Lab 6: Habitat Modeling for Blue Rockfish, Carmel Bay, California

Yutian Fang

Introduction

Blue rockfish is an endangered fish species live off the coast of California due to large amounts of recreation and commercial fishing activities. Since those activities have been shut down, managers of the Carmel Blue Rockfish Conservation Association (CBRCA) want to establish an efficient marine protected area (MPA) to protect the endangered blue rockfish species. However, they don't have enough knowledge about the blue rockfish habitat distributions in Carmel bay. Therefore, this lab focuses on using the statistical models in R to predict and ArcGIS pro to map the potential habitats of blue rockfish in Carmel bay, California. The habitat prediction will be used in determining the proper site for the MPA that can effectively protect the blue rockfish species.

Method

- **Analysis Data and variables**

  The dependent variables in the statistical models are blue rockfish presence points that collected from Carmel bay, and pseudo random absence points that generated in ArcGIS. The pseudo absence points are necessary to contrast with the actual presence points in order to determine whether a place is habitat or not. The independent variables, which are the predictors of the models are habitat raster types (habras10), benthic complexity (botc10_8ws), distance from the shelf break (dist_100m), distance from kelp beds (dist_kelp), and bathymetric depth (bathy).

- **Statistical Model Prediction in R**

  The statistical models are being used to predict the probability that a given pixel is likely to have blue rockfish based on the independent variables we think are important, then we decide on the levels of probability that likely to represent a habitat. Both the generalized linear model (GLM) and generalized additive model (GAM) are being used here to predict the result.

  The first step is to use ***Sample*** tool in ArcGIS pro to generate an environmental sample table. This table contains the information of independent variables corresponds to each presence points (as "presence.dbf") and random absence points (as "absence.dbf"). Those two tables are the base of the model prediction (model 1).

  - *Generalized Linear Model (GLM)*

    GLM is a type of binomial logistic regression model. GLM is needed here as the prediction results are expected to be binomial (habitat or not habitat). GLM tends to fit a general linear line to capture the relationship between independent and dependent variables. The model is being adjusted based on the percent of data could be explained by the model (% of deviance explained) and the Akaike Information criterion (AIC). % of deviance explained is calculated by (null deviance-residual deviance)/null deviance. The AIC analysis is conducted by running stepwise AIC, and find out remove which independent variable will produce the lowest AIC. The final model decision is to remove dist_100m, and keep other predictors.

  - *Generalized Additive Model (GAM)*

    GAM is also a type of binomial logistic regression model. Different from GLM, GAM tends to fit a smooth spline curve (rather than linear line) to capture the relationship between independent and dependent variables. The DF (degrees of freedom) of GAM is used to adjust the smoothness of the fitted spline curve. For this study, auto DF is used when generating the GAM. The model is also being examined for % of deviance explained and AIC to determine the best one for the habitat prediction, as the step like GLM above. The final model decision is to keep all the predictors.

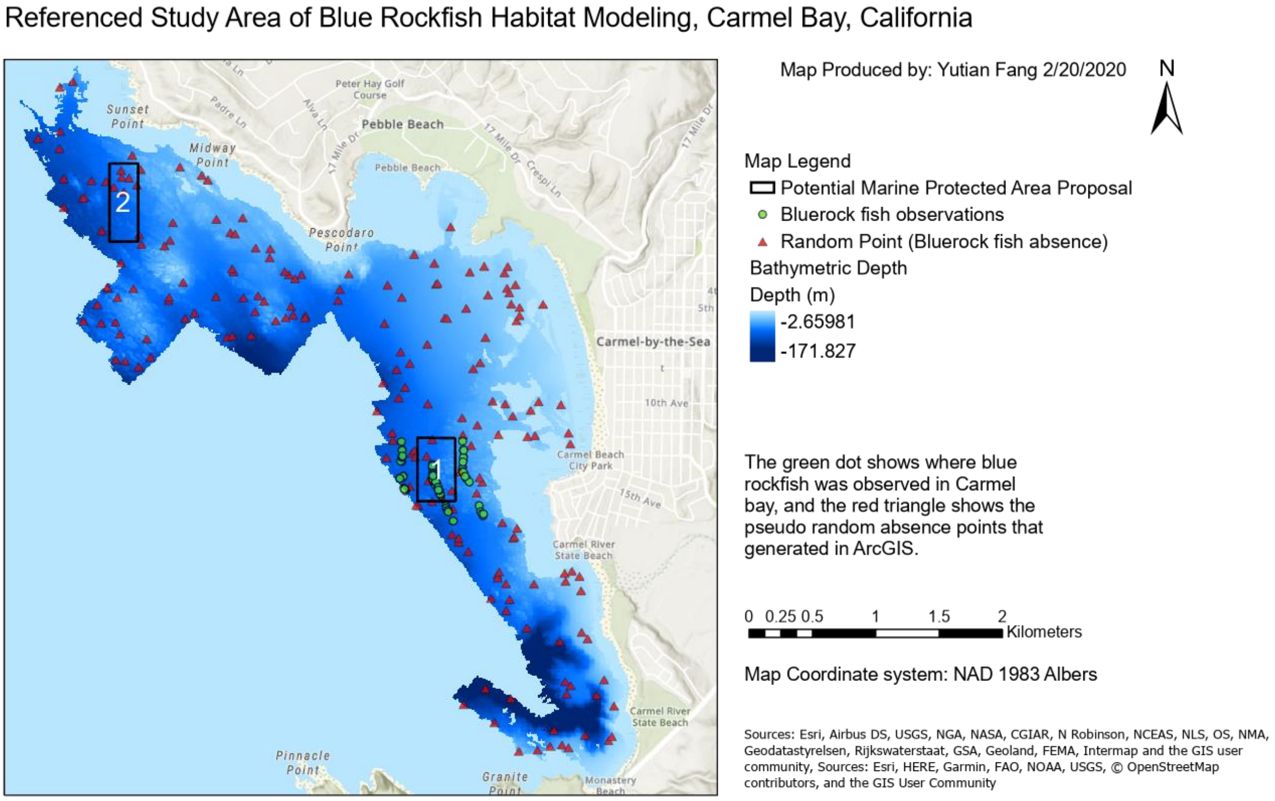
- **Result Presentation in ArcGIS pro**
  - *Generalized Linear Model (GLM)*

