Advantages and Disadvantages of Random Forest in Modeling Species Distribution: From Theory to Practice

Yuting Vicky Lin

2021/10/19

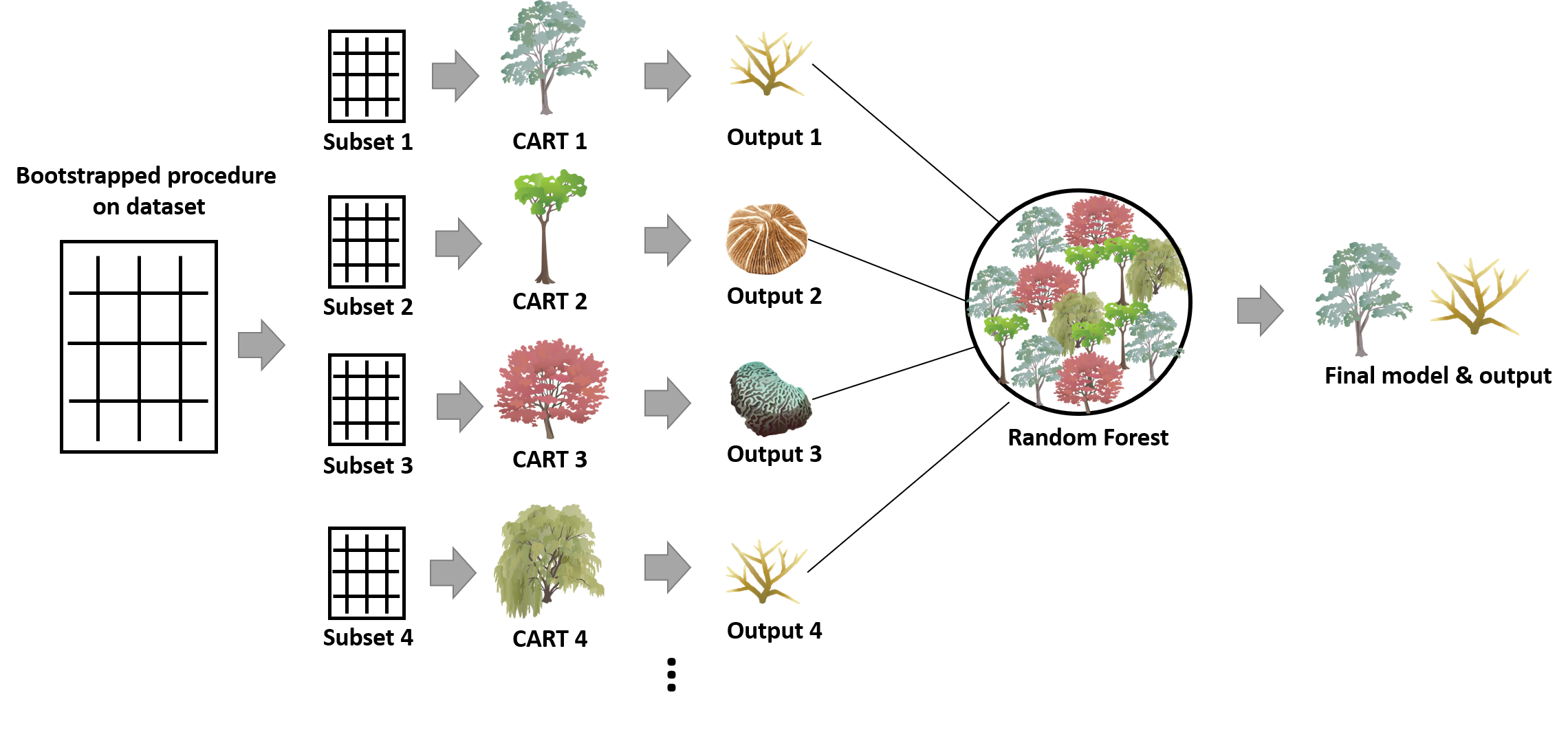
Table of Contents

# What is a random forest?

Random forest (RF) is a supervised learning algorithm that needs data to train. As its name, RF is an assemble of many individual decision trees which act as classifiers. Each tree can make a decision, and the final decision is made based on the major votes (classification tree) or the average (regression tree) from all decision trees. It is commonly used to deal with classification, regression, or prediction tasks and also to measure the relative importance of the input variables. In addition, it is easy to use and the results it produces are usually good and quite straightforward. Combing all these features, it becomes one of the most used algorithms in machine learning.

