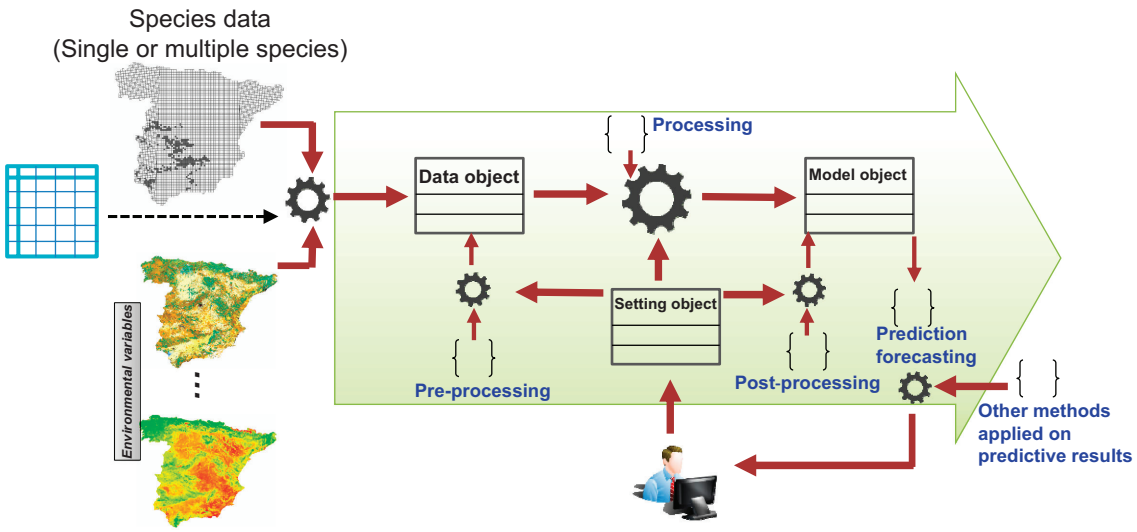

Figure 2. A schematic representation of a chain including the main classes and pre-processing, processing, and post-processing procedures for species distribution modelling in sdm.

collinearity for it may cause instability in parameter estimation and biases in inference statistics (Dormann et al. 2013). Several approaches have been provided in the statistical literature to detect collinearity. Pairwise correlation coefficients and the variance inflation factor (VIF) (Marquardt 1970) are, perhaps, the most widely used approaches. The Pearson (r) or Spearman (ρ) correlation coefficients between a pair of variables can simply show whether two variables are correlated and, if so (usually when its value is greater than a threshold e.g. 0.7), having both variables in the modelling procedure may cause problems of collinearity. The VIF is a more precise method as it measures how strongly each predictor can be explained by the rest of predictors: if all information regarding a predictor is provided by other predictors why keep the predictor? The VIF is based on the square of the multiple correlation coefficient (R^2) resulting from regressing the predictor variable against all other predictor variables. A VIF greater than 10 (as a rule of thumb) is a signal that the model has a collinearity problem (Chatterjee and Hadi 2006). All of the above measures are implemented in sdm and can be used to detect collinearity. To avoid collinearity in the modelling, one approach is to remove the collinear variable prior to model fitting. We developed two stepwise procedures to detect and exclude collinear variables: one based on VIF measure; the other using both correlation coefficients and VIF. The former approach calculates VIF for all predictors and excludes the one with the greatest VIF (if it is greater than a threshold). The procedure is repeated until all strongly collinear variables are excluded. The second approach calculates the correlation coefficients between variables and identifies a strongly correlated pair with the highest coefficient. Then the variable with a highest VIF is excluded from the pair, and the procedure is repeated until no strongly correlated pair remains.

Principle component analysis (PCA) can be used as a data reduction technique to reduce dimensionality in predictor variables (Heikkinen et al. 2006) and is available in sdm.

Positional uncertainty – increasing amounts of species data, especially presence-only data from museum or herbarium collections (Graham et al. 2004) or from volunteer observation networks (Wood et al. 2011), are becoming available on the Internet. One of the problems with these data is the uncertainty regarding the exact position of the occurrence records (Graham et al. 2004, Rowe 2005). Examining spatial autocorrelation in predictor variables is one possible strategy to investigate whether positional uncertainty in species occurrences is problematic (Naimi et al. 2011, 2014). Spatial autocorrelation in predictors can give insight into how similar the nearby locations are to the uncertain species location. Strong spatial autocorrelation indicates that the errors in species locations matter less, because nearby locations have similar environmental characteristics to the true location. Spatial autocorrelation can be measured globally, over the entire study area (e.g. using a variogram; Naimi et al. 2011), or locally at each species location (e.g. using a local spatial autocorrelation measure; Naimi et al. 2014). The former can give insight into the level of positional uncertainty under which the models will be sensitive (by assuming that the spatial structure is the same over the study area), while the latter leads to identify the species locations that are likely to be problematic as a consequence of positional uncertainty. sdm, implements the two methods.