    In this step, the prediction result of GLM in R is being converted to map presentation in ArcGIS pro. The estimated coefficients of each independent variables and intercept are being used to compose the formula of the logit prediction (table 2). The formula is calculated in ***raster calculator*** and the result is a logit prediction output. This logit prediction is then being converted back into probability raster of habitat, which ranges from 0 to 1 (the probability of being habitat from 0% to 100%), by using **raster calculator** again. Finally, we run the **raster calculator** to categorize the area that has probability >0.5 as potential habitat, and <0.5 as not habitat (model 2).
  - *Generalized Additive Model (GAM)*

    Because the formula for GAM is too complicated to type in raster calculator, the raster is being processed in R and then directly export the result raster back into ArcGIS pro. Generally, we first get the logit prediction raster, then converts it back to probability raster, then categorize them into habitat or not habitat like the GLM step above.

Result

Map 1: Study area with blue rockfish observation and pseudo random absence points



The green dot shows where blue rockfish was observed in Carmel bay, and the red triangle shows the pseudo random absence points that generated in ArcGIS.

Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Map 2: GLM and GAM probability prediction

Map Legend

🟥 Potential Marine Protected Area Proposal

● Bluerock fish observations

Probability Distribution

Habitat Probability

🟫	≤0.1
🟫	≤0.2
🟧	≤0.3
🟨	≤0.4
🟩	≤0.5
🟩	≤0.6
🟩	≤0.7
🟦	≤0.8
🟦	≤0.9
🟦	≤1

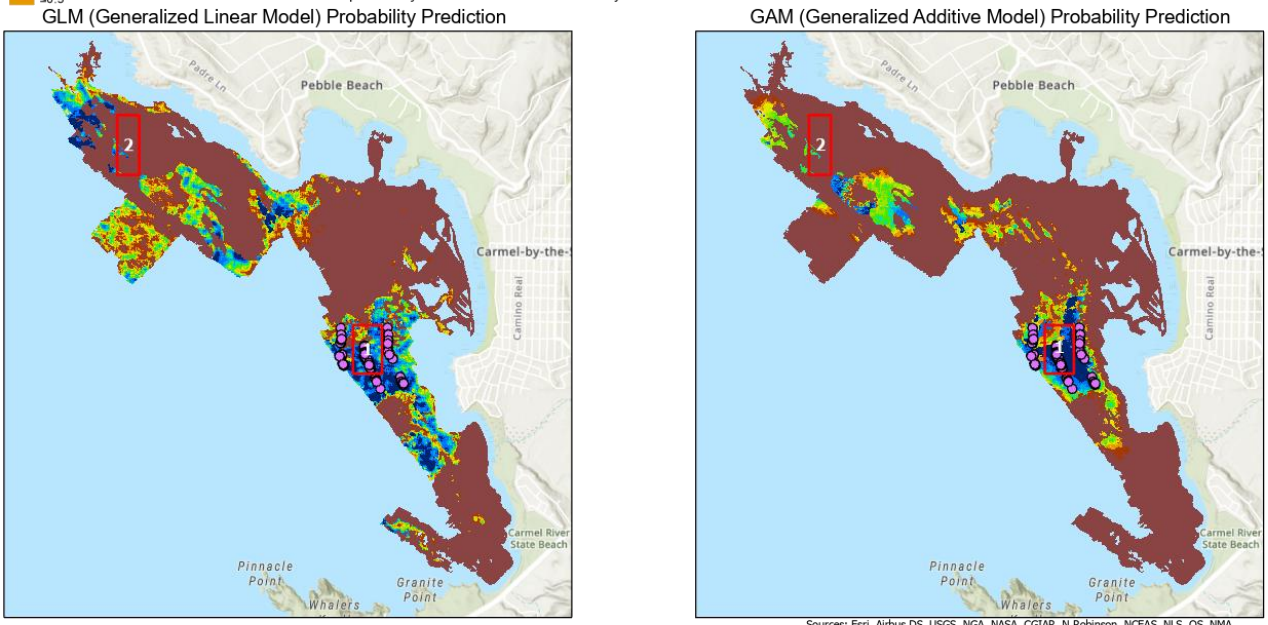
MPA recommendation: I would recommend MPA 1 as the MPA to be established, as it can do a better job in protecting blue rockfish.