# How does it work?

RF uses a bootstrapped procedure on the dataset to repeatedly create many subsets. Each of the subsets is later used to fit either a classification or a regression tree, and the final decision is made based on the votes or the average from those trees (Figure 1). That is, an optimal model can be arrived at by converging many classification and regression trees (CARTs) fitted by the bootstrapping samples. Based on this best model, we can make a prediction of the new observations. Here, I will explain step by step how does this algorithm work and terminologies should be known to understand this algorithm.



**Figure 1.** The framework of the Random Forest algorithm

## - Step

Assume I have a dataset (N) and the independent variables (M) that I think are important features to delineate those observations. By applying RF to this dataset, I can **(1) find out the best model to fit this dataset, (2) predict the new observation according to this model, and (3) calculate the importance of those feature variables.** These three tasks are the target to obtain to answer the research questions when most people use RF. The procedure of this algorithm is:

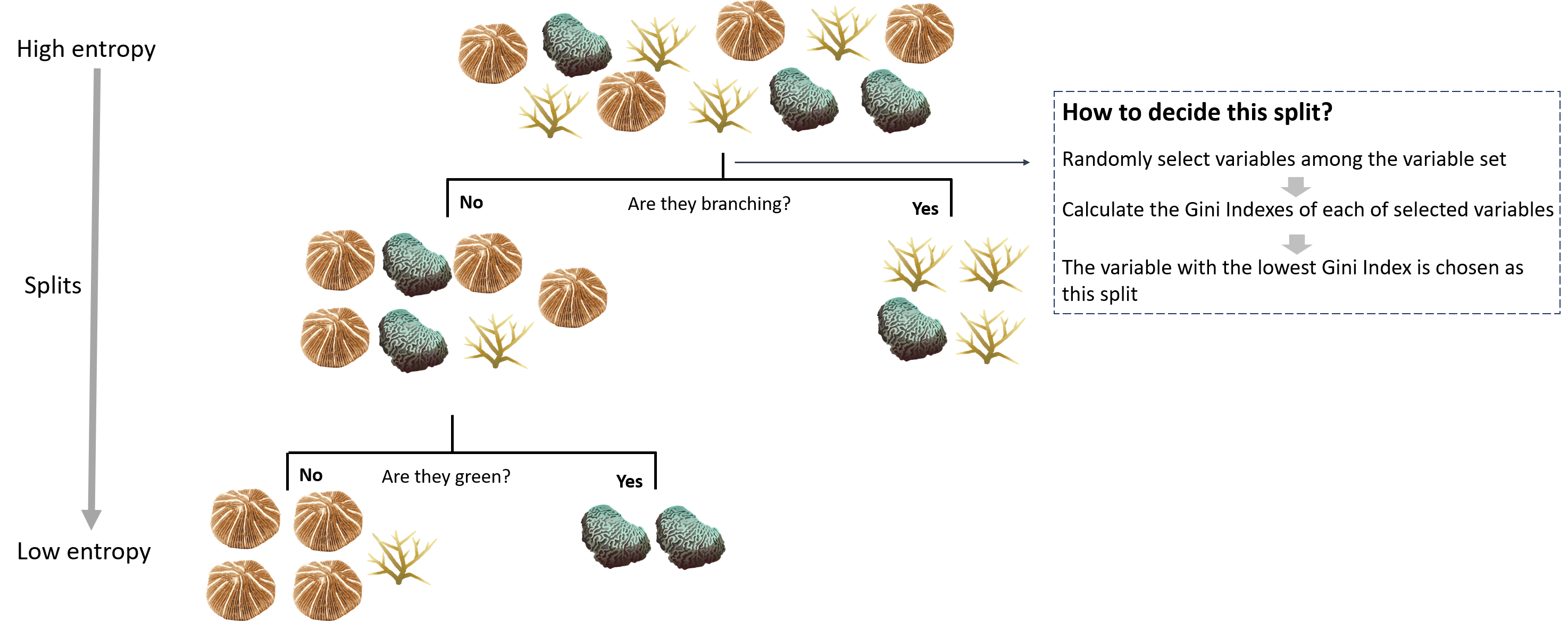
1. Build n Bootstrap replicates from the dataset N with the consistent amount of observations (with replacement). For each replicate, a CART will be created. That is, n is also the number of CARTs to grow.
2. Divide each Bootstrap replicate into a training and a testing set. The amount of data in two sets can be 9:1, 8:2, 7:3, or any ratio, there is no fixed rule but depending on the performance of your resulted model.
3. For each replicate, fit a CART to the training set. At each node of the tree, a permutation procedure randomly selecting m feature variables from the variable set (M) is conducted to find out the best split based on the Gini Index that can quantify the within-node entropy. After this CART is successfully grown, it would be validated with the testing set. The prediction error of the classification tree is measured by our-of-bag (bootstrapping) error while the one of regression tree is calculated by mean squared error.
4. Ensemble n CARTs to find out the best model.
5. Based on this model, a prediction can be made for the new observations.