Feature construction – before processing the models, user-defined features are established that determine how species distributions data are related to environmental variables. Several features are available in sdm including linear, quadratic, polynomial, product, hinge, threshold, spline, and factor that can be extended by the user according to the needs. sdm treats features as model-independent, which is an important advantage over other SDM platforms as it makes it possible to use and compare unique set of feature classes across all models (subject to the being supported by modelling algorithm). The ability to set common features across different models helps overcoming one of the main drawbacks of existing model comparisons: not controlling for varying features across models. For example, while Maxent software (Phillips et al. 2006) supports hinge and threshold features in fitting a

maximum entropy algorithm, the other SDM software do not support them.

Processing (model fitting)

Model fitting is a step in modelling species distributions, whereby one or several model(s) is fitted to relate response variables (species distributions) to predictor (environmental) variables. A user can select any (or all) of available methods (modelling algorithms). Several instances of a model may be used with different settings, and/or ensembles of several models can also be generated for each species to generate a consensus among them. Currently, sdm supports 15 modelling methods including generalized linear model (GLM; McCullagh and Nelder 1989), generalized additive model (GAM; Hastie and Tibshirani 1990), classification and regression trees (CART; Breiman et al. 1984), boosted regression trees (BRT; Friedman 2001), multivariate adaptive regression spline (MARS; Friedman 1991), mixture discriminant analysis (MAD; Hastie et al. 1994), random forests (RF; Breiman 2001), support vector machine (SVM; Vapnik 1995), artificial neural networks (ANN; Rosenblatt 1958), environmental niche factor analysis (ENFA; Hirzel et al. 2002), maximum entropy (Maxent; Phillips et al. 2006), maxlike (Royle et al. 2012), Bioclim (Busby 1991), Domain (Carpenter et al. 1993), and Mahalanobis (Farber and Kadmon 2003). Furthermore, several community-based models (Baselga and Araújo 2009) and consensus techniques (Garcia et al. 2012), derived from fitting multiple (i.e. ensembles) of models (Araújo and New 2007), are implemented in sdm. Most of these modelling methods were available through different packages in R (e.g. GAM, BRT, SVM). sdm depends on and uses these packages to fit the models based on such methods that are selected by a user. Several modelling methods (Table 1) as well as all of the procedures in the pre- and post-processing (e.g. multicollinearity test, variable importance, model evaluation), are implemented in the sdm package. The programme also provides some facilitator functions enabling the user to include (and use) new methods as they become available. The new method, or the specific settings for using an existing one, can then be exported and published (for example on Internet) for other users.

Table 1. A list of implemented modeling methods in the first release of the sdm package and their dependent packages.

Modelling methods	Depends on
Generalized linear models (GLM)	stats
Generalized additive models (GAM)	mgcv; gam
Boosted regression trees (BRT)	gbm
Support vector machine (SVM)	kernlab
Classification and regression trees (CART)	tree
Multivariate adaptive regression spline (MARS)	earth
Mixture discriminant analysis (MAD)	mda
Random forests (RF)	randomForest
Artificial neural networks (ANN)	nnet; neuralnet
Environmental niche factor analysis (ENFA)	adehabitatHS
Maximum entropy (maxent)	Java software: maxent.jar
Maxlike	maxlike
Bioclim	NONE
Domain	NONE
Mahalanobis	NONE
Ensemble modelling	NONE
Community-based models	gdm; mda

Post-processing

When the models are fitted, there are several additional processes that can be employed, including model evaluation, prediction, and variable importance assessment. sdm also offers specific functions to analyse geographically the outputs when multiple species are modelled (e.g. calculation of species richness, beta diversity, and niche similarity), or to assess the temporal changes when species records are available in multiple time periods.

