Calvin Tsang

SIO209

16 Jun 2016

Predicting Coral Colony Fate with Random Forest

Summary

Coral reef change rapidly over time. Within a single year, there are multiple transitions for a single colony between growth and shrinkage that eventually lead to its survival or death. To date, there is a substantial amount of literature attempting to reason the causes for changing coral fate and with different approaches. One approach to monitor this change over time is large-area imagery. Preservation of the coral reef in a high-resolution 3D model allows analyzing data in time series on a large scale (Dornelas, 2016). In this novel and exploratory study, the use of random forest, an ensemble machine learning method, is combined with large-area imagery in order to predict how a coral colony's health will be in the future. The aim is to predict, given a colony is alive the next year, will it: 1) Grow or 2) Shrink? The study finds that random forest is an adequate modelling technique to discover trends, but much of the predictions are biased towards Growth because it has higher frequencies. The major variables in predicting coral colony fate are typically measurements of the colony in question, such as its area, perimeter, and species as opposed to how many neighbors it encounters.

Introduction

Coral reefs are a constantly changing environment. At its lowest level, coral reefs are composed of thousands of coral colonies, colonial organisms that maintain numerous polyps within its skeletal structure. These ecosystem engineers are responsible for creating and significantly modifying the area around them. Not only are they a source of habitat for benthic fauna, they provide protection from wave motion and food as well as nitrogen fixing. As the individual coral colony develops and seeks to grow and survive, it affects the surrounding colonies and their ability to survive. This will determine whether it will live or die. Through the life of a coral colony, if a single polyp or living tissue dies this does not mean the entire coral colony dies. This only indicates the colony has shrunk. Fortunately, coral colonies can clone polyps and build its colony by creating more skeletal structure, which provides a basis for growth. These two outcomes have many influences, which are still ambiguous. One potential influence of coral fate lies with the biology of corals. Since they are sessile organisms, they are always in the presence of their immediate neighbors and effects in the surrounding environment. The effect of coral competition has a detrimental condition on the corals interacting. While hard coral competition may not be a cause for death, it does have a negative effect on growth rates and thus, the ability to reproduce and sustain health in the future (Lang, 1973; Tanner, 1996).

Numerous studies have attempted to find the combination of factors responsible for coral colony fate but lack the means to study phenomena across sites and in large-scale. Classic methods

include the basic 1 meter by 1 meter quadrat, supplemented by underwater cameras to grasp an idea of the health of coral reefs. The lacking qualities of this method is that is does not capture large-scale effects and requires an extraordinary amount of effort to generate a large amount of data. It is also improbable that it captures the information about large corals and their neighbors. An improvement upon this method is the large-scale image. With relatively cheap cameras such as the go-pro, divers can capture thousands of images of the benthic floor that can then be stitched together using identical features that overlap in the photos to create a 3D model (Delparte et al., 2015; Guo et al., 2016). The algorithm for creating these models uses cues from the pictures such as shadows and contrasting objects (think pink coral colony against green algae) to create a 3 dimensional representation of the coral reef. An image is taken from a particular angle of the model, preferably an aerial view that is able to preserve the entire model and high enough that features are relatively undistorted, which is termed an orthoprojection (Figure 1). Repeat for enough years and sites and we have time series across multiple study sites and reefs.

