Do Marine Protected Areas Protect Fish? A Study on South Pacific Transects

Victor Cheung

February 22, 2016

Abstract

Objective: To determine the efficacy of MPAs in protecting fish.
Setting: South Pacific coral reefs.
Methods: Data was gathered as transects from one MPA and one non-MPA. Probability sampling was used.
Results: Sandy ocean floor and the area of coral reefs relative to the length of the transect do not affect biomass. The median height and the presence of coral reefs is associated with biomass; estimates of the relationships between these two factors and biomass is skewed by sparsity of data. MPAs and non-MPAs have substantial differences in average biomass after controlling for different characteristics of the transects.
Conclusion: MPAs may be effective tools for conservation or restoring ecosystems to their natural state.

Introduction

Human activity, along with other stressors, can seriously damage fish ecosystems in the ocean.

Marine Protected Areas, as defined by a United States Executive Order 13158, is "Any area of the marine environment that has been reserved by federal, state, tribal, territorial, or local laws or regulations to provide lasting protection for part or all of the natural and cultural resources therein".

Thus, the core idea motivating the existence of Marine Protected Areas is to protect a specific area against stressors, whether commercial fishing activity, other extractive activities, indirect pollution through carbon dioxide emissions due to industry, or so on.

There exist studies that show that MPAs generate economic value through the preservation of aesthetics, employment of coastal protection personnel and uplifting fisheries productivity[1].

[1]Studies done by IUCN in Vanuatu and Fiji evaluating community-based Marine Protected Areas

What about in terms of biomass? MPAs are also meant to preserve the biodiversity and health of the ocean ecosystems. It's key then to evaluate the effectiveness of MPAs on protecting fishes, as measured by biomass. If MPAs protect only aesthetics, but fail to protect the fishes, then we might consider alternative measures of conservation that can fulfill both roles. This possibility is concerning, given the number of MPAs that have been set up around the world and continue to grow in number.

Our study will therefore explore the relationship between biomass and a variety of other factors, including whether an area is an MPA.

Data

Our data is collected from the South Pacific ocean from two separate reefs, one protected, the other not.

We resort to transects as the unit of analysis. Transects are lines of a certain length across the ocean floor. We know that transects were randomly generated and hence randomly sampled from all possible transects along the ocean floor. Divers were sailed over a specific spot, pre-chosen, and from there they make a dive to the bottom. They swim along the transect, and with electronics, estimate the amount of biomass as the sum of the expected mass of the the different fishes captured within some predefined distance of the transect.

The data collected are as follows. Mass, which is the estimated fish biomass along the transect, given in kilograms. Treatment, which denotes whether the data point is from the MPA or not. Prop.HC, refers to the proportion of hard coral along the transect, and Prop.SD, meaning proportion of sandy floor along the transect. Chain, which is a measure of the ratio between the reef surface contour to the length of the reef. And finally, Median.Height, which is the median height of the reef measured along the transect in meters.

Type of Analysis

I expect to conduct a level II analysis. As the population is well-defined, then inferences would be made to the rest of the South Pacific ocean floor. That is, we can therefore generalize to other transects in the South Pacific, such that given the various predictors, we can make a prediction about the expected biomass along that transect.

Within the "wrong model persective", we will consider the data generation process, the originating joint probability distribution, the estimation target, and effects of data snooping. Subject matter expertise informs us that there are a variety of variables, including seasonality, water temperature, migration patterns, climate patterns, etc. that should be included in a full model. Therefore, our model is incomplete and wrong.

This is justified as the population is well-defined in this case. It is the set of all possible transects along the South Pacific ocean floor. We also have that the data was generated via a probability sampling process. The originating joint probability distribution therefore includes the biomass, the treatment, Prop.HC, Prop.SD, Chain, and Median.Height, as well as other unmeasured variables. Thus it seems fair to assume that the observations were generated independently from a joint probability distribution.

Our estimation target will be an approximation to the true response surface that will be determined empirically. We will leave to our algorithms the task of determining the approximation.

As we are conducting level II analysis, data snooping will seriously damage the generalizability of our fitted model. We would be overconfident in our inferences as a result of fitting noise. As I intend to make use of bootstrapping to optimize tuning parameters and to select a model based on minimization of the RMSE, I will hold out a portion of the data for an "honest" assessment of its out-of-sample performance.

Univariate Statistics

We will conduct exploratory data analysis through univariate statistics and histograms. We start with the entire data set. Then we will consider the two reefs separately.