The threshold being used here is 0.5. The probability above 0.5 will be viewed as high probability and classified from green to blue. The probability below 0.5 will be viewed as low probability and classified from brown to yellow.

GLM and GAM Model Prediction of Blue Rockfish Probability Distribution in Carmel Bay, California

Map Produced by: Yutian Fang 2/20/2020

N


GLM (Generalized Linear Model) Probability Prediction
GAM (Generalized Additive Model) Probability Prediction
Padre Ln
Pebble Beach
2
1
Carmel-by-the-
Camino Real
Carmel River State Beach
Pinnacle Point
Granite Point
Whalers

0 0.250.5 1 1.5 2 Kilometers

Map Coordinate System: GCS GRS 1980 IUGG 1980

Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Map 3: GLM and GAM Habitat and Not Habitat Prediction

Map Legend

□ Potential Marine Protected Area Proposal

● Bluerock fish observations

Predicted Habitat

🟧 Not Habitat

🟩 Habitat

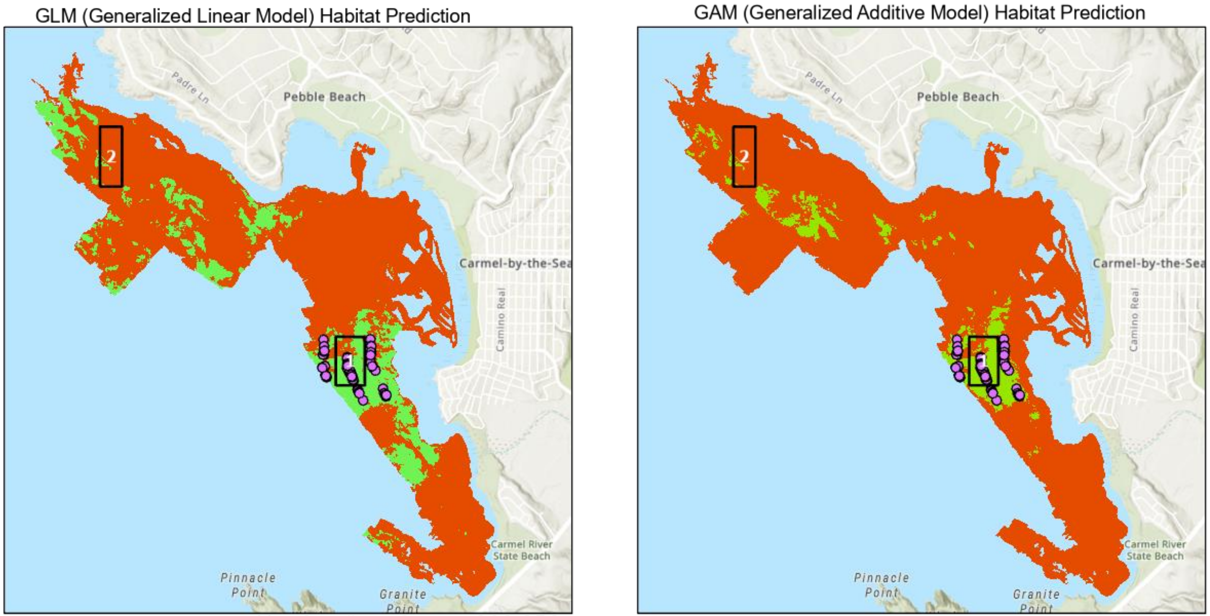
MPA recommendation: I would recommend MPA 1 as the MPA to be established, as it can do a better job in protecting blue rockfish.