## - Classification and regression tree (CART)

CART is a classifier used to sort the observations based on recursive binary decisions made by the given feature variables having the most information. However, how does the CART decide the binary decisions on each split? Before answering it, we have to know what is a Gini Index. Gini Index can measure the probability of a specific variable being incorrectly classified when randomly selected. This index is entropy-based, so their interpretation is similar. Gini Index ranges between 0 to 1, while 0 denotes that all the elements belong to a single class or there is only one class existing (pure), 1 denotes that all the elements randomly distribute among the classes (impure). Mathematically, it is expressed as:

where pi denotes the probability of an element being classified to a specific class.

In each split of a tree, Gini Indexes for each of the variables randomly picked up from its variable set are calculated, and the one with the lowest Gini Index is chosen as the given split (Figure 2). For the variable part, the more the variable contributes to the prediction process, the more chances it would be retained as the split. While for the observation, the more binary decisions the CART makes, the more similar the feature variables of the subset groups (Figure 2). Nonetheless, more splits on a tree do not always let the model better, too many splits may lead to the overfitting of the model.



**Figure 2.** A CART trying to classify corals based on their morphology and color. The entropy of the dataset decreases when each time a split is made.

## - Variable importance

An excellent piece of information that we can derive from RF is the relative importance of each feature variable on the prediction.

There are two kinds of measures. The first one measures the mean decrease in the accuracy. It measures the loss of accuracy when excluding the particular variable from the prediction model. Specifically, for each tree, the out-of-bag error (or mean squared error for the regression tree) is computed from its testing data. Then, the error is calculated again after permuting the values of each feature variable in testing data. The difference between the two error values is later averaged over all the trees. After normalization, the value obtained is the decrease in prediction accuracy.

The second measure is based on the decrease in Gini Index when a variable is selected as a split. In particular, it calculates for each variable the sum of the decrease in the Gini Index from each tree when it is chosen as the split. Then, this sum is divided by the total number of trees built to gain the decrease in the Gini Index of that variable.

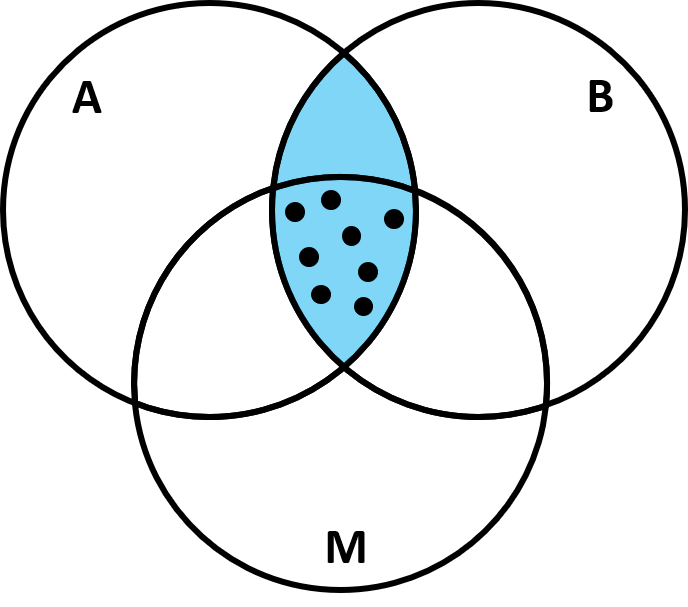
## - Hyperparameter tuning

Hyperparameter tuning refers to a set of parameters that needs to be decided before starting the algorithm. Those parameters can be adjusted either to make the model have better performance or to fasten the model. These are some common parameters that can be set before conducting the RF in R.