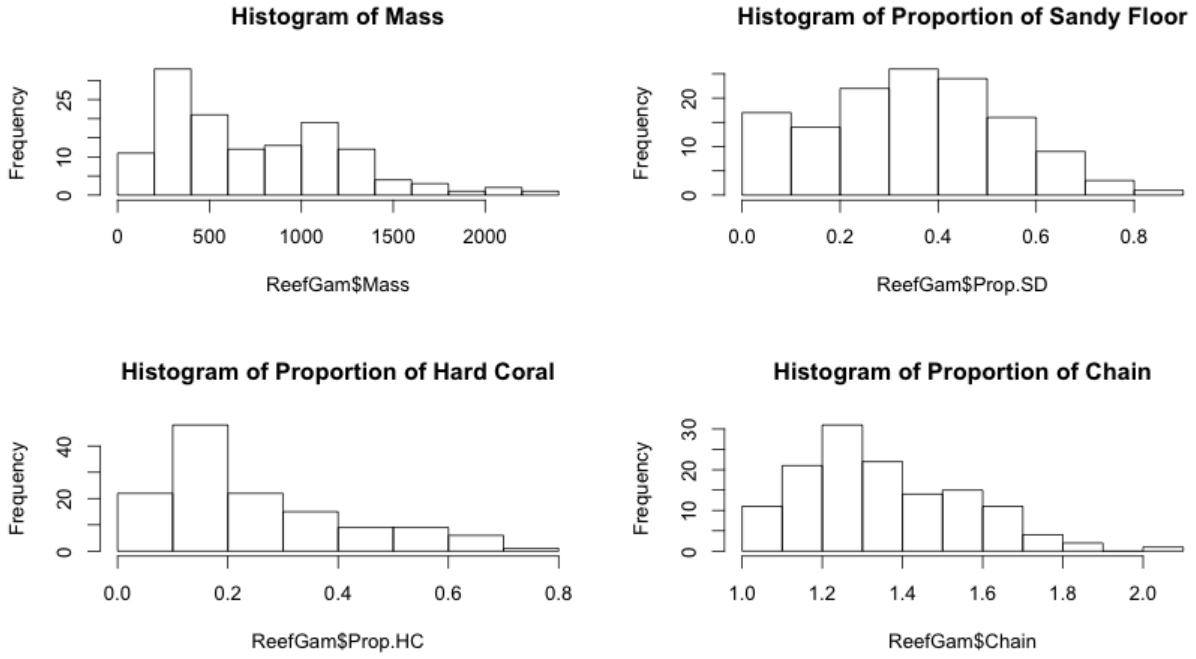

Figure 1: Histograms for ReefGam Data 1

First, we see from the histograms that data for all predictors and responses are right skewed. This implies that data is sparse for extreme large values. Given we only have 132 observations on 5 predictors, this is troubling. If the relationship between these predictors and the responses are nonlinear under the true response surface for those values which are sparse, we will fail to approximate the response surface.

Second, the data looks clean. Proportions data do not exceed 1 and do not go under 0. However, we are missing realizations of Prop.SD past 0.875, and realizations of Prop.HC past 0.725. This implies we have no idea how the true response surface behaves with high

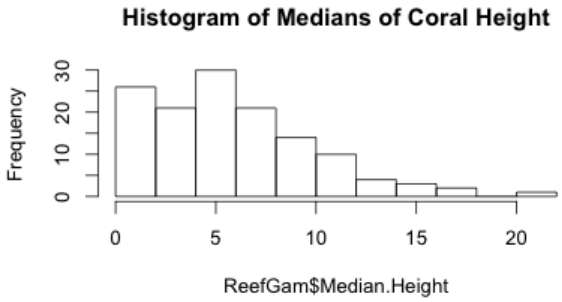

Figure 2: Histograms for ReefGam Data 2

Table 1: Summary Statistics for ReefGam

Statistic	N	Mean	St. Dev.	Min	Median	Max
Mass	132	735.379	489.863	81.000	618.500	2,390.000
Prop.HC	132	0.260	0.169	0.000	0.200	0.725
Prop.SD	132	0.359	0.192	0.000	0.350	0.875
Chain	132	1.357	0.203	1.030	1.320	2.040
Median.Height	132	6.182	4.331	0	6	21

proportions of either sandy floor or hard coral; it limits our ability to generalize to other transects with those extreme characteristics.

Third, in terms of central tendency, the mean and median for the response and the predictors are fairly close, with the exception of mass. As the histogram indicates, since mass is right skewed, the mean is pulled farther to the right when compared with the median.

Table 2: Summary Statistics for Non-MPAs

Statistic	N	Mean	St. Dev.	Min	Median	Max
Mass	76	475.066	335.914	81.000	338.500	1,510.500
Prop.HC	76	0.262	0.180	0.000	0.200	0.725
Prop.SD	76	0.324	0.216	0.000	0.300	0.875
Chain	76	1.370	0.225	1.030	1.330	2.040
Median.Height	76	5.737	4.509	0	5.5	21

Consider now the non-protected reefs. The summary statistics for Prop.HC, Prop.SD, Median.Height and Chain look very similar to the pooled statistics. However, the mean of the non-MPAs is below the pooled mean, with a lower standard deviation as well. We note also that the highest proportions of hard coral and sandy floor appear in the transects of the non-protected reefs as well. Finally, the mean and medians of the response and the predictors are fairly close, suggesting somewhat symmetrical distributions.

Table 3: Summary Statistics for MPAs

Statistic	N	Mean	St. Dev.	Min	Median	Max
Mass	56	1,088.661	443.815	227.000	1,101.500	2,390.000
Prop.HC	56	0.258	0.154	0.025	0.200	0.625
Prop.SD	56	0.407	0.143	0.100	0.400	0.750
Chain	56	1.339	0.168	1.080	1.305	1.730
Median.Height	56	6.786	4.039	0	6	18

Consider now the protected reefs. The mean is substantially higher than that of the non-protected reefs. The standard deviation thereof is larger as well. We note that the maximum observed biomass of the pooled transects was realized in the protected reefs. Furthermore, we note that Prop.HC and Prop.SD do not reach their maximum observed values in the protected reef. The means and medians of the response and the predictors are also close, again suggesting a symmetrical distribution.

Consider now the histograms for the protected and non-protected reefs separately.