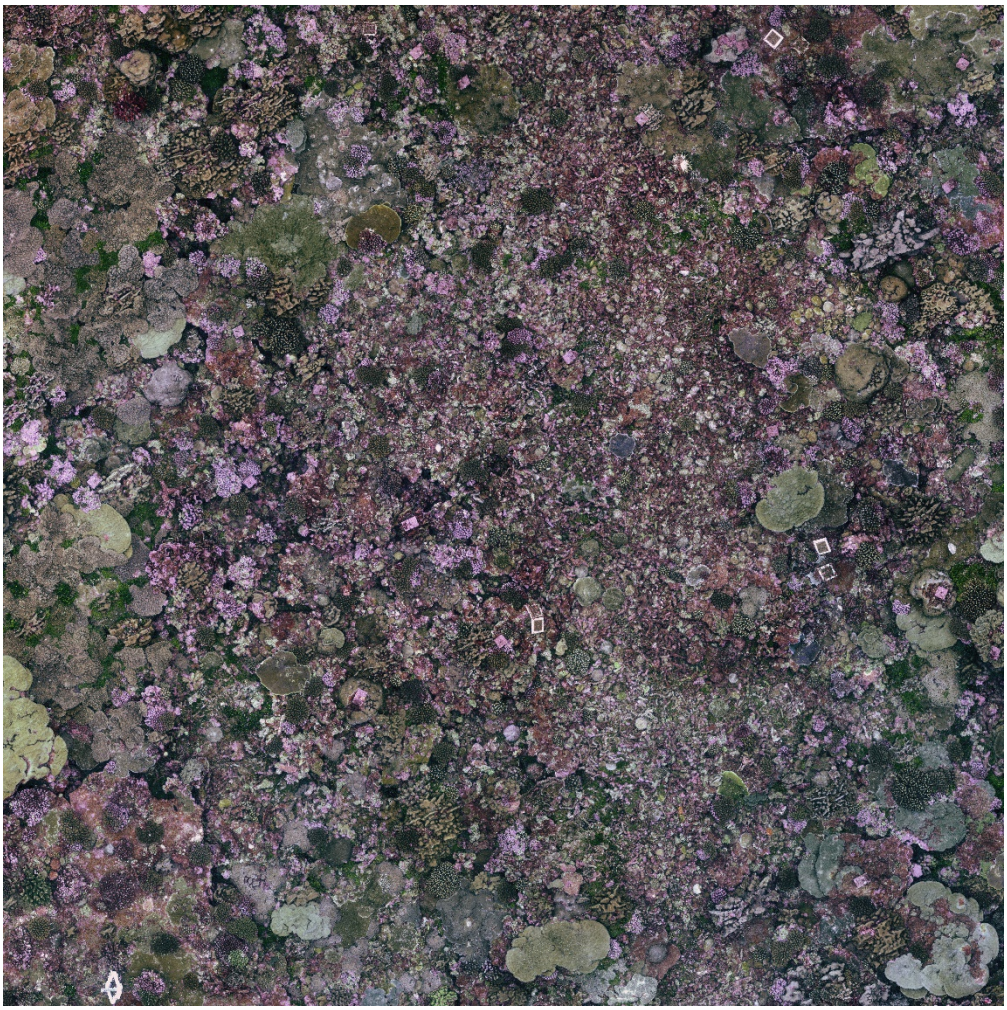

Figure 1: Orthoprojection (100m^2) – Palmyra FR3 2013

After processing the data from the orthoprojections, random forest will be tested to see how well it can predict coral growth and shrinkage. Random Forest is an ensemble machine learning method that creates binary decision trees using bootstraps of the training data. The “random” part of random forest is due to the selection of a random subset of the predictors among every branch in the tree; the larger the difference between the two outcomes in a predictor, the more likely it will be selected as the predictor for the split (Breiman, 2001). The “forest” part is due to repeating the creation of the bootstrapping trees until a desired limit. The decision trees create a “forest”. Furthermore, random forest includes variable importance determination by the creation of a tree model. For each bootstrapped tree, approximately 2/3 of the data is required for the model, but 1/3 is never selected. This is termed “out-of-bag”, which is then used as a test set for the model created. Once the prediction of this out-of-bag test data is obtained, a single predictor among the test set is randomly permuted and the test data is run through the model again. The prediction will then change accordingly depending on how important the variable was. This variable is then given a numerical value reflecting the amount of prediction loss. If the amount of prediction significantly dropped, it will have a bigger number indicating that is relatively important. This process is repeated with all other predictors to find how well they measure against each other. For classification, this is called “Mean Decrease Accuracy” while in regression it is called “Percent Increase Mean Square Error” (%IncMSE) instead.