1. Number of trees (ntree)
2. Number of feature variables randomly sampled as candidates at each split. (mtry)

# Species distribution model (SDM)

The concept of species distribution is deeply connected to the niche theory that begins to be formulated in the early 90’ (Johnson 1910; Grinnell 1917; Elton 1927) and was later innovated by Hutchinson in 1957. He first differentiated the niche into the fundamental and the realized one. The former measures the geographical area where the species can have a positive population growth considering all the abiotic factors, while the latter encompasses not only the abiotic effect but also the competition among species. Therefore, it made the realized niche smaller than the fundamental one. However, nowadays, we know that some biotic effects such as facilitation (positive interspecific interactions) can help the population positively grow (Bruno et al. 2003). Until 2007, Soberon integrated the third critical factor, species dispersal or movement capacity, in the niche theory to consider if a species can reach a given area. He argued that before assessing if a particular species can thrive in the geographical area of interest where both abiotic and biotic conditions can satisfy the requirement of species, the species’ ability to arrive at that area must be examined. This is so-called biotic-abiotic-movement (BAM) concept that incorporates the interplay among these three factors into the assessment of species niche (Figure 3).



**Figure 3.** A BAM diagram suggests the three factors affecting the distribution of a species, abiotic factors (A), biotic factors (B), and dispersal or movement ability (M). It is adapted from Soberon (2007). The blue area between A and B indicates the area where the species can potentially have a positive growth considering its abiotic requirement and the prevailing biotic process, or also called “realized niche”. The intersection among A, B, and M (close circle area) represents the area where the species occupied can move in and increase its population

Now, although we know that the SDM originated from the niche theory, the dispute still exists in which niche we are trying to study in SDM, is it the realized niche (the blue area in Figure 3), or the occupied niche (the close circle area in Figure 3), or none of both? There might not be a correct answer for now. Depend on the interest of research, the selected factors can be largely different, resulting in the distinctive niches measured. Indeed, according to the knowledge of the ecologist on the targeted species, the availability of abiotic data, and the sampling scale, the selection of predicting variables in SDM can be very different. Common climate data including in the models are often temperature, precipitation…etc. If your targeted species is the tropical scleractinian corals, factors other than those common ones such as light intensity and sedimentation should also be considered because of the reliance of corals on zooxanthellate. However, those two parameters are much less important in regulating the distribution of pelagic fish species which do not largely rely on light.

# Advantages and disadvantages

After understanding both RF and SDM, we can start to talk about why RF becomes more and more popular in fitting SDM recently. It has several advantages when combing both:

1. **No need for assumptions before conducting the models** (e.g. independence of variables, normal distribution). This is a very great asset of this non-parametric model, which makes it outcompete to other linear-based models (e.g. generalized linear model). One of the required assumptions before conducting the traditional linear-based models is the independence of variables. However, since some environmental data are naturally spatial- or temporal-autocorrelation, additional statistical steps have to be implemented to deal with this issue before starting the linear-based models. Another required assumption when building linear-based models is the normality, nonetheless, those environmental data are not usually normally distributed. In this condition, RF is a non-parametric model that can still work when the normality of the data is not meet.
2. **Flexibility to address several types of analysis.** As mentioned earlier, RF can address many kinds of statistical analysis or models such as regression, classification, prediction. This flexibility makes RF very useful in modeling species distribution. For example, common types of environmental predictors include categorical and continuous, and the species data can be either presence-absence or abundance. The good news is, RF is able to integrate all of them.
3. **Identify non-linear relationships.** The response of species to the environmental condition or change is not always linear, it can be exponential, concave…etc. Sometimes, these non-linear interactions between the environmental factors and species can also be meaningful to affect species distribution. The advantage of RF is that it can discover non-linear relationships between both that the linear-based models cannot find. In addition, another limitation when using the linear-based model to fit species distribution is that those linear-based models can only test the priori hypothesis and are, therefore, hard to detect the relationship out of this priori expectation. On the other hand, RF is also able to figure out the complex interactions between feature variables. This is because the RF is an assemble of many CARTs, so it can address the environmental factors that interact with non-linear or with hierarchy.
4. **Allowing missing value.** RF automatically produces proximity, which is the measure of similarity among data points. This proximity matrix can be used to impute the missing data that could be obtained when ecologists have to simultaneously sample species and environmental data.
5. **Less overfitting.** CART tends to overfit the data because it would continuously split the nodes to minimize the within-node entropy. However, this issue is overcome in RF by choosing the best CART from many of them through the out-of-bag procedure/averaging many CARTs and also by randomly selecting the candidate variables in each node of trees. Overfitting is the last issue we would like to encounter when constructing a model since it could make the model perform badly in prediction, so does the SDM.