Examination confirms our most of our initial suspicions. Biomass for non-protected reefs appear right skewed, whereas biomass for protected reefs is more symmetrically distributed. However, the protected reefs also contains a higher frequency of moderately sandy

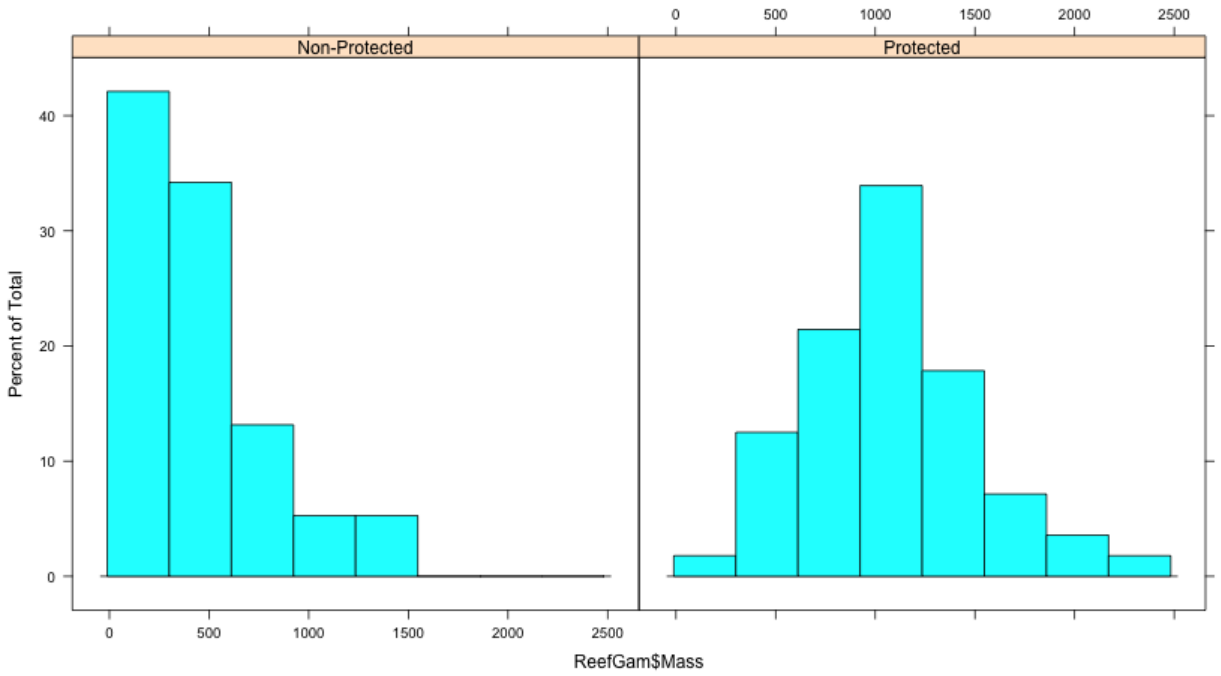

Figure 3: Comparison of Mass across Treatment Groups

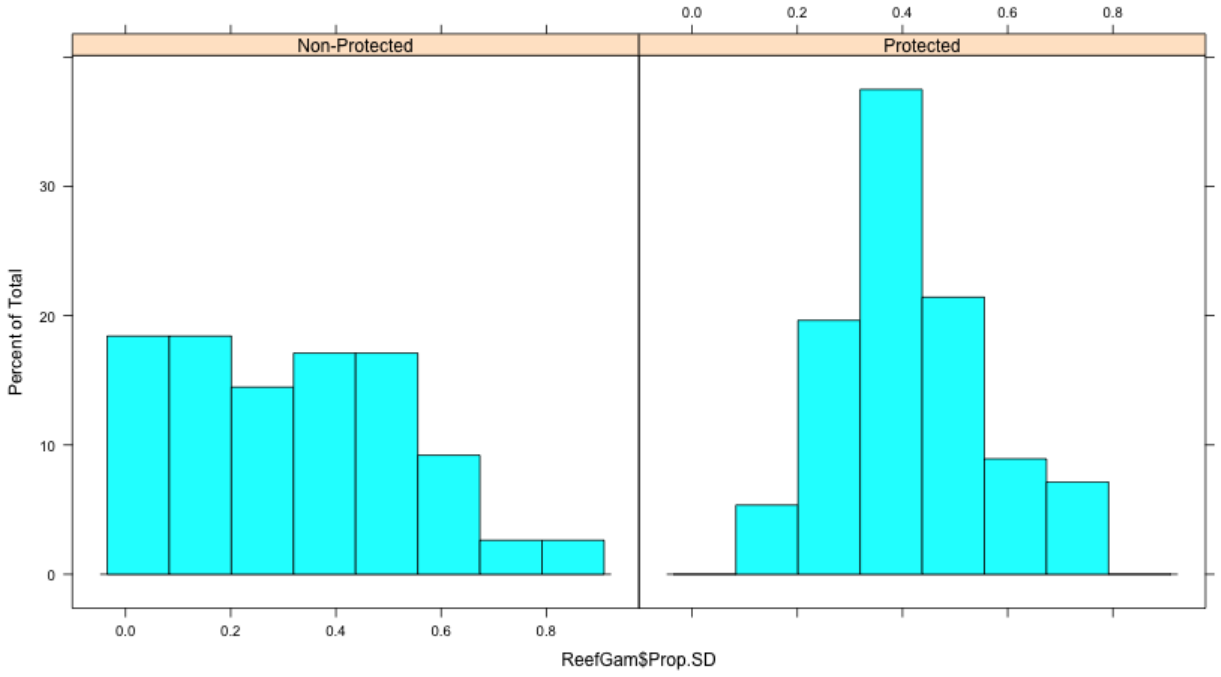

Figure 4: Comparison of Prop.SD across Treatment Groups

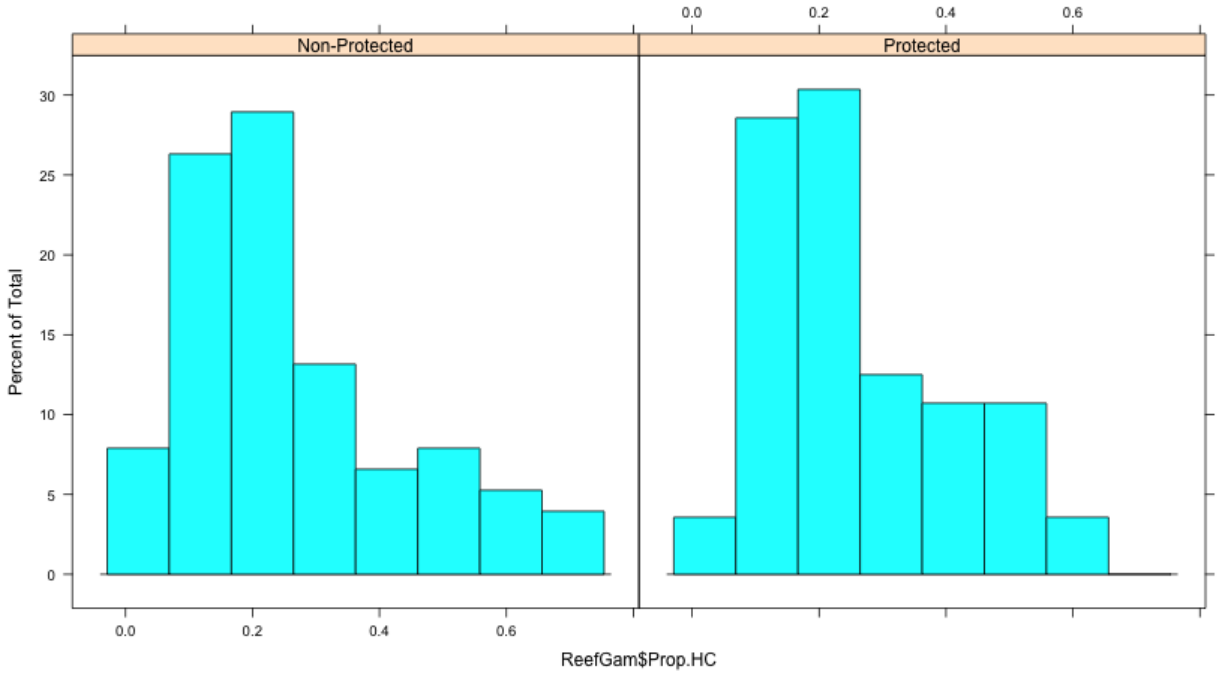

Figure 5: Comparison of Prop.HC across Treatment Groups

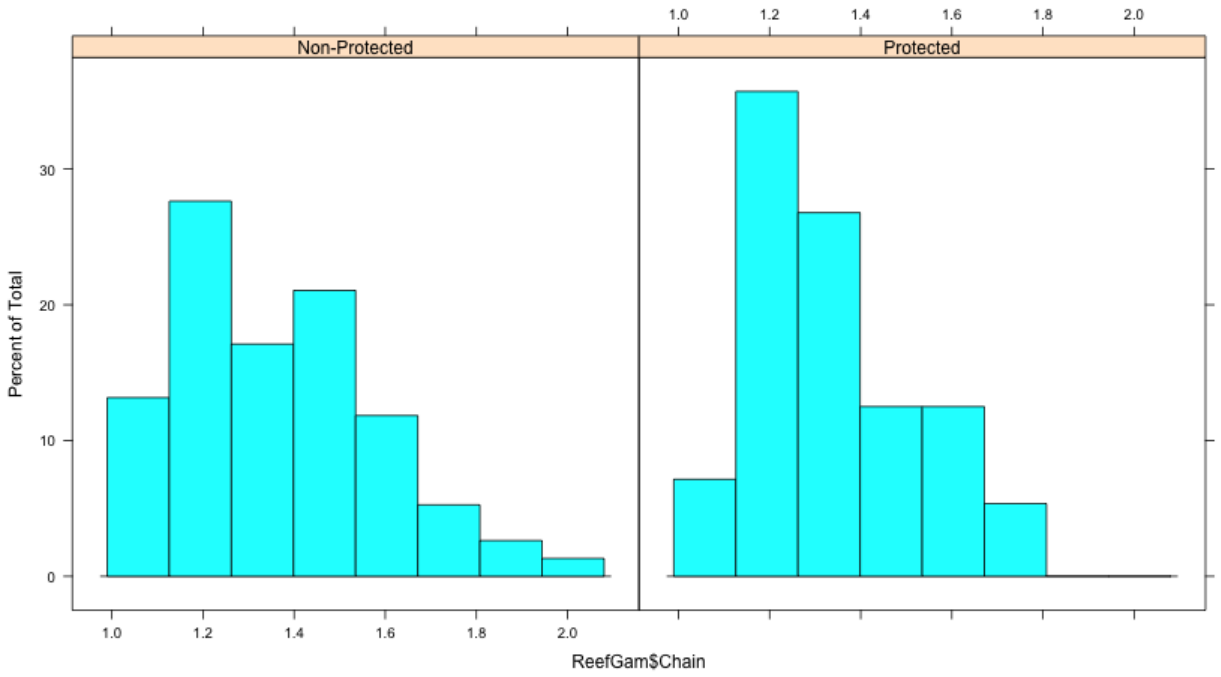

Figure 6: Comparison of Chain across Treatment Groups

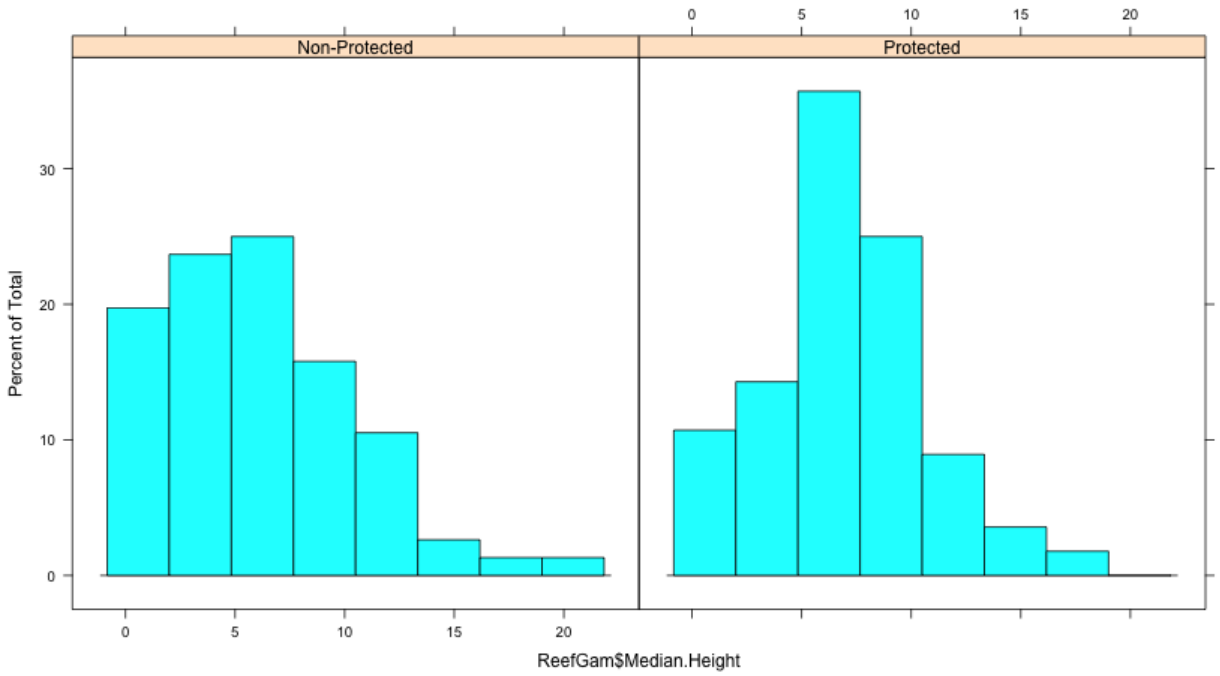

Figure 7: Comparison of Median Height across Treatment Groups

ocean floors. Distributions for proportions of hard coral floor are about the same, except non-protected reefs contain realizations of higher proportions overall. Chain is distributed similarly for both reefs. For Median.Height, protected reefs again contain higher frequencies of moderate median heights.

These differences in distribution may make it difficult to separate the effects of protection from having different proportions of sandy floor or median heights of the reefs.

GAM Model

We used the `caret` library to decide on the best model. We also partition the data into two sets, training data and testing data. With the training data, we bootstrapped for feature selection and tuning parameter optimization. The objective was to minimize the root mean square error. The model below is the best model as outputted.

For our model, about 57.5% of the deviance are explained by the predictors Treatment, Median.Height, Prop.HC and Chain. We note that Prop.SD has been eliminated as a predictor after feature selection.

We will consider each of the terms in turn. We warn that training on a small sample ($n = 108$) means that smooths will be heavily swayed by noise. Therefore, we will treat with skepticism apparent turns and bends for the smooths and defer to subject matter expertise instead.

Prop.SD Prop.SD dropped out of the model as its degrees of freedom shrank to zero. This