Model evaluation (accuracy assessment) – a comprehensive set of model evaluation procedures are implemented in sdm. Ideally, statistically independent data (test data) should be used to evaluate model predictions (Araújo et al. 2005a), otherwise a data-splitting method is often used as an alternative by which a randomly drawn sample of the data are used to train the models and the remaining data are used for model evaluation (but see for alternative approaches, Madon et al. 2013). A one-time data-splitting has been widely used for this purpose, although it may introduce a bias to the parameter estimation (Araújo et al. 2005a). This issue can be overcome by using a family of resampling methods including random subsampling, K-fold cross-validation, Jackknife (leave-one-out), and bootstrapping (Hastie et al. 2009). Subsampling repeats the random data splitting into training and testing proportions K times (uses sampling without replacement). K-fold cross-validation, first, splits the data into K roughly equal-sized parts, and then fits the models K times. Each time one part is used as test data and the other K – 1 parts of the data are used as training data. Leave-one-out is equal to the K-folds cross-validation when K is equal to the number of observations. This means that only one observation is used to evaluate the model at each run. Bootstrapping repeats a sampling with replacement method, each time a sample with equal size as the original data is drawn and used for training data. The observations that are not selected are used for the evaluation at each run. In sdm, all these procedures are implemented and can be used (one or all) in the evaluation procedure. Many state-of-the-art statistics for evaluating SDMs (Fielding and Bell 1997) are implemented that include threshold-dependent statistics (e.g. TSS, Sensitivity, Specificity), threshold-independent statistics (e.g. AUC, COR), and methods developed to calculate p-values through Jackknife for data sets with small sample size (Pearson et al. 2007).

Variable importance and response curve – determining the role of predictor variables in explaining the species distribution is of practical relevance to researchers concerned with interpreting the outputs of the models. Evaluating how important each variable is (Murray and Conner 2009) and/or visualizing the predicted response of species to the predictor variable (Elith et al. 2005) are two known methods to determine predictor variable importance. In sdm, several model-independent techniques were implemented to evaluate the importance of variables and visualize species response curves. In sdm, response curves are generated according to the procedure proposed by Elith et al. (2005). Additional