It seems it is perfect to model the distribution of species with RF. Nonetheless, several disadvantages still exist when conducting it:

1. **A large number of trees make the algorithm slow to run.** A large enough dataset, as well as number of trees, has to be prepared and built to gain a robust RF model. This process can make the algorithm time-consuming.
2. **Blackbox approach.** The ecologists have very few controls on what the algorithm does, basically, they can only try with different settings of hyperparameters (see the section of hyperparameter tuning) and different ratios of training and testing sets, then check the model afterward. This blackbox effect can sometimes make the results hard to interpret, however, this may not be a big issue if the ecologists already have a basic understanding of the species they are working on.

# Demonstration in R

To demonstrate how to apply RF on SDM in R, I chose three abiotic factors, sea surface temperature (SST), light at the bottom, and chlorophyll *a*, to model the distribution of a scleractinian coral species, *Stylophora pistillata*. The cover data of the coral species from 209 transects sampled in the north, east, and south Taiwan during 2015-2020 were derived from my Ph.D. research project. Each transect sampled the area of 5.25m2 and one unit value in species cover represent the area of 0.01 among 5.25 m2. As for the three abiotic parameters, I computed their 2014 mean annual values averaging the 12 months data downloaded from the ERDDP server managed by National Oceanic and Atmospheric Administration (NOAA), USA (dataset ID for temperature is NOAA\_DHW, for chlorophyll *a* is erdMH1chlamday, for light at the bottom are erdMH1par0mday and erdMH1kd490mday1). Since the abiotic variables I chose are all continuous ones, so I will demonstrate the RF with regression trees instead of classification ones.

1 light at the bottom = PAR x exp (-kd490 x z) where PAR is photosynthetically available radiation, kd49 means diffuse attenuation k490, and z represents the in situ depth.

Package installation

# install.pckages(c('readxl','caTools','randomForest','rgdal','ggplot2'))  
library(readxl)  
library(caTools)  
library(randomForest)  
library(rgdal)  
library(ggplot2)

Data importation

stpi <- read\_excel("Stylophora\_pistillata.xlsx") # import Stylophora pistillata cover and environmental data from 209 transects  
stpi <- stpi[ ,-c(1,5,6)] # remove the column with transect, min SST and max SST information  
cord <- read\_excel("cord.xlsx") # import the coordinate of sampling transects  
taiwan <- readOGR('TWN\_adm0.shp') # import the OGR file of Taiwan map

After importing all the files, let’s start to build the species distribution model with RF. The first thing I do is to divide the whole dataset into training and testing sets with the ratio of 7:3, this ratio can be changed if the new model performed better than this one.

sample.stpi <- caTools::sample.split(stpi$SP, SplitRatio = .7)  
stpi.train <- subset(stpi, sample.stpi == TRUE)  
stpi.test <- subset(stpi, sample.stpi == FALSE)  
dim(stpi.train)

## [1] 148 4

dim(stpi.test)

## [1] 61 4

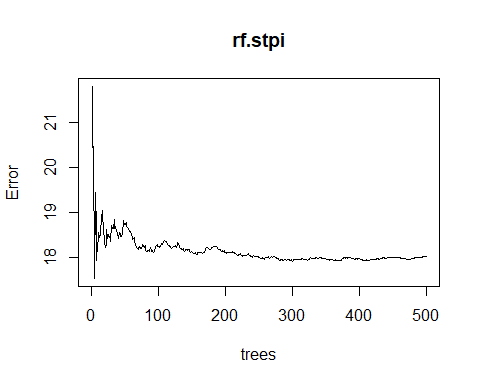
Then, I can start to create the model.

set.seed(11) # set a random seed to make the model reproducible  
rf.stpi <- randomForest(SP ~ .,data=stpi.train, importance = T) # build the random forest model  
rf.stpi

##   
## Call:  
## randomForest(formula = SP ~ ., data = stpi.train, importance = T)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 1  
##   
## Mean of squared residuals: 18.03562  
## % Var explained: 17.6

The result of this model shows that the three factors can together explain 24.88% of the total variance. This is not high, but with only three variables, it is also not bad. Still, more abiotic factors are recommended to add to this model to catch more variance. I use 500 trees (R’s default) to build this model, and caution is needed for the number of trees since too many trees may make the model overfitting. Therefore, checking the mean square errors of models created with the different number of trees is recommended.

plot(rf.stpi) # check the mean square error of models



From the plot, we know that as more and more trees are built, the errors decrease and eventually become stable. It seems that when the number of trees is around 150, we arrived at the lowest error. We can check it with the “which.min” function.

which.min(rf.stpi$mse)

## [1] 5

It turns out that we can have the lowest error in the RF model with 133 trees, so let’s rebuild the model with this number of trees by using the argument “ntree”.