Note: If probability of being blue rockfish habitat is >0.5, then the area counted as potential habitat. Otherwise, it counted as not habitat

GLM and GAM Habitat Model Prediction of Blue Rockfish Habitat in Carmel Bay, California

Map Produced by: Yutian Fang 2/20/2020

N

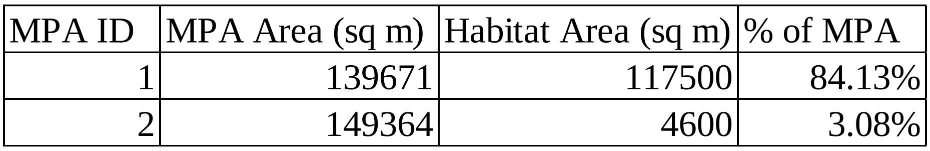


0 0.250.5 1 1.5 2 Kilometers

Map Coordinate System: GCS GRS 1980 IUGG 1980

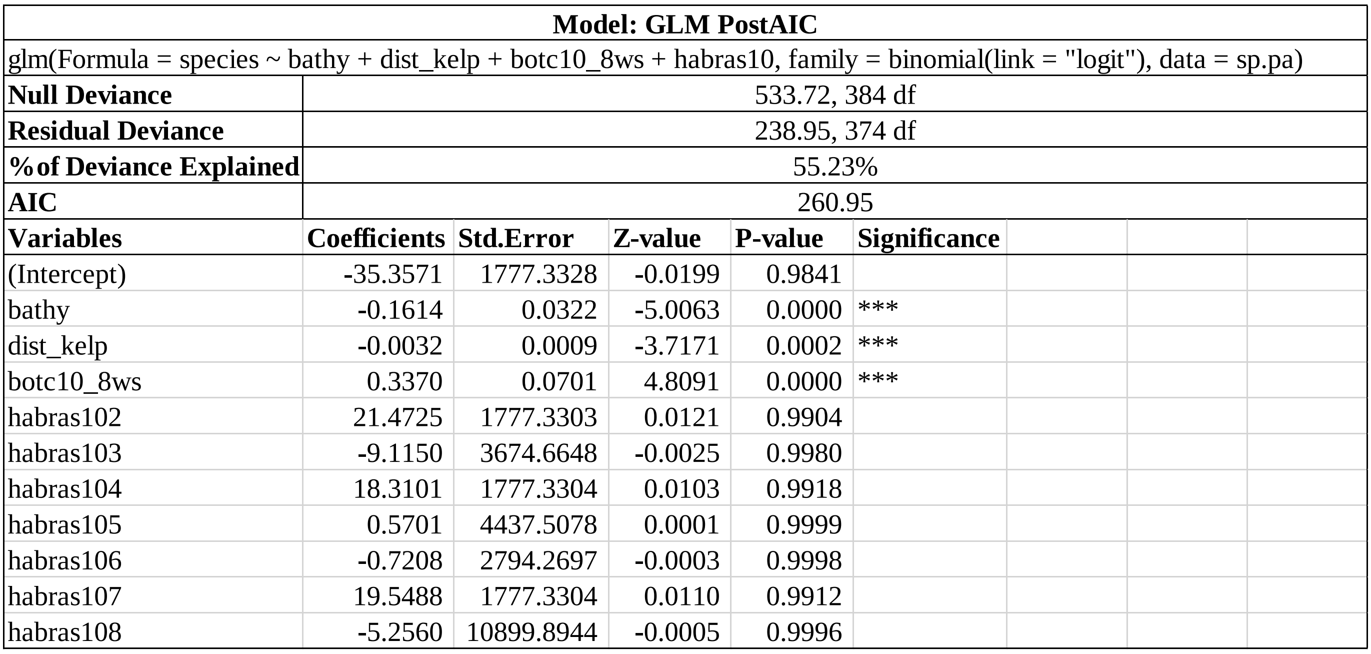
Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Table 1: MPA result for GAM model (best model chose)