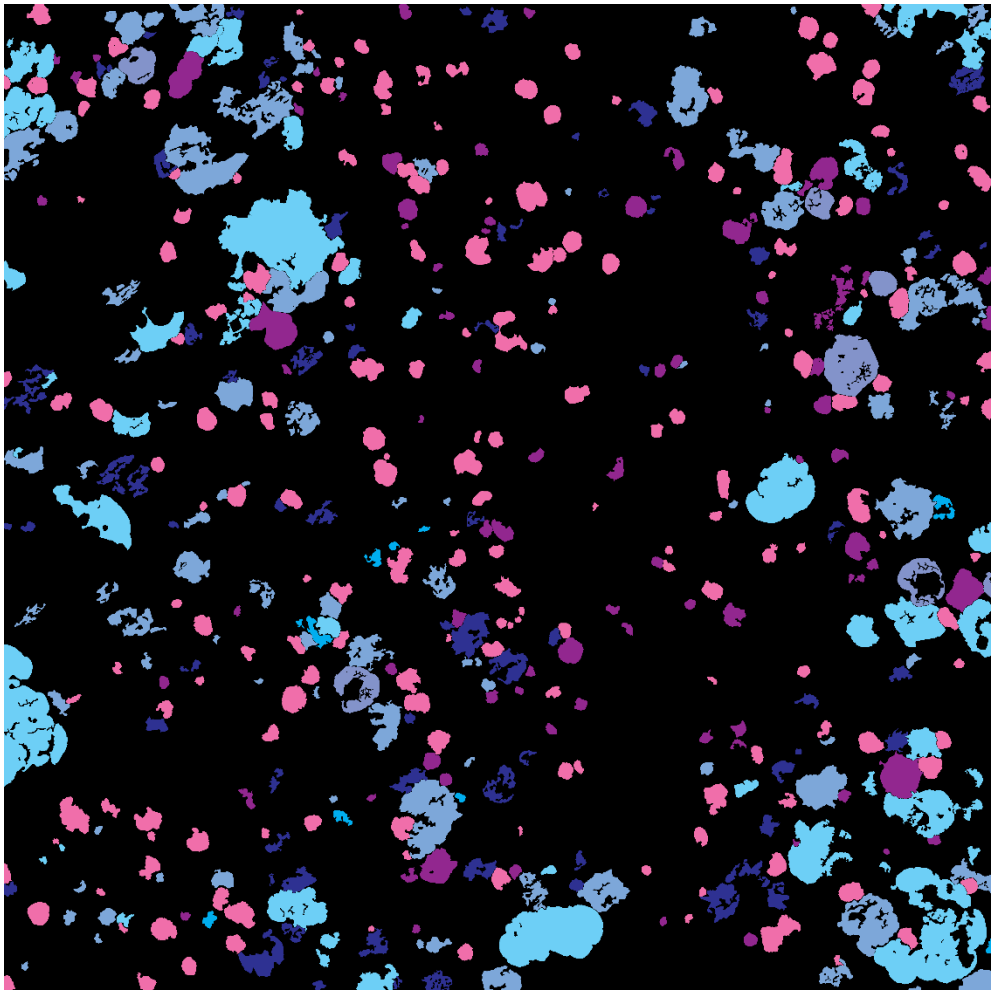

Figure 2: Digitized Orthoprojection (100m^2) – Palmyra FR3 2013

Physical and Mathematical Framework

From an orthoprojection, a subset of species are selected to analyze, which are then traced and color coded to represent that particular species. In this study, we choose 7 genus-morphology level species to study: *Favia, Hydnophora, Montastrea, Pocillopora, Pavona, Porites*, and *Stylophora* (Figure 2). These species represent a large portion of the potential hard corals in the study site, Palmyra atoll. Four sites (FR3, FR5, FR7, and FR9) are chosen to analyze in 2013 and 2014. The purpose of choosing a single time series is to maximize effect size of the variables and decrease the potential for pseudo replication seen in merging coral colonies. To current date, only competition-based predictors and measurements on the focal colony (colony in question) have been extracted. A buffer of 10 centimeters serves as the immediate surroundings that will be analyzed on the colonies because the maximum reach of coral aggression techniques are within 10 centimeters (Sheppard, 1981).

