*Supporting Information for:*

# "m2b" package in R: deriving multiple variables from movement data to predict behavioural states with random forests.

Andréa Thiebault[1*], Laurent Dubroca[2], Ralf Mullers[3], Yann Tremblay[4], Pierre Pistorius[5]

*1. Department of Zoology, Nelson Mandela University, South Campus, PO Box 77000, Port Elizabeth 6031, South Africa*
*2. Datacall Response Unit (CREDO), IFREMER, Avenue du Général de Gaulle, 14520, Port-en-Bessin-Huppain, France*
*3. Department of Biodiversity, University of Limpopo, Private Bag X1106, Sovenga 0787, South Africa*
*4. Institut de Recherche pour le Développement, UMR MARBEC 248: Marine Biodiversity, Exploitation and Conservation, Avenue Jean Monnet CS 30171, 34203 Sète cedex, France*
*5. DST/NRF Centre of Excellence at the Percy FitzPatrick Institute, Department of Zoology, Nelson Mandela University, South Campus, PO Box 77000, Port Elizabeth 6031, South Africa*
* Corresponding author: andrea.thiebault@gmail.com

## Supporting information 2. The random forest algorithm.

## The random forest algorithm with categorical response variable

A random forest is based on the classification and regression tree (CART) algorithm (figure 1), where a tree is constructed by recursively splitting the dataset into two subsets (Breiman (2003)). Starting from the root node that contains all the data, the splitting process stops when the terminal nodes are homogeneous, i.e. they contain only one class of data. To split the data at each node, all the predictor variables are tested individually and the couple [predictor , splitting value] that provides the two most homogeneous subsets with respect to the response variable is selected (Prasad, Iverson, and Liaw (2006)). Once the tree is fully grown, it can be pruned as in the CART algorithm, although this step is skipped to construct a random forest (Breiman (2003)).

Classification trees are unstable predictors, being very sensitive to small changes in the dataset (Breiman (1996)). To improve the accuracy of prediction, the random forest algorithm adds two levels of randomization in the construction of the trees (Breiman (2001)). Firstly at each node, a subset of variables is randomly selected to split the data. Secondly, several trees are grown from bootstrap samples of the dataset, where a bootstrap sample is constructed by randomly selecting data with replacement from the training (original) dataset. The dataset created in this way is of the same size as the training dataset, but it contains replicates while other data are missing (Breiman (1996)). All the trees are then aggregated for prediction, a method called “bagging” for bootstrap aggregating (Breiman (1996)). The strength of the random forest algorithm lies in the aggregation (bagging) of non-correlated (double randomization) unstable predictors (classification trees), to improve the accuracy of prediction (Breiman (2001)). The use of numerous trees in the forest ensures a stabilization of the results, as shown by the convergence of the error of prediction

(Breiman (2001)).

Once the model is built based on the training dataset, the class for new unlabelled data can be predicted. To do so, the model put the data down all the trees of the forest to get as many predicted classes as there are trees. These predictions are then aggregated, and the final result (one predicted class for this new unlabelled data) is chosen from a majority vote (Breiman (2003)).