Then, I can start to create the model.

set.seed(1)  
rf.stpi.min <- randomForest(SP ~ .,data=stpi.train, importance = T, ntree=133)  
rf.stpi.min

##   
## Call:  
## randomForest(formula = SP ~ ., data = stpi.train, importance = T, ntree = 133)   
## Type of random forest: regression  
## Number of trees: 133  
## No. of variables tried at each split: 1  
##   
## Mean of squared residuals: 17.71559  
## % Var explained: 19.06

The variance explained slightly increase and the error slightly decrease, so this model performs a little bit better than the one with 500 trees. In addition to this hyperparameter, we can also adjust the number of the variables randomly selected at each split with the argument “mtry” to make the model perform better.

rf.stpi.min.2 <- randomForest(SP ~ .,data=stpi.train, importance = T,ntree=133, mtry=2)  
rf.stpi.min.2

##   
## Call:  
## randomForest(formula = SP ~ ., data = stpi.train, importance = T, ntree = 133, mtry = 2)   
## Type of random forest: regression  
## Number of trees: 133  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 18.21749  
## % Var explained: 16.76

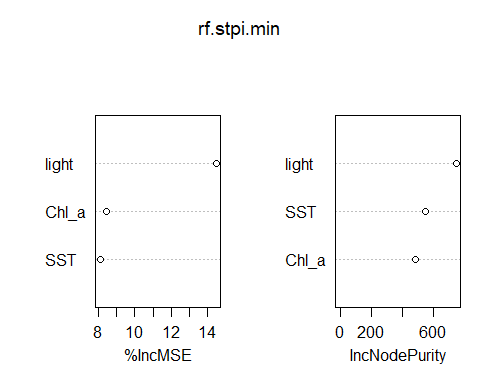
rf.stpi.min.3 <- randomForest(SP ~ .,data=stpi.train, importance = T,ntree=133, mtry=3)  
rf.stpi.min.3

##   
## Call:  
## randomForest(formula = SP ~ ., data = stpi.train, importance = T, ntree = 133, mtry = 3)   
## Type of random forest: regression  
## Number of trees: 133  
## No. of variables tried at each split: 3  
##   
## Mean of squared residuals: 17.95728  
## % Var explained: 17.95

However, the models with the different number of variables do not perform better than the original one. Hence, we will use the one with mtry =1.

Now, we can check the importance of those variables in this model with the function “varImpPlot” and” importance”.

varImpPlot(rf.stpi.min)



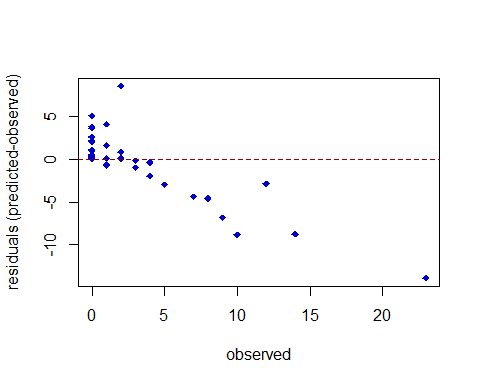
importance(rf.stpi.min)

## %IncMSE IncNodePurity  
## light 14.468713 744.0756  
## SST 8.114595 544.8524  
## Chl\_a 8.459833 479.1057

The result shows the importance of variables calculated with the mean decrease in accuracy and with the Gini Index. Both suggest that light is the most important factor in driving the distribution of *S. pistillata* followed by SST and the chlorophyll *a*.