The predictors obtained from the orthoprojection are (numerics are in centimeters & the right section gives a description of the predictor and the intuition behind selecting it):

Table 1: Biologically relevant predictors and descriptions

1	Site	What site it was located in. This covers any potential effects like temperature or salinity that isn't capture by the images
2	Morph	The morphology of the colony. Tells of the life history of the colony and how its shape is.
3	Species	The species of the colony. Different species have different growth rates and life histories.
4	Neighbor Count	Number of neighbors in its buffer. More neighbors should have a negative effect on the colony.
5	Neighbor Diversity	Number of species in its buffer. More types of species could have an effect on colony.
6	Total Neighbor Area	Accumulated area of its neighbors. More total neighbor area could have a negative effect on colony.
7	Mean Neighbor Area	Average area of its neighbors. The bigger the average neighbor, the higher chance it will have a negative effect on colony.
8	Area	Area of the focal colony. Representation of coral health and its cover on reef.
9	Perimeter	Perimeter of the focal colony. The length it is exposed to surrounding neighbors
10	Perimeter-Area Ratio	P/A Ratio of the focal colony. Reflection of size and shape of the coral, which is health.
11	Circularity Factor	Deviation from a perfect circle - $(4\pi A / P^2)$. Meant to correct for the exponentially increasing area in perimeter-area ratio.
12	Minimum Distance	The shortest distance from any neighbor. The closer a neighbor is, the more potential it has for aggression.
13	Percent Occupation	The amount of area within its buffer. The higher, the more aggression it is exposed to in its immediate environment.
14	Largest Neighbor	The biggest neighbor area. Bigger neighbors should have stronger advantage or cause stress.
15	Encounter *Favia*	Did it encounter *Favia*?
16	Encounter *Hydnophora*	Did it encounter *Hydnophora*?
17	Encounter *Montastrea*	Did it encounter *Montastrea*?
18	Encounter *Pavona*	Did it encounter *Pavona*?
19	Encounter *Porites*	Did it encounter *Porites*?
20	Encounter *Pocillopora*	Did it encounter *Pocillopora*?
21	Encounter *Stylophora*	Did it encounter *Stylophora*?

All analyses were run in R (RStudio). Special packages included in analyses were: rgeos and randomForest. For the classification of Shrink and Growth, the number of predictors that were subset at each break was 5 and the node side was 1. One thousand trees were run per prediction and random forest was repeated 1000 times for the confidence intervals.

A null distribution model through a simple permutation was created by taking the data set and partitioning it into a training and test set. The training set would be counted for its frequencies and the test set would be permuted to match the proportion of the training set. This would simulate a random binary assignment for the response variable to create the simplest model possible. The test set was then compared to the permutation and a confusion matrix was drawn to find the total accuracy and recall of the outcomes. This was then repeated 1000 times with the same split to find a null distribution. Random forest was then run once to find if a single repetition would be considered significantly better than random at a 95% confidence interval. This was then run 100 times where a rate of 80% or better would indicate the random forest prediction did better than random. If the model was significant, it would then be selected for further analysis.

Results

The frequencies of Shrink to Growth are given in Table 2. Out of 1474 focal colonies to examine, 568 experienced shrinkage totaling to 39% of the total data. For *Hydnophora* and *Montastrea*, there is a low amount of samples noted. With the exception of these two, all the rest have sufficient sample sizes above 150 and *Pocillopora* especially has 612. *Pocillopora* and *Pavona* have at least 45% of their proportion to be shrinkage.