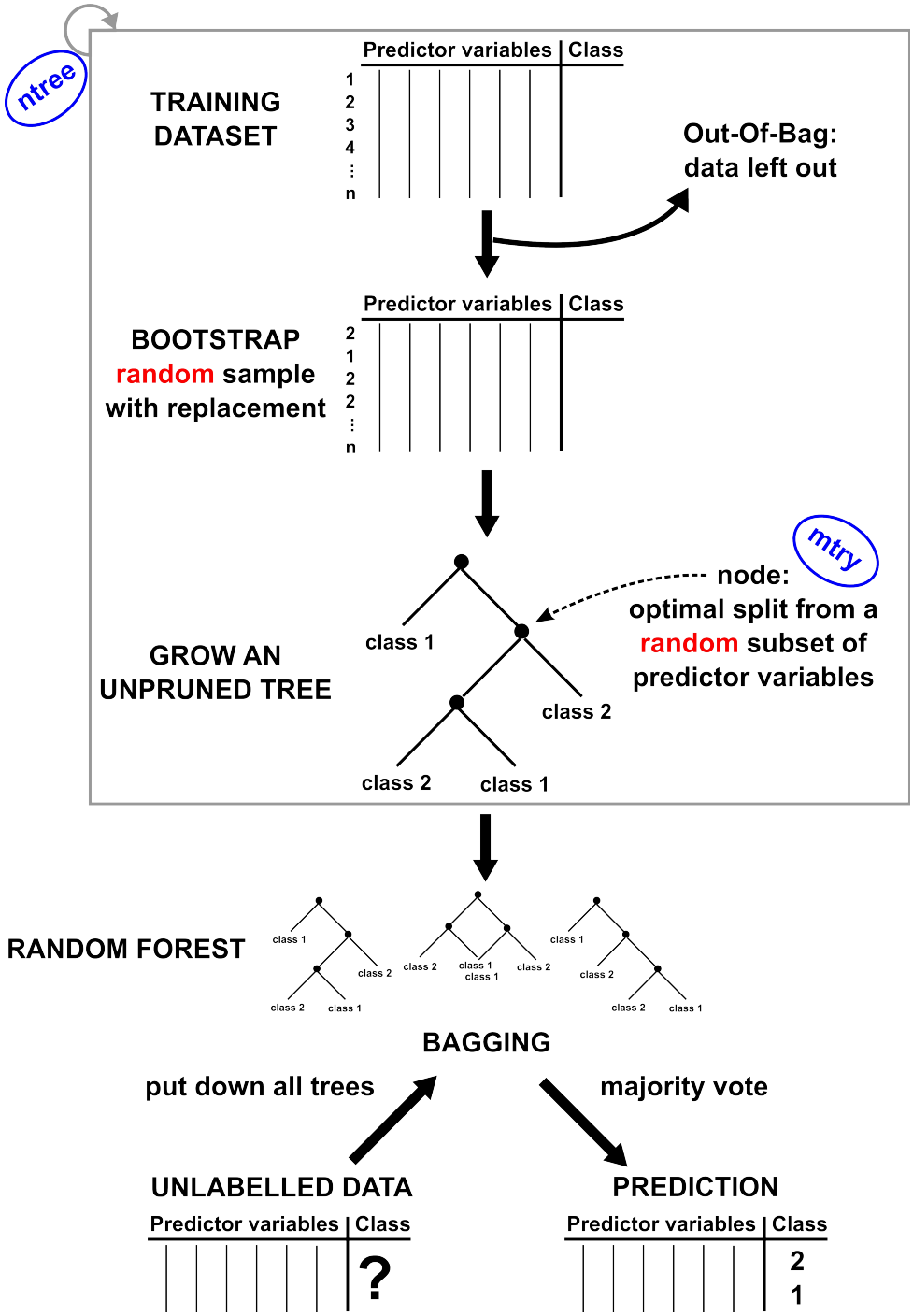


Figure 1: Random forest algorithm for classification. Based on the classification and regression tree algorithm, random forest adds two levels of randomness: (i) at each node a subset of the predictor variables is randomly selected to split the data, (ii) several trees are grown on bootstrap samples of the original dataset. The parameter «mtry» sets the number of variables to be selected at each node from the pool of predictors, and the parameter «ntree» sets the number of trees to be grown. The out-of-bag data created at each bootstrap are used to estimate the error of the model.

## *Parameters*

The random forest algorithm requires two parameters to be set.

The parameter «ntree» (number of trees to be grown) must be large enough for the results to converge (Breiman (2003)). The convergence of the algorithm can be checked by looking at the decrease of the OOB error rate with the increase of the number of trees grown.

The parameter «mtry» (number of predictor variables to be randomly selected at each node) must be large enough to minimize the error rate for each tree, but small enough to minimize the correlation between trees. To set this parameter, a nested cross-validation procedure can be used, as proposed by Svetnik et al. (2004) and implemented in the function «rfcv» in the «randomForest» package (Liaw et al. (2014)).

## *Imbalanced data*

Data in ecology are often imbalanced, meaning some classes (or behaviours) are observed more often than others. To deal with these data and increase the prediction accuracy for the rare classes, the ”balanced random forest” (Chen, Liaw, and Breiman (2004)) can be used. This procedure down samples the abundant classes: data are independently sampled from each class using a bootstrap process (random selection with replacement) to reach the size of the rare class. All the classes are then represented with the same number of cases to grow each tree.

## *Intrinsic estimation of error*

The random forest algorithm provides an internal measure of error, from the bootstrapping process (Breiman (2003)). At each bootstrap step, the data that were not included to grow the tree, or out-of bag (OOB) data, are put down this given tree to get a predicted class. At the end of the entire process, each data point gets a number of predictions (for as many times as they were left out from the bootstrap), and the final predicted class is chosen based on a voting process. The predicted classes for all the data points are then compared to the real classes to get an error rate, either global or for each class. This measure is called the OOB estimate of error rate (Breiman (2003)).

## *Variable importance*

In addition to a measure of error, the OOB data provide an estimate of the importance of predictor variables (Breiman (2003)). To do so, the same process as for the measure of the error rate is followed, except that all the values for the variable to be tested are randomly permuted before the OOB data is put down the tree. The difference between the accuracy measured using the original OOB data and using the data with randomized predictor values is then calculated (the larger the difference is, the more important the variable was for prediction). At the end of the entire process, the differences in accuracy calculated at each tree for a given variable are averaged and normalized by dividing them with the standard deviation so the measure can be compared among the various variables (Liaw and Wiener (2002)). If all predictor variables are continuous, this measure is assured to be unbiased (Strobl et al. (2007)).

# References


Breiman, L. (1996) Bagging Predictors. *Machine Learning* **24** (2) (August): 123–140.

Breiman, L. (2001) Random Forests. *Machine Learning* **45** (1) (October): 5– 32.

Breiman, L. (2003) Manual – Setting up, Using, and Understanding Random Forests V4.0. Report available at: http://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0. pdf.

Chao, C., Liaw, A. and Breiman, L. (2004) Using Random Forest to Learn Imbalanced Data. University of California, Berkeley.

Liaw, A. and Wiener, M. (2002) Classification and Regression by RandomForest. *R News* **2** (3): 18-22.

Liaw, A., Wiener, M., Breiman, L. and Cutler, A. (2014) random Forest: Breiman and Cutler's Random Forests for Classification and Regression. R-Package available at : http://cran.r-project.org/web/packages/randomForest/index.html

Prasad, A. M., Iverson, L. R. and Liaw, A. (2006) Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **9** (2): 181–199.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007) Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* **8** (1): 25.

Svetnik, V., Liaw, A., Tong, C. and Wang T. (2004) Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In *Multiple Classifier Systems*, edited by Fabio Roli, Josef Kittler, and Terry Windeatt, 334–343.