Now, we can try to validate this model with the testing set by calculating the residuals and the mean squared error.

pred.stpi <- predict(rf.stpi.min, newdata=stpi.test)  
rf.resid.stpi <- pred.stpi - stpi.test$SP # the residuals  
plot(stpi.test$SP,rf.resid.stpi, pch=18,ylab="residuals (predicted-observed)", xlab="observed",col="blue3") # residual plot  
abline(h=0,col="red4",lty=2)



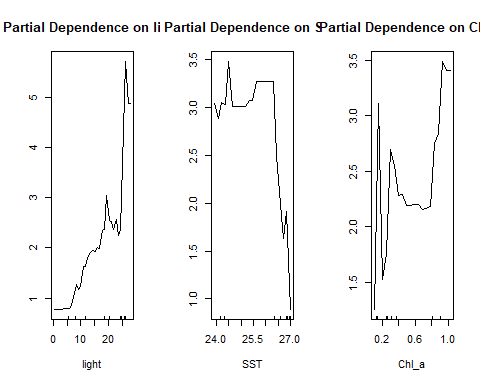
mean(rf.resid.stpi^2) # mean squared error

## [1] 10.83963

The residual plot shows that the model makes around 5% cover error when predicting the distribution of *S. pistillata*, and more error would be found in the area with higher cover. This may be because there are too little data in our dataset especially in the area with the higher cover of *S. pistillata*. The value of mean squared error can be used to compare to other models and figure out the best one.

Then, we can have a look on how the individual abiotic variable affecting the model with partial plots with the function ”partialPlot”.

par(mfrow=c(1,3))   
partialPlot(rf.stpi.min, as.data.frame(stpi.train), light)  
partialPlot(rf.stpi.min, as.data.frame(stpi.train), SST)  
partialPlot(rf.stpi.min, as.data.frame(stpi.train), Chl\_a)



dev.off()

The purpose of the partial plot is to display how the prediction of coral cover is changing with the change of each of the three abiotic variables. If the line is closer to zero, it means that the given variable is not affecting the coral cover at all, while the more the line deviating from zero, this variable influences the coral cover more. Instead of looking at the value on the y-axis, the relationship between the y-axis and the given variable across a specific range is what we should focus on. For example, in the partial dependence plot on light, when the light intensity becomes stronger, we have more chances to see a high cover of *S. pistillata*. This increase of chances is even more obvious when the light intensity is between 5 to 15 W/m2/day and over 25 W/m2/day. Look at the partial dependence plot on SST, we found that the cover of *S. pistillata* would dramatically decrease when the 2014 mean SST is above 26 OC and its cover remains quite consistent when it is below 26 OC. Lastly, for the partial dependence plot on chlorophyll a, it peaks when the concentration of chlorophyll *a* is at around 0.2, 0.4, and 1 mg/m3. By checking the tendency and the value in the y-axis on these three partial plots, it may explain why light intensity is the most important factor in predicting the distribution of *S. pistillatav*.

Now, let’s also draw the distribution map of *S. pistillata* with cover information to visualize it.

cord <- cbind(cord,stpi$SP)  
colnames(cord) <- c('Tr','Lat',"Lon",'Cover')  
lat <- as.factor(cord$Lat)  
lon <- as.factor(cord$Lon)  
coords <- SpatialPoints(cord[, c("Lon", "Lat")])   
stpi\_cord <- SpatialPointsDataFrame(coords, cord)   
proj4string(stpi\_cord) <- CRS("+proj=longlat +no\_defs +ellps=WGS84", doCheckCRSArgs = F)   
head(stpi\_cord)

## coordinates Tr Lat Lon Cover  
## 1 (121.8046, 25.14455) N\_CJ\_5\_T1 25.14455 121.8046 12  
## 2 (121.8046, 25.14455) N\_CJ\_5\_T2 25.14455 121.8046 1  
## 3 (121.8046, 25.14455) N\_CJ\_5\_T3 25.14455 121.8046 4  
## 4 (121.8046, 25.14455) N\_CJ\_5\_T4 25.14455 121.8046 0  
## 5 (121.8046, 25.14455) N\_CJ\_5\_T5 25.14455 121.8046 3  
## 6 (121.9146, 25.1257) N\_BT\_5\_T1 25.12570 121.9146 0

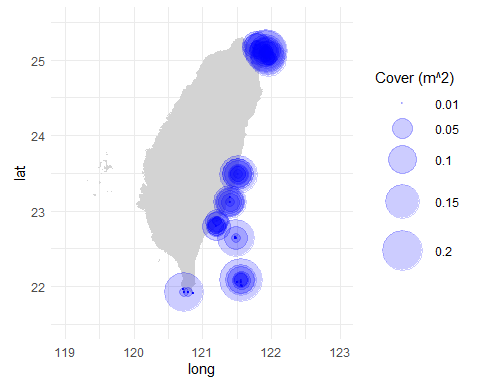
class(stpi\_cord)

## [1] "SpatialPointsDataFrame"  
## attr(,"package")  
## [1] "sp"

taiwan\_map <- fortify(taiwan)