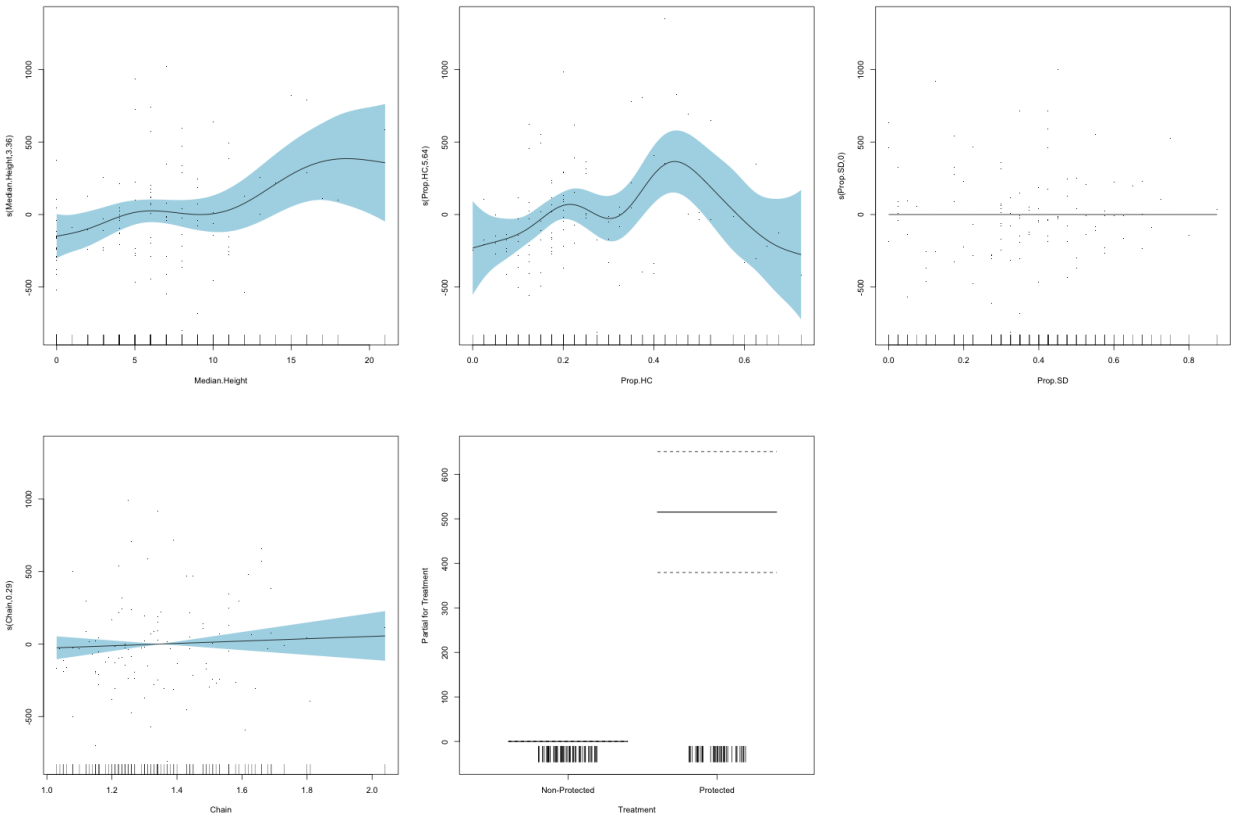

Figure 8: Correlates of biomass for South Pacific Reefs

suggests that for transects, whether or not the ocean floor is sandy has very little relevance with respect to the expected biomass.

Prop.HC Prop.HC has a clear nonlinear relationship with biomass. It is significant with p-value < 0.001. With transects as the unit of analysis, if below approximately 20% of the ocean floor is hard coral, this has an almost linearly positive relationship with biomass. Between about 20% and 45%, the relationship with biomass varies but is generally positive. After around 45%, the positive relationship decreases in a linear way until it becomes negative once again. All this suggests that middle ranges of proportions of hard coral are conducive to more biomass, whereas extreme high and extreme lows are associated with negative relationships on biomass.

Median.Height Median.Height has a nonlinear relationship with biomass. The plot shows something resembling a positive and close to linear relationship with biomass. That median heights of the reefs is correlated with higher biomass makes sense. Higher median heights suggest a more complex coral reef environment, and hence home to a larger number of species, including fishes. This in turn could lead to higher biomass on average.