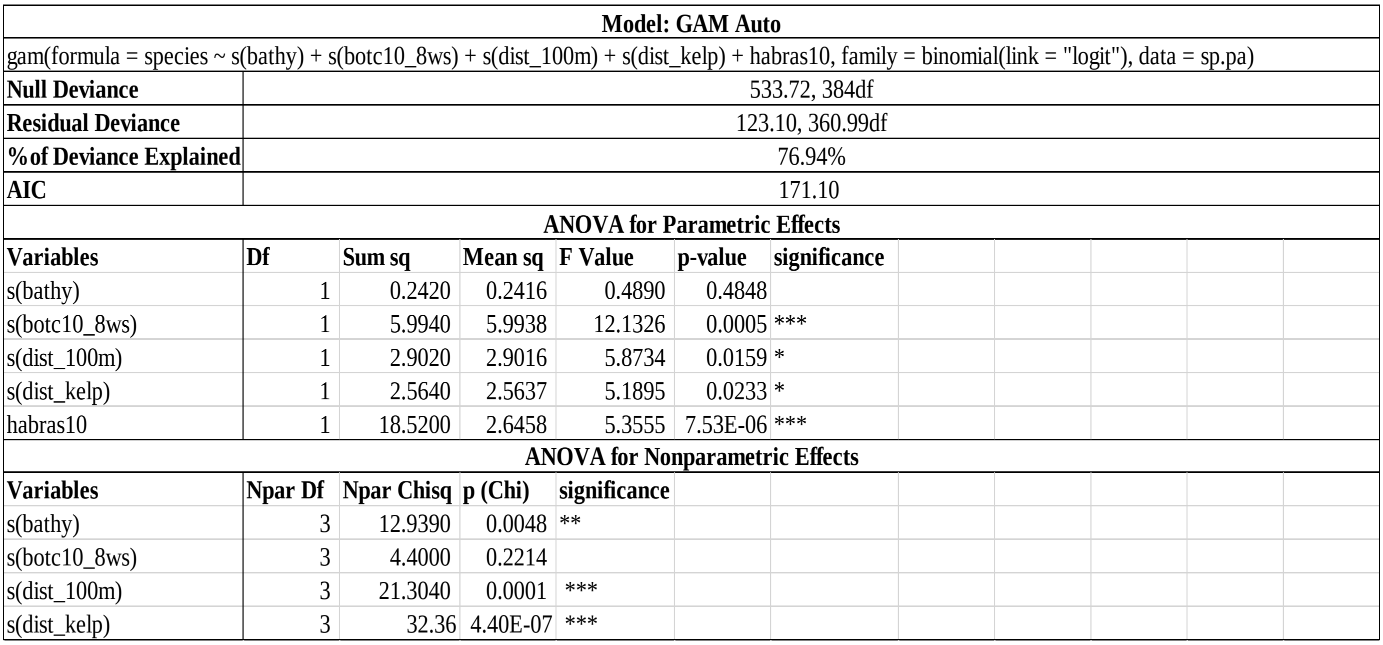
MPA ID	MPA Area (sq m)	Habitat Area (sq m)	% of MPA
1	139671	117500	84.13%
2	149364	4600	3.08%

Table 2: Regression Result for best GLM model in R



Model: GLM PostAIC					
glm(Formula = species ~ bathy + dist_kelp + botc10_8ws + habras10, family = binomial(link = "logit"), data = sp.pa)					
Null Deviance	533.72, 384 df				
Residual Deviance	238.95, 374 df				
% of Deviance Explained	55.23%				
AIC	260.95				
Variables	**Coefficients**	**Std.Error**	**Z-value**	**P-value**	**Significance**
(Intercept)	-35.3571	1777.3328	-0.0199	0.9841	
bathy	-0.1614	0.0322	-5.0063	0.0000	***
dist_kelp	-0.0032	0.0009	-3.7171	0.0002	***
botc10_8ws	0.3370	0.0701	4.8091	0.0000	***
habras102	21.4725	1777.3303	0.0121	0.9904	
habras103	-9.1150	3674.6648	-0.0025	0.9980	
habras104	18.3101	1777.3304	0.0103	0.9918	
habras105	0.5701	4437.5078	0.0001	0.9999	
habras106	-0.7208	2794.2697	-0.0003	0.9998	
habras107	19.5488	1777.3304	0.0110	0.9912	
habras108	-5.2560	10899.8944	-0.0005	0.9996	

Table 3: Regression Result for best GAM model in R



Model: GAM Auto						
gam(formula = species ~ s(bathy) + s(botc10_8ws) + s(dist_100m) + s(dist_kelp) + habras10, family = binomial(link = "logit"), data = sp.pa)						
Null Deviance	533.72, 384df					
Residual Deviance	123.10, 360.99df					
% of Deviance Explained	76.94%					
AIC	171.10					
ANOVA for Parametric Effects						
Variables	**Df**	**Sum sq**	**Mean sq**	**F Value**	**p-value**	**significance**
s(bathy)	1	0.2420	0.2416	0.4890	0.4848	
s(botc10_8ws)	1	5.9940	5.9938	12.1326	0.0005	***
s(dist_100m)	1	2.9020	2.9016	5.8734	0.0159	*
s(dist_kelp)	1	2.5640	2.5637	5.1895	0.0233	*
habras10	1	18.5200	2.6458	5.3555	7.53E-06	***
ANOVA for Nonparametric Effects						
Variables	**Npar Df**	**Npar Chisq**	**p (Chi)**	**significance**		
s(bathy)	3	12.9390	0.0048	**		
s(botc10_8ws)	3	4.4000	0.2214			
s(dist_100m)	3	21.3040	0.0001	***		
s(dist_kelp)	3	32.36	4.40E-07	***		