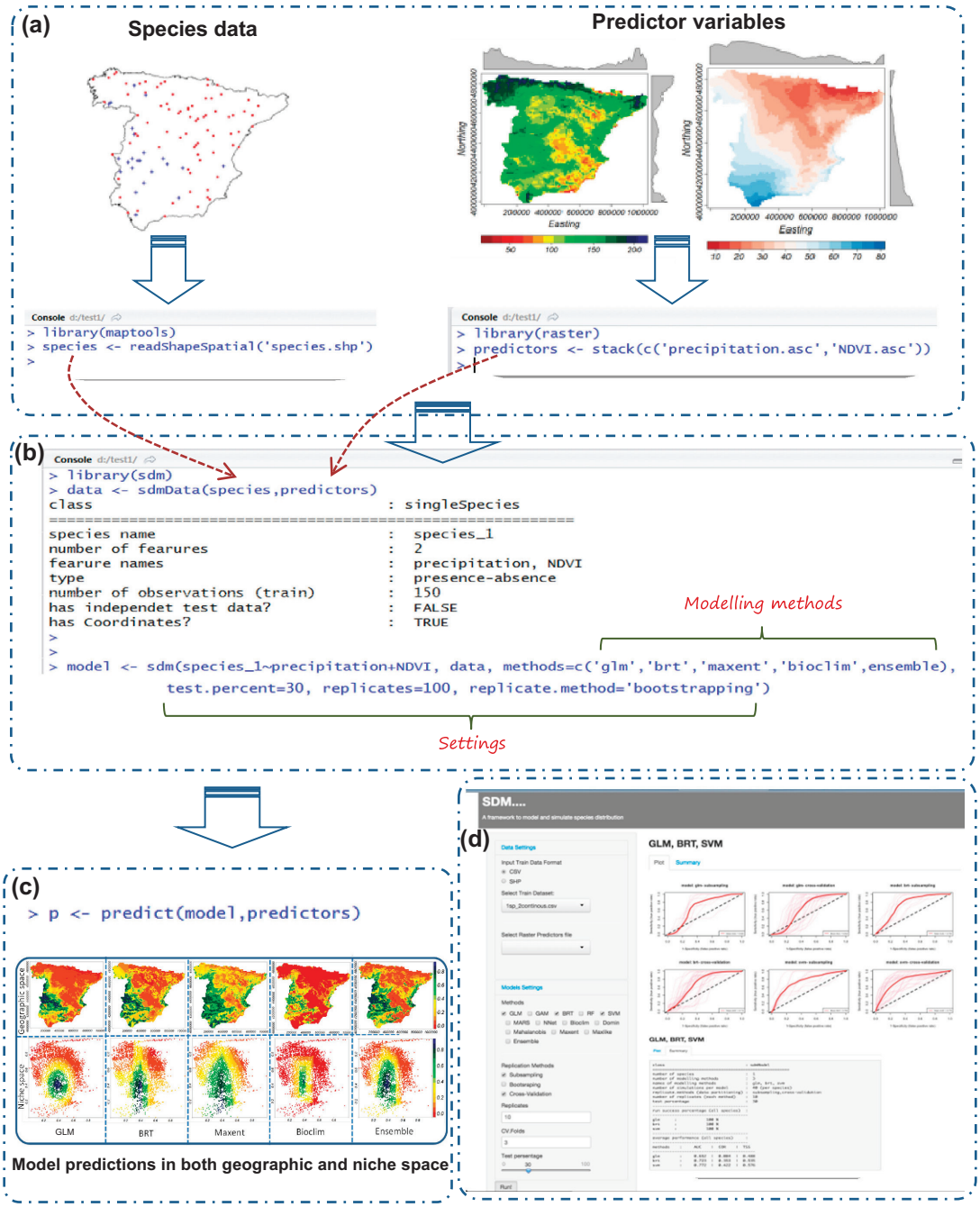

Figure 3. An example of using sdm package in R that demonstrating reading data, fitting species distribution models and predicting; (a) species dataset is provided as a spatial shapefile including presence–absence records, and two predictor variables (NDVI and precipitation) are in the format of raster Ascii files; (b) shows how the data are loaded and the model is fitted in the command line interface; (c) shows how the predict function can be used, and its outputs are visualized in geographic and niche space; (d) shows an example of GUI interface (as an alternative to the command line interface).

model-independent techniques were also implemented to evaluate the relative variable importance. For example, a method is to calculate the improvement of the model performance over inclusion of each variable comparing to when the variable is excluded through a cross-validation procedure. Another method is a randomization procedure that measures the correlation between the predicted values and predictions where the variable under investigation is randomly permutated. If the contribution of a variable to the model is high, then it is expected that the prediction is more affected by a permutation and therefore the correlation is lower. Therefore, '1 – correlation' can be considered as a measure of variable importance (Thuiller et al. 2009).

Other capabilities

Apart from the functionalities corresponding to the main steps in species distribution modelling, sdm implements additional functions and classes for different purposes

including controlling the data and procedures for handling the errors, facilitating the extensibility and reproducibility of the methods and procedures (by allowing to include or modify a method or procedure and distribute to the wider community), providing graphical user interface (GUI) and making the framework easy-to-use, generating dynamic reports, and implicitly parallelize the procedures to boost them through high performance computing, etc. Figure 3 provides a simple example on interfacing sdm through command line and GUI as well as some outputs. A tutorial contains further examples is provided with the package as a vignette, demostrating the main capabilities of sdm (as listed in 'design of the sdm package').

Conclusion

sdm is an object-oriented reproducible and extensible framework for species distribution modelling in R that unified different implementations of SDMs in a single framework. sdm provides an easy-to-use comprehensive framework to perform the entire modelling process within the same environment using different state-of-the-art approaches. The software is designed such to enable users to extend it and share the new data, methods or procedures to reproduce them by other users.

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

Naimi, B. and Araújo, M. B. 2016. sdm: a reproducible and extensible R platform for species distribution modelling. – Ecography 39: 368–375 (ver. 0).

References

Alfons, A. et al. 2010. An object-oriented framework for statistical simulation: the R package simFrame. – J. Stat. Softw. 37: 1–36.

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Araújo, M. B. and Peterson, A. T. 2012. Uses and misuses of bioclimatic envelope modeling. – Ecology 93: 1527–1539.

Araújo, M. B. et al. 2005a. Validation of species–climate impact models under climate change. – Global Change Biol. 11: 1504–1513.

Araújo, M. B. et al. 2005b. Reducing uncertainty in projections of extinction risk from climate change. – Global Ecol. Biogeogr. 14: 529–538.

Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. – Ecol. Model. 200: 1–19.

Barbet-Massin, M. et al. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.

Baselga, A. and Araújo, M. B. 2009. Individualistic vs community modelling of species distributions under climate change. – Ecography 32: 55–65.

Bivand, R. S. et al. 2008. Applied spatial data analysis with R. – Springer.

Breiman, L. 2001. Random forests. – Mach. Learn. 45: 5–32.