Table 2: Shrink Counts in Samples

	Favia	***Hydnophora***	***Montastrea***	***Pavona***	***Pocillopora***	***Porites***	***Stylophora***	**Total**
% Shrink	38 %	22 %	43 %	45 %	48 %	30 %	19 %	39 %
Counts	$\frac{70}{186}$	$\frac{11}{50}$	$\frac{22}{51}$	$\frac{82}{181}$	$\frac{291}{612}$	$\frac{45}{151}$	$\frac{47}{243}$	$\frac{568}{1474}$

Using the null distribution significance test through permutations, we find for random forest, *Hydnophora, Montastrea, and Pavona* do not meet the minimum criteria of being significant 80% of the time (Table 3). The full model including all species data and *Stylophora* have significant prediction but only for Growth and the Overall accuracy. *Favia* and *Porites* only display significance for Growth. The insignificant Shrink prediction is coupled with the observation made earlier that the frequencies of the outcomes are biased for these four models.

Pocillopora is the only model that significantly predicts all outcomes, which is noted to have nearly equal frequencies, and therefore no bias in the proportion.

Table 3: Model Significance

RF	Full	*Favia*	*Hydnophora*	*Montastrea*	*Pavona*	*Pocillopora*	*Porites*	*Stylophora*
Overall	X					X		X
Growth	X	X				X	X	X
Shrink						X		

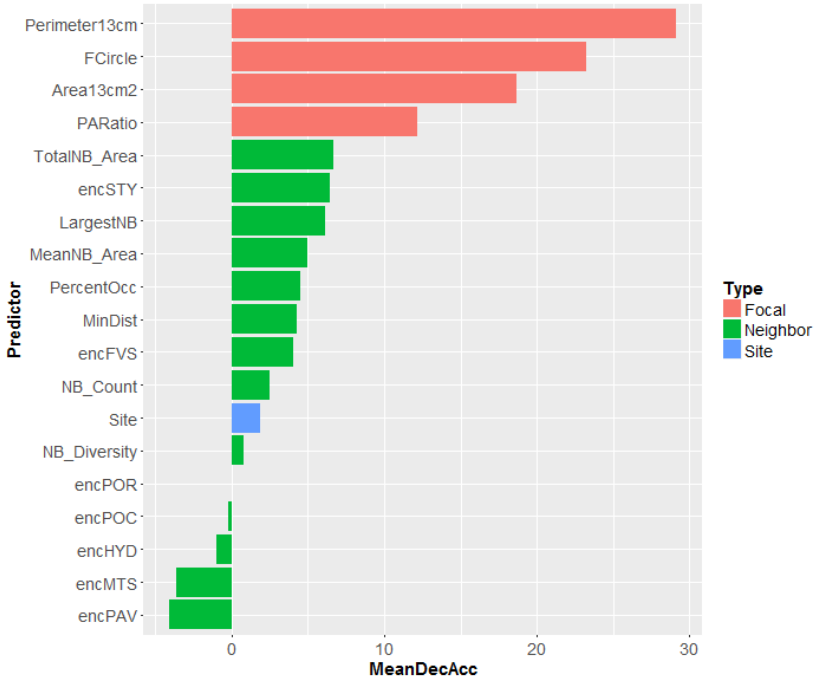

Figure 3: Random Forest Pocillopora Variable Importance

The variable importance calculation random forest is displayed in Figure 3. This figure describes the mean decrease accuracy each variable has when randomly permuted and its importance in prediction of the outcomes. Focal measurements such as perimeter, circularity factor, area, and perimeter-area ratio have the largest values, indicating they have more importance than neighborhood related predictors.

