Distribution of R scripts for building and applying a RIVPACS-type predictive model

John Van Sickle

US Environmental Protection Agency, US Environmental Protection Agency, National Health and Environmental Effects Laboratory, Western Ecology Division, 200 SW 35th St., Corvallis, OR 97333 USA.
VanSickle.John@epa.gov

Version 4.2, Dec. 30, 2010 -- Upgrade Random Forest modules only.

The accompanying text files contain R-language scripts and functions for building and applying RIVPACS-type predictive models. The code can be implemented by experienced R programmers. All files within this distribution can be opened in any text editor or word processor.

Documentation consists either of comment lines within the scripts, or of external Help files, as noted below. I have run the scripts using R 2.10.0 and Windows XP, but they should work on earlier R versions as well.

The scripts are provided free of charge, and must be employed at the user's own risk. The scripts have not yet been subjected to EPA's peer and administrative review. Please contact the author when you find bugs.

CONTENTS –

model.build.v4.1.r -- Master program for model building. Includes code for calculating assemblage dissimilarities, for cluster analysis, and for calling functions that build discriminant function models (DFMs). Execute this code line by line, modifying as needed., and read comment lines carefully. Note alternative blocks of code for different dissimilarity measures, dendrogram pruning methods, etc. Internal documentation.
** NEW in Version 4.0 :
– Store your final model as an R object, or export it to text files for submission to the Western Center for Environmental Monitoring as a custom model.
--Tools to assess model performance.
-- Examples of making predictions for new data.
-- Compare the list of expected and observed taxa at a single site.

*NEW in V.4.1. Option to remove rare taxa before clustering. Improved documentation.

model.predict.v4.1.r – Predicts O, E, O/E, and BC, and their statistics, for any set of samples and/or sites. Also outputs predicted capture probabilities for all modeled taxa in each sample. Internal documentation.
* NEW in V. 4.1.
-- Inputted site-by-taxa data can contain any set of taxa (columns) in any order.
-- Predicted group membership probabilities for all test sites added to function output.

assess.one.sample.4.1.r – ** NEW in Version 4.0 - Compares predicted occurrence probabilities and observed occurrences for all taxa for a chosen test sample. Documented internally, and also in **model.build.v4.r.**

dfa.allsub.v4.r, and **dfa.allsub.v4.help.txt** – All-subsets selection of linear discriminant function (DF) models, given a set of candidate predictors and a candidate grouping of model calibration sites. Output includes statistics of O/E and BC for selected “best” models.
** NEW in Version 4.0 – Outputs cross-validated confusion matrices of all “best” DF models.

dfa.step.r and **dfa.step.help.txt --** Stepwise model selection for linear discriminant function models, given a set of candidate predictors and a candidate grouping of model calibration sites.

Random Forests. **Upgraded in Version 4.2.
Scripts **model.build.RanFor.4.2.r** and **model.predict.RanFor.4.2.r** replace the DF model with a Random Forest model, for predicting membership in site classes. Internal documentation. Version 4.2 upgrade allows out-of-bag or in-bag predictions for calibration data.

rarify.r , **rarify.help.txt**, and **rarify.examples.r.txt –** Randomly subsamples from a sample assemblage, without replacement, to create a fixed-count subsample.

matrify.r – Create a site-by-species matrix from an assemblage data set in 'list' form (i.e., each data line is the abundance of one species at one site).

align_test_ref.r – Reconstruct a site-by-species matrix (e.g., for validation or test sites) to have the same columns (species) as another site-by-species matrix (e.g., for calibration sites).

dapply.r, and **dapply.help.txt –** Generalized dyadic product, used to calculate a dissimilarity matrix when given a user-specified dissimilarity measure. Used in **model.build.v4.1.r**.

Enviros27Jul04.txt , **Bugs_matr_27Jul.txt –**
Example data sets of environmental predictor variables and macroinvertebrate taxa abundances, respectively. Apply the scripts to these data sets to see how the process works. These files also illustrate a convenient format for inputting data to **model.build.v4.1.r** or **model.build.RanFor.4.2.r .** This data is supplied courtesy of the Oregon Department of Environmental Quality.