Map 1 shows the reference map of the Carmel bay with blue rockfish observation points and pseudo random points locations. From map 1, we can see that blue rockfish observations are clustered in a specific region, which may indicate the potential habitat of them. The base layer shows one of the independent variables, which is bathymetric depth.

Map 2 shows the blue rockfish presence probability prediction from both GLM and GAM.

The low probability is demonstrated as brown to yellow color, and high probability is demonstrated as green to blue color. From map 2, we can see that 1) GLM has higher blue rockfish presence probability prediction overall than GAM, and 2) MPA proposal 1 seems to do a better job in protecting blue rockfish than MPA 2 as the blue rockfish presence probability is higher in MPA 1 for both models (most of the color in MPA 1 ranges from green to blue, which represents a high probability).

Map 3 shows the categorized habitat result for both GLM and GAM model. For pixels where blue rockfish presence probability is greater than 0.5 (shown as colors above from brown to yellow in map 2), they are being categorized as habitat, and demonstrated as green color in map 3. For pixels where blue rockfish presence probability is less than 0.5 (shown as colors from green to blue in map 2), they are being categorized as not habitat, and demonstrated as red color in map 3. From map 3, we can see that 1) GLM has more habitat prediction than GAM, and 2) MPA proposal 1 seems to do a better job in protecting blue rockfish than MPA 2 as MPA 1 contains more habitats than MPA2.

Table 1 shows the MPA result for GAM, as GAM is a better model prediction here than GLM (see discussion section for detailed reason). From table 1, we can clearly see that MPA 1 has larger habitat area and percent of habitat covers than MPA 2. Therefore, MPA 1 should be more efficient in protecting blue rockfish. This also prove our direct observation result from map 2 and 3.

Table 2 and 3 show the regression results for GLM and GAM respectively. The coefficients are used in composing the prediction formula of the models, and the p-value demonstrate the significance of the relationship between independent and dependent variables. The comparison between two tables tells us which of them is better. For % of deviance explained, we can see GAM is higher than GLM (76.94%>55.23%), which means more deviance could be explained by GAM than GLM. For AIC, we can see GAM is lower than GLM (171.1<260.95), which means GAM has better quality in model predication than GLM.

Discussion

The assumption of the GLM and GAM models above is the random absence points of blue rockfish. Different from the blue rockfish presence points, those absence points are being generated in ArcGIS, rather than being collected in reality. In order for the models to make prediction, we have to assume that blue rockfish are not present at those points, so our models can contrast the independent variables in presence and absence locations, and tell us whether there is a difference exist between them. Furthermore, the null hypothesis of the model is that the independent variables do not have correlation with the dependent variables, which means the predictors we select cannot tell us the difference between our dependent variables, which is the presence and absence points.

The difference between presence vs. random absence points rather than absence points is worth noticing in this study. Random absence points means we randomly generate them, and they do not necessarily represent the actual absence points of blue rockfish in reality. In fact, to collect the actual absence points of blue rockfish is very hard. Different from sessile animals like corals, blue rockfish are mobile animals, which means we cannot really tell if they are truly “absent” in one location or they are simply not present at the time when we collect the data. Since the absence points in this study are randomly generated, they could be changed if we use a different set of randomly generated points.

Our model prediction is prescribed to the area we conduct the study, the observation and absence points we have, and also the independent variables we selected. In other places, the blue rockfish may present in different locations, and the parameters of the independent variables may be changed. Therefore, the model result could be changed correspondingly. If we select different independent variables, the model result could also be changed. Furthermore, one thing to notice is that we assume the absence points of blue rockfish in this study. This assumption also increase the uncertainty of the prediction result, as the blue rockfish could still be present at those points in reality.