## Regions defined for each Polygons

Taiwan.map <- ggplot() +  
 geom\_polygon(data= taiwan\_map, aes(x= long, y= lat, group= group), fill='light grey') +  
 scale\_x\_continuous(limits = c(119, 123)) + # set x and y limt  
 scale\_y\_continuous(limits = c(21.5, 25.5)) +  
 theme\_minimal()+ # background   
 coord\_fixed(1.1)+ # fix the coordinate  
 geom\_point(cord, mapping = aes(x= Lon, y= Lat, size = Cover),alpha=0.2, color="blue") + # add the location of sampling transects  
 theme(legend.position="right")  
  
Taiwan.map+ scale\_size\_continuous(name = 'Cover (m^2)' , range = c(0, 15),breaks = c(0,5,10,15,20), labels = c('0.01','0.05','0.1','0.15','0.2')) # adjust the size of bubble plot



According to this map, *S. pistillata* has higher cover in the north of Taiwan while the east and the south have less cover. Based on its distribution and the RF model, we can deduce that *S. pistillata* tends to live in the water with higher light intensity and lower SST, concentrating in north Taiwan. The deduction for the light makes sense because this coral species needs light to support the growth of their zooxanthellate, which can in turn provide the coral species energy. Despite tropical scleractinian corals usually favor warm and bright areas, *S. pistillata* is more abundant in the north where the SST is lower. This might be because this species is not competitive enough to other coral species when the thermal environment is suitable for growth, so they are not abundant in the east and the south where the thermal condition can allow many coral species to develop. However, there is another possibility that this dataset may suffer from oversampling in the shallower water of the north (26% of the transects sampled at < 20 meters in the north, 19% of the transects sampled at < 20 meters in the east, 17% of the transects sampled at < 20 meters in the Orchid Island, Green Island and Kenting are less than 17%). It may make the algorithm have more chances to sample the transects in the north when bootstrapping, hence, the results would bias toward the environment of the north shallow water where is colder and lighter. More sampling effort needs to be added to the areas other than the shallow water in the north of Taiwan to make sure if this can be an issue or not.

Overall, this RF model suggests that the three abiotic variables, SST, light at the bottom, and chlorophyll *a*, all positively contribute to the prediction of the distribution of *S*. pistillata\* with the light does it better than the other two variables. With these three factors, we can already catch 25.95% of the total variance. Nonetheless, considering the number of observations and also the number of variables, this dataset is too small to build a very robust RF model. Further sample effort is needed to make this prediction stronger especially in places other than the shallow water in the north of Taiwan.

# Summary

RF can integrate different types of variables as well as a large amount of data, and seek for the complex relationship between the predicted variables and the response variables that sometimes the linear-based model cannot find. These features allow RF to stand out as one of the most popular approaches in modeling species distribution, in which the complex and dependent relationship commonly exists within and among environmental factors. Furthermore, based on a robust RF model, a prediction on the future alteration of species distribution considering the human-driven changes in any environmental parameters can be further estimated. This information could be quite helpful in conservation.

# Reference

Bruno JF, Stachowicz JJ, Bertness MD (2003) Inclusion of facilitation into ecological theory. Trends Ecol Evol 18:119-125

Cutler DR, Edwards TC, Beard KH, Catler A, Hess KT, Gibson J, Lawler JJ (2007) Random forest for classification in ecology. Ecology 88:2783–2792

Elton CS (1927) Animal ecology. Sedgwick & Jackson, London, U.K Grinnell J (1917) The niche-relationships of the California trasher. The Auk 34:427-433

Hutchinson GE (1957) Concluding remarks, cold spring harbor symposium. Quantitative Biology 22:415-427

Jeffrey S. Evans, Melanie A. Murphy, Zachary A. Holden, Cushman SA (2011) Modeling species distribution and change using random forest. In: Drew CA, Wiersma YF, Huettmann F (eds) Predictive species and habitat modeling in landscape ecology. Springer New York Dordrecht Heidelberg London,

Johnson RH (1910) Determinate evolution in the color-pattern of the lady-beetles. Carnegie Inst. of Washington, Washington, D.C., USA

Soberon J (2007) Grinnellian and Eltonian niches and geographic distributions of species. Ecol Lett 10:1115-1123 [10.1111/j.1461-0248.2007.01107.x]

<https://builtin.com/data-science/random-forest-algorithm>

<https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

<https://damariszurell.github.io/SDM-Intro/#1_Background>