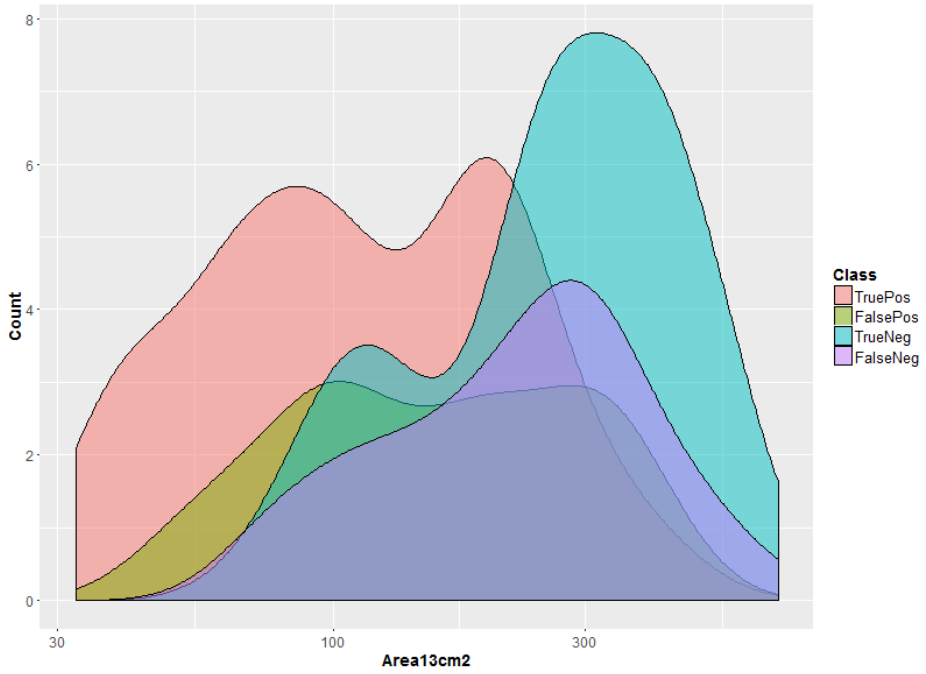

Figure 4: Confusion matrix of random forest prediction – Pocillopora

The confusion matrix in Figure 4 describes the amount of times the outcomes are correctly and incorrectly predicted. This is a density graph plotted with log area distributions of the focal colony. Blue is true negative (correctly labeled Shrink), red is true positive (correctly labeled Growth), purple is false negative (incorrectly labeled Shrink), and green is false positive (incorrectly labeled Growth). The graph shows a separation of Growth and Shrink around 150 cm^2. Biologically, this indicates growth has a pattern of mainly being in juveniles and as a colony grows larger, the higher the chance one will shrink. The false negatives share similar patterns to true negative, which shows the cues of area push the prediction towards separating by area size. Same for false positives. The overlap shows there is still correct predictions along the area distribution but they become rarer as one diverges too far from the main proportion.

From the simulations, the prediction for *Pocillopora* in overall accuracy is 64 +/- 4 % with an average of 8% better prediction than the null. Growth recall is 66 +/- 9 % with an average of 9% better prediction than the null. Shrink recall is 61 +/- 10 % with an average of 8% better prediction than the null. These are performance significant, but perhaps biologically insignificant.

Works Cited

Breiman, L. 2001. *Random forests*. Machine Learning 45:15–32.

Delparte, D., Gates, R. D., Takabayashi, M., & Burns. (2015). *Utilizing Underwater Three-dimensional Modeling to Enhance Ecological and Biological Studies of Coral Reefs.* International archives of photogrammetry and remote sensing, XL(5), 61-66.

Dornelas M, Madin JS, Baird AH, Connolly SR. 2017 *Allometric growth in reef-building corals*. Proc. R. Soc. B 284: 20170053.

Guo T, Capra A, Troyer M, Gruen A, Brooks AJ, Hench JL, et al. *Accuracy Assessment of Underwater Photogrammetric Three Dimensional Modelling for Coral Reefs*. ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2016; XLI-B5(June):821–828.

Lang, 1. 1973. *Interspecific aggression by scleractinian corals. 2. Why the race is not only to the swift.* Bull. Mar. Sci. 23:260-79.

RStudio Team (2015). *RStudio: Integrated Development for R.* RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Sheppard, C.R.C. 1981. *"Reach" of aggressively interacting corals, and relative importance of interactions at different depths*, pp. 363–368, *in* F.O. Gomez etal. (eds.). Proceedings Fourth International Coral Reef Symposium, Manila, Vol. 2: The Reef and Man Marine Sciences Center, University of the Philippines, Manila.

Tanner JE (1997) *Interspecific competition reduces fitness in scleractinian corals*. J Exp Mar Biol Ecol 214: 19–34.