Breiman, L. et al. 1984. Classification and regression trees. – Wadsworth International Group, Belmont, CA, USA.

Busby, J. R. 1991. BIOCLIM – a bioclimate analysis and prediction system. – Plant Protection Quarterly, Australia.

Carpenter, G. et al. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. – Biodivers. Conserv. 2: 667–680.

Chambers, J. M. 2014. Object-oriented programming, functional programming and R. – Stat. Sci. 29: 167–180.

Chatterjee, S. and Hadi, A. S. 2006. Regression analysis by example. – John Wiley and Sons.

de Souza Muñoz, M. et al. 2009. openModeller: a generic approach to species' potential distribution modelling. – GeoInformatica 15: 111–135.

Diniz-Filho, J. A. F. et al. 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. – Ecography 32: 897–906.

Dormann, C. F. 2011. Modelling species' distributions. – In: Jopp, F. et al. (eds), Modelling complex ecological dynamics. Springer, pp. 179–196.

Dormann, C. F. et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. – Ecography 36: 27–46.

Elith, J. et al. 2005. The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. – Ecol. Model. 186: 280–289.

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

Farber, O. and Kadmon, R. 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. – Ecol. Model. 160: 115–130.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.

Friedman, J. H. 1991. Multivariate adaptive regression splines. – Ann. Stat. 19: 1–67.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. – Ann. Stat. 29: 1189–1232.

Garcia, R. A. et al. 2012. Exploring consensus in 21st century projections of climatically suitable areas for African vertebrates. – Global Change Biol. 18: 1253–1269.

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

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.

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

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

Hastie, T. et al. 1994. Flexible discriminant analysis by optimal scoring. – J. Am. Stat. Assoc. 89: 1255–1270.

Hastie, T. et al. 2009. The elements of statistical learning. – Springer.

Heikkinen, R. K. et al. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. – Prog. Phys. Geogr. 30: 751–777.

Hijmans, R. J. and Elith, J. 2013. dismo: species distribution modeling with R. – R project, species distribution modeling with R.

Hirzel, A. H. et al. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? – Ecology 83: 2027–2036.

Joppa, L. N. et al. 2013. Troubling trends in scientific software use. – Science 340: 814–815.

Kuhn, M. 2008. Building predictive models in R using the caret package. – J. Stat. Softw. 28: 1–26.

Madon, B. et al. 2013. Community-level vs species-specific approaches to model selection. – Ecography 36: 1291–1298.

Marquardt, D. W. 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. – Technometrics 12: 591–612.
McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall.
Murray, K. and Conner, M. M. 2009. Methods to quantify variable importance: implications for the analysis of noisy ecological data. – Ecology 90: 348–355.
Naimi, B. and Voinov, A. 2012. StellaR: a software to translate Stella models into R open-source environment. – Environ. Model. Softw. 38: 117–118.
Naimi, B. et al. 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. – J. Biogeogr. 38: 1497–1509.
Naimi, B. et al. 2014. Where is positional uncertainty a problem for species distribution modelling? – Ecography 37: 191–203.
Pearson, R. G. et al. 2006. Model-based uncertainty in species range prediction. – J. Biogeogr. 33: 1704–1711.
Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.
Peterson, A. T. et al. 2011. Ecological niches and geographic distributions: e-book. – Princeton Univ. Press.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Randin, C. F. et al. 2006. Are niche-based species distribution models transferable in space? – J. Biogeogr. 33: 1689–1703.
Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. – Psychol. Rev. 65: 386.
Rowe, R. J. 2005. Elevational gradient analyses and the use of historical museum specimens: a cautionary tale. – J. Biogeogr. 32: 1883–1897.
Royle, J. A. et al. 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. – Methods Ecol. Evol. 3: 545–554.
Segurado, P. and Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1568.
Soetaert, K. et al. 2010. Solving differential equations in R. – In: Psihoyios, G. and Tsitouras, C. (eds), Numerical analysis and applied mathematics, vol. I–III. Am. Inst. Physics, pp. 31–34.
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.
Thuiller, W. et al. 2004. Do we need land-cover data to model species distributions in Europe? – J. Biogeogr. 31: 353–361.
Thuiller, W. et al. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.
VanDerWal, J. et al. 2011. SDMTools: species distribution modelling tools: tools for processing data associated with species distribution modelling exercises. – R package.
Vapnik, V. 1995. The nature of statistical learning theory. – Springer.
Wood, C. et al. 2011. eBird: engaging birders in science and conservation. – PLoS Biol. 9: e1001220.