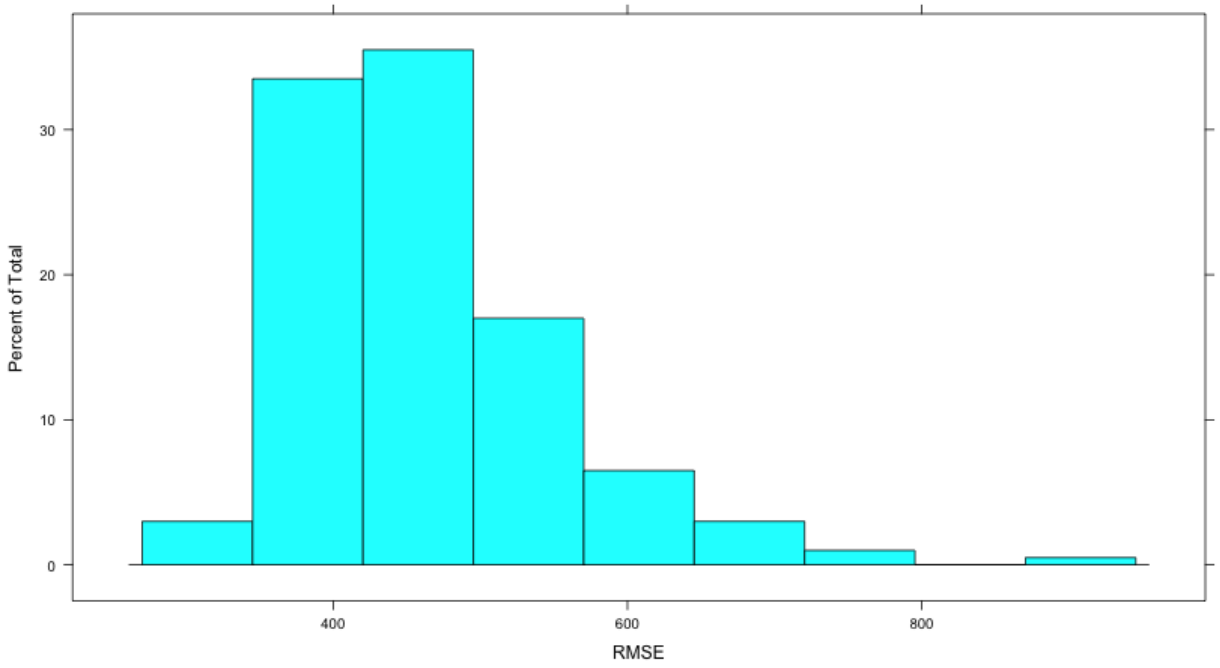

Figure 9: Distribution of RMSEs for Cross Validation

Chain Chain has an effective degree of freedom of 0.2852. Consequently, it exhibits a linear relationship with biomass. The positivity of the association is very much in doubt, as the standard error bands allow for no correlation and negative correlation as well.

Treatment Treatment denotes whether the reef was protected or non-protected. There are dramatic differences in the mean biomass. Protected zones realize, on average, an additional $515kg$ in biomass in comparison with non-protected zones. And even 2 standard errors below that conditional mean we are still anticipating, on average, an additional $400kg$ in biomass. This suggests that MPAs have a dramatic effect on the mean biomass of the transect.

We evaluate the fit by predicting on a test data set.

It's immediately obvious that our model has limited predictive power. Even for the training data, our model fails to make accurate predictions, evident by the deviations from the 45% degree line. We note with interest however that the smoothed fit of the scatter plots lies very close to the reference line. This suggests that on average, our predictions are unbiased; some unexplained factor or irreducible noise stands in the way of better performance.

It's also apparent from the training plot that at lower observed values of biomass, our model tends to overestimate; simultaneously, at high values of biomass, our model tends to underestimate. This suggests the presence of unexplained variables not captured in our model.

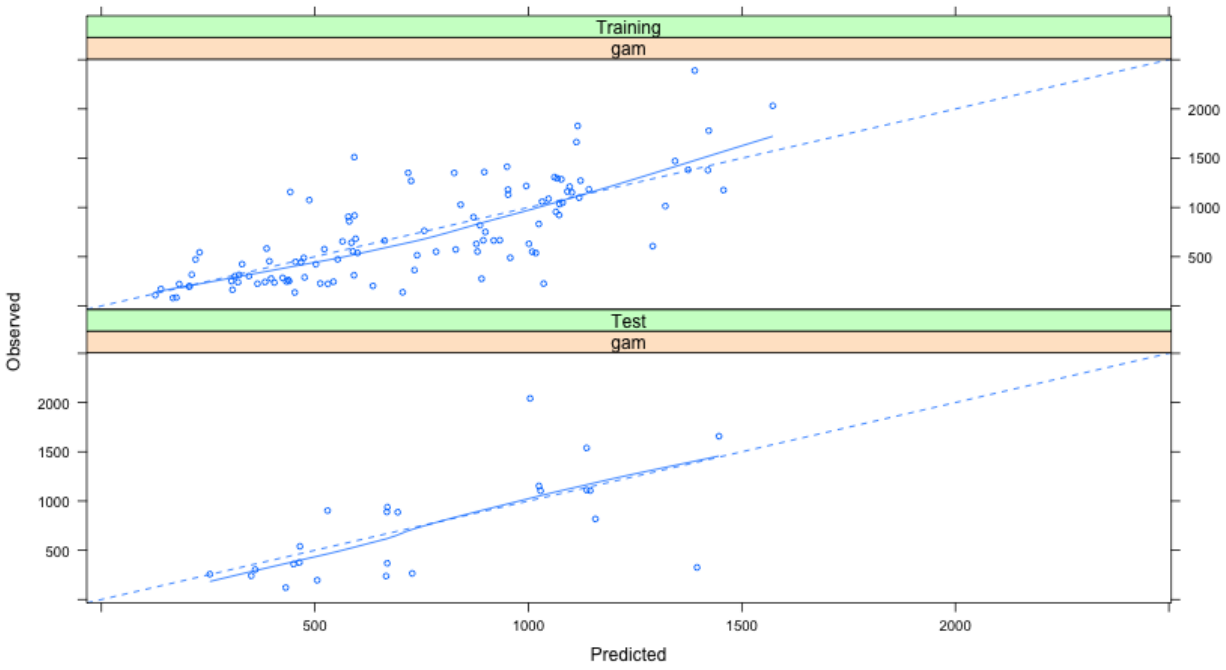

Figure 10: Training and Test Observations against Predictions

A quick calculation of the RMSE shows that RMSE for predicting on the test data set is 386.3236, which falls in the left tail of the distribution shown in figure 3. This suggests that our out-of-sample fit is fairly good, although we can almost certainly consider the test RMSE to be overly optimistic, considering the distribution of RMSEs during cross validation. Nonetheless, this "honest" assessment of our model's performance suggests that our model captures something essential about the relationship between biomass and the various predictors.

Discussion

Supposing that our assumptions regarding the data generation process holds, that our originating joint probability distribution exists, and that our model is a fair approximation of the true response surface, it appears that MPAs have an enormous impact on the biomass.

From the univariate statistics it was immediately apparent that the biomass captured in protected zones was substantially higher than in non-protected zones. This is to be expected. If stressors have previously depleted the biomass in that specific reef, then the prohibition of fishing activity, the prohibition of further industrial contamination or of human activity in general would help to restore the ecosystem to its original state.