The binomial GLM result we have tells us the correlation between independent variables and dependent variables. From table 2, we can see that bathymetric depth (bathy), distance to kelp bed (dist_kelp), and benthic complexity (botc10_8ws) show significant relationship with the presence of the blue rockfish. The negative/positive sign shows the direction of this correlation. For example, bathymetric depth shows negative sign, which means that the deeper the water depth, the less likely the blue rockfish will be present. Benthic complexity shows positive sign, which means that the more complex the benthic layer, the more likely the blue rockfish will be present. Furthermore, the null/residual deviance and the % of deviance explained is also noticeable in table 2. Deviance is a measure of good fitness to the model, and high number always means bad fit. Our null deviance is 533.72, which measures how much total deviance we have in the model. Residual deviance is 238.95, which shows how much deviance in the model that cannot be explained by the independent variables we selected. Therefore, the % deviance explained tell us the percent of the deviance that could be explained by our model. Accordingly, the higher the % of deviance explained, the better the prediction of the model.

Our GAM prediction gives us higher % of deviance explained compared to GLM prediction. Furthermore, GAM also gives us less places to be predicted as habitat than GLM prediction. One key difference between GLM and GAM is that GLM fits a general linear line between independent and dependent variables, whereas GAM fits a spline smooth curve between those variables. This means that GAM is more specific to the places we study, and tend to capture the small variances that exist in the environment. In contrast, GLM is more general, and tend to overlook the small variance exist in the specific environment. Therefore, since GAM can give us more specific result, they will have less habitat prediction than GLM.

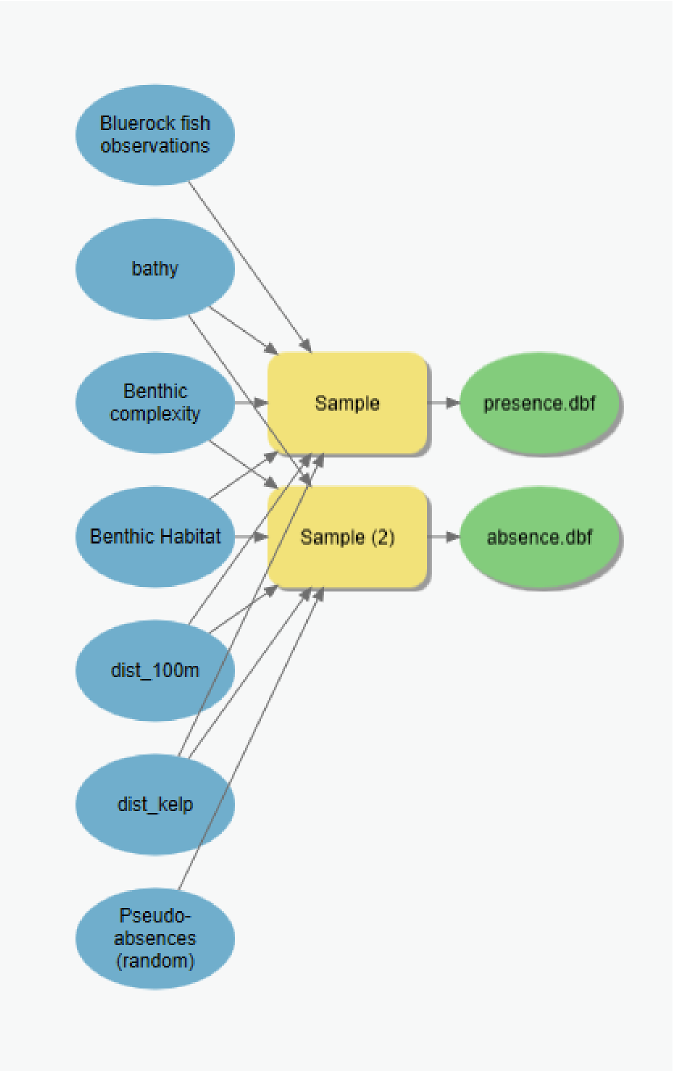
Both of the models did a good job in capturing the actual obervation locations of blue rockfish, as demonstrated in both map 2 and map 3. However, GAM is a better model than GLM for the purpose of this study. Like said above, GAM has higher % of deviance explained than GLM, which means more deviance in the data could be explained by GAM rather than GLM. Furthermore, GAM has lower (and better) AIC, which is another indicator of the model quality. Also, since this study only limited to the region of Carmel bay, which is a specific area, GAM is probably did a better job than GLM in the model prediction. GLM is more suitable for a larger area to make general prediction, whereas GAM is more suitable for a specific area to make specific prediction.

The study could be improved through several ways. For example, we can change a different set of random absence points to compare with the presence points and run the model again. If our model prediction result changes a lot, this could mean our model have serious problems. We should probably collect more blue rockfish presence points in the field to improve our model result. More blue rockfish presence points are also useful for us to do the model accuracy assessment. We could randomly choose some points to run the model, and some points to test the result. Furthermore, we could also select different sets of independent variables to run the model and compare the prediction result with the model we run in this study. This could give us a better sense about the quality of our model, and also makes our prediction to be more accurate.