Subject matter informs us of the following. We would expect that the biomass in both reefs are either the same, or that the biomass in the protected areas be on average higher than the biomass in the non-protected areas.

```
Parametric coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          515.10      43.08  11.957  < 2e-16 ***
TreatmentProtected   515.16      67.84   7.594 1.96e-11 ***

Approximate significance of smooth terms:
                       edf Ref.df     F p-value
s(Median.Height) 3.361e+00      9 1.160  0.0128 *
s(Prop.HC)       5.635e+00      8 2.366  0.0031 **
s(Prop.SD)       1.087e-07      9 0.000  0.5912
s(Chain)         2.852e-01      4 0.107  0.2179

R-sq.(adj) =   0.53   Deviance explained = 57.5%
GCV score = 1.2313e+05  Scale est. = 1.1027e+05  n = 108
```

Figure 11: Final GAM Coefficients and Tests

The first scenario occurs when the MPAs are in place for any reason, including overfishing (tragedy of the commons type scenarios), industrial activity, and so. As we know, MPAs can also be imposed to preserve environmental aesthetics. In these cases, protection may restore biomass to its natural levels. However, for this scenario to materialize, we would need that the non-protected reef is not being over-fished or contaminated.

The second scenario occurs when the MPAs are in place for any reason once again; the non-protected areas are contaminated or being over-fished. In that case, the depletion of fish would result in decreased biomass.

Our model point towards this second scenario, where being non-protected results in observations of severe drops in biomass. We must refer to subject matter experts for the exact mechanisms of depletion. However, our models show that none of the other measured predictors can explain the large differences in mean biomass. Our model suggests that the proportion of the transect as sandy floor is not associated with biomass in anyway. The role of chain (the ratio between coral reef area to the length of the reef) is also similarly unclear. We found that proportions of hard coral is significantly positively associated with biomass; we also found that the median height (and thus height) of the reef is positively associated with biomass.

This suggests that the presence of coral reefs may be conducive for increased biomass. This makes sense from a subject matter perspective. Coral reefs are home to creatures that make up the bottom of the ocean food pyramid: sponges, tuncats, crustaceans, worms, mollusks and fish. As the foundation of the ecosystem that serves as home to thousands of difference species and organisms, one could reasonably expect the presence of coral reefs to signal increased biomass, which is what our models suggest.

From our comparison of histograms for the MPAs and non-MPAs, we know that the distribution of Prop.HC for both groups are similar; and in fact non-protected areas contains

the transects with the highest proportion of coral reefs. This may explain the downward trend for the extreme values we observed for the partial response of Prop.HC in figure 8. Even though we expect biomass to increase with Prop.HC, the few observations with the highest values of Prop.HC are all from the non-MPAs, and thus, overfishing or other unexplained stressors have confounded the relationship.

Conclusion

MPAs have been advertised as an effective method of conservation. Studies done by the ICUN have proven its economic value. In this study, we built a forecasting model and found evidence that MPAs also help preserve biomass at natural levels or restore biomass after overfishing. We controlled for the effects of different characteristics of the transects, and found that treatment (MPA status) is associated with, on average, a $515kg$ increase in biomass along the transect The exact mechanism of action would be an interesting topic for future study.

We warn again about the small sample size of data used; the sparsity of values for certain variables as well is cause for concern.

Nonetheless, our study suggests that that fishing prohibition is effective. It appears that imposing MPAs, at least for South Pacific coral reefs, is an effective method of increasing biological activity and improving the health of the ecosystem, insofar as biomass is a proxy for these factors.

R Code

```
load("~/Desktop/ReefGam.rdata")
library(caret)
library(stargazer)
library(lattice)

par(mfrow=c(2,2))
hist(ReefGam$Mass, main="Histogram of Mass")
hist(ReefGam$Prop.SD, main = "Histogram of Proportion of Sandy Floor")
hist(ReefGam$Prop.HC, main="Histogram of Proportion of Hard Coral")
hist(ReefGam$Chain, main = "Histogram of Proportion of Chain")
hist(ReefGam$Median.Height, main = "Histogram of Medians of Coral Height")
summary(ReefGam)
histogram(~ReefGam$Mass | ReefGam$Treatment)
histogram(~ReefGam$Prop.SD | ReefGam$Treatment)
histogram(~ReefGam$Prop.HC | ReefGam$Treatment)
histogram(~ReefGam$Median.Height | ReefGam$Treatment)
histogram(~ReefGam$Chain | ReefGam$Treatment)
treated <- subset(ReefGam, Treatment == "Protected")
```

```
nontreated <- subset(ReefGam, Treatment == "Non-Protected")
summary(treated)
summary(nontreated)

inTrain <- createDataPartition(ReefGam$Mass, p = 4/5, list = F)
trainReefGam <- ReefGam[inTrain,]
testReefGam <-ReefGam[-inTrain,]
trainReefGam$Mass <- NULL
trainClass <- ReefGam$Mass[inTrain]
testClass <- ReefGam$Mass[-inTrain]
gamControl <-trainControl(number = 200)
caretgamFit <- train(trainReefGam, trainClass, method = "gam", scaled = F,
 trControl = gamControl, tuneLength = 5)
predictions <- extractPrediction(list(caretgamFit), testX = testReefGam,
testY = testClass)
plotObsVsPred(predictions)
pred <- predict.gam(caretgamFit$finalModel, testReefGam)
testRMSE <- RMSE(pred, testClass)
testRMSE
```