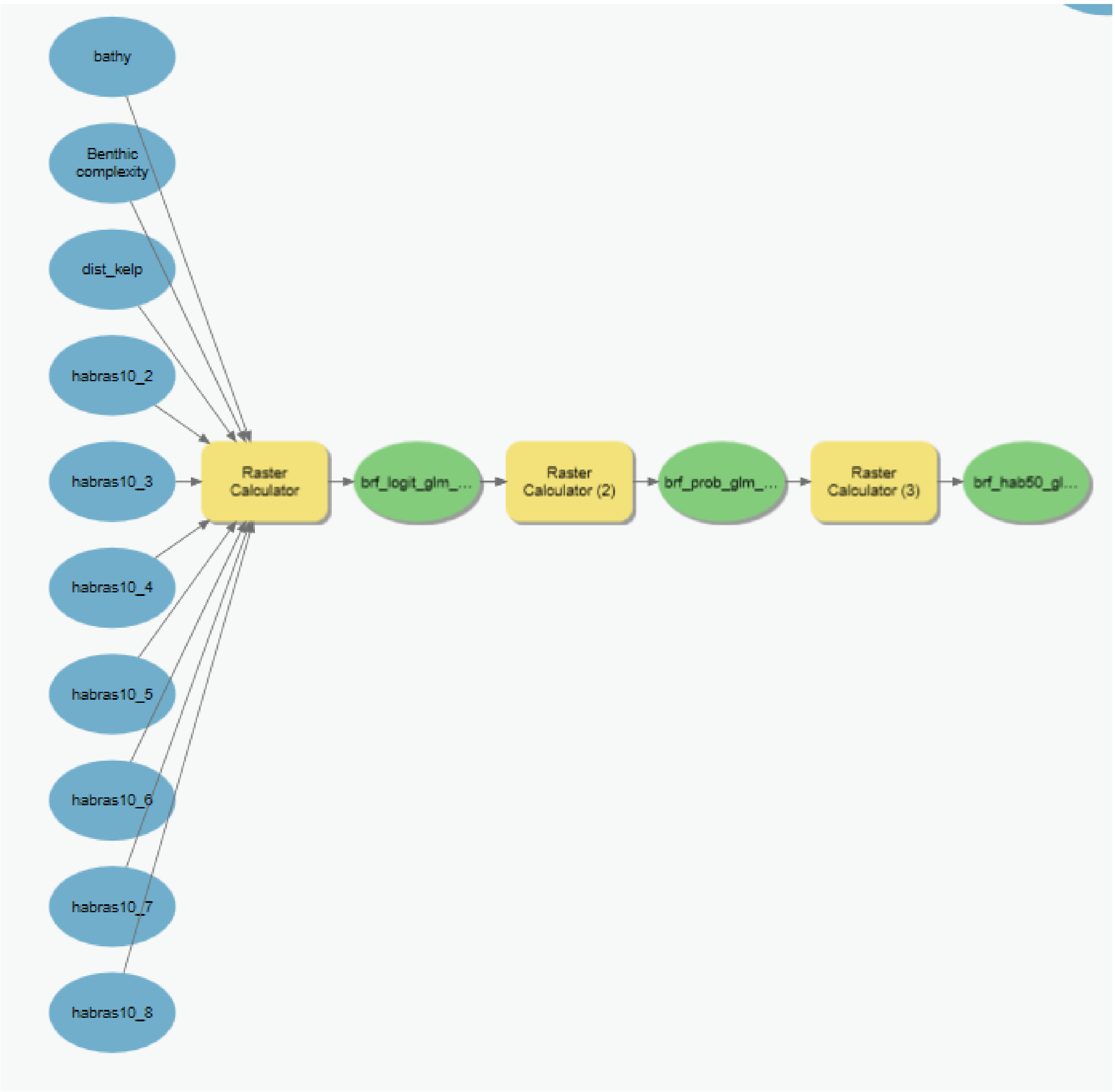
Finally, based on the result of the model prediction and the map representation, I would recommend CBRCA to establish the marine protected area in MPA 1. This area has overall higher probability of blue rockfish presence, and is highly likely to be the potential habitat of blue rockfish. Therefore, this area could satisfy the requirement of the association, and would do a better job in protecting the endangered blue rockfish species.

Appendix

Model 1: Generating the environmental model table



Model 2: Representing the GLM habitat prediction in ArcGIS pro


bathy
Benthic complexity
dist_kelp
habras10_2
habras10_3
habras10_4
habras10_5
habras10_6
habras10_7
habras10_8
Raster Calculator
brf_logit_glm_...
Raster Calculator (2)
brf_prob_glm_...
Raster Calculator (3)
brf_hab50_gl...