Yuqing Xia
Archipelagos, Institution of Marine Conservation

Predicting Common Dolphin Delphinus delphis distribution in Aegean Sea area by machine learning models

Introduction

With sighting data and tracking data in 2016 and 2017, we want to explore how Chlorophyll, Salinity, Sea Surface Height, Sea Surface Temperature, Distance from shore and Slope will influence common dolphin Delphinus delphis' appearance in Aegean Sea area. On the base of key factor analysis, we want to use machine learning models to predict the common dolphins appearance based on known environmental factors' values. We also provided with alternatives for future researchers in order to reach higher accuracy of predicting common dolphins' active area.

Methodology

I. Dataset Introduction

Our dataset composes of two parts, common dolphins appearance points and common dolphin absent points with corresponding environmental variables values. The sample size is 82.

There are 43 common dolphins sighting spots from Archipelagoes, Institute of Marine Conservation from April 2016 to May 2017. With transect data, we have around 76000 GPS points along our regular boat survey routes, we randomly sampled 1 point as absent points on that day's transect if there is one dolphin appearance point.

II. Variables

Variables Name	Meaning	Data Attribute	Resource
P_A	Dolphins' appearance vs dolphins' absence	daily	Archipelagoes, Institute of Marine Conservation boat survey
SST	Sea Surface Temperature	daily	marine.coprenicus.eu
SSH	Sea Surface Height	daily	marine.coprenicus.eu
SAL	Salinity	daily	marine.coprenicus.eu
CHL	ChlorophyII	daily	marine.coprenicus.eu
DIS	Distance from Shore	static	calculated in ArcGIS 10.4 by Near tool
SLO	Slope	static	http://www.gebco.net/

Table 1 Research Variables Introduction

III. Data pre-processing

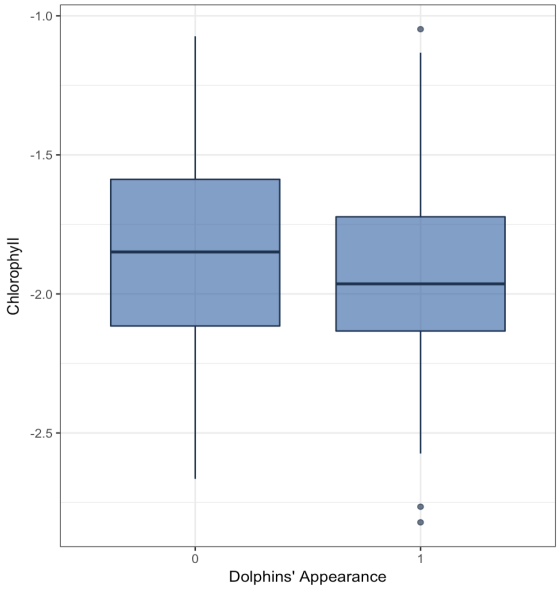
1. Data Collection and geographical data pre-processing

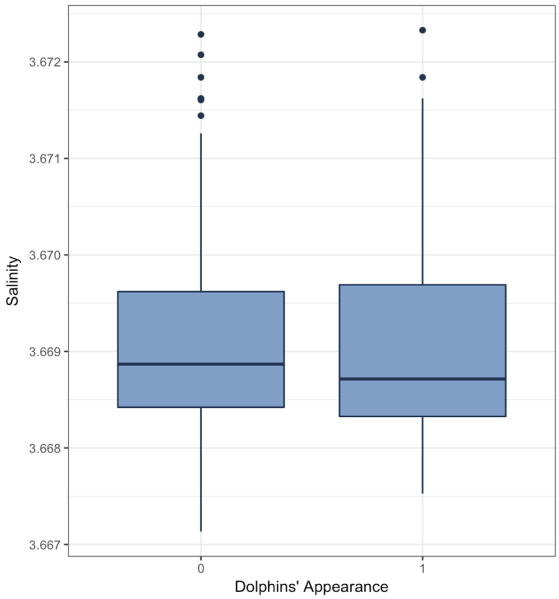
The environmental variables SST, SSH, SAL, CHL are collected from marine.coprenicus.eu. And SLO data are processed based on bathymetry data (downloaded from General Bathymetric Chart of the Ocean), while DIS data are calculated in ArcGIS 10.4 by Near tool.

Some variables, like SST(Sea Surface Temperature), SSH(Sea Surface Height), SAL(Salinity) have missing data on our desired location and thus require us to use interpolation. The interpolation method we used is Inverse Distance Weighted(IDW). Compare to nearest neighbor, which assign unknown points value to the nearest known point value directly, IDW gives the weight of each known points contribution to the unknown points as an inverse proportion to their distance. After interpolation, we then used extract and merge tool to produce shapefiles with all 6 variables data.

2. Data Exploration

From the box plots in Figure 1, we can see that there is no strong relationship between dolphins appearance data with certain factors because those variables' range remain no significant changes when dolphins appearance data change from 0 to 1. From those plot, we can also observe some outliers in the plot, however there is no need to exclude them because some variables(like Slope, Distance from shore) changes dramatically if the spots are observed far away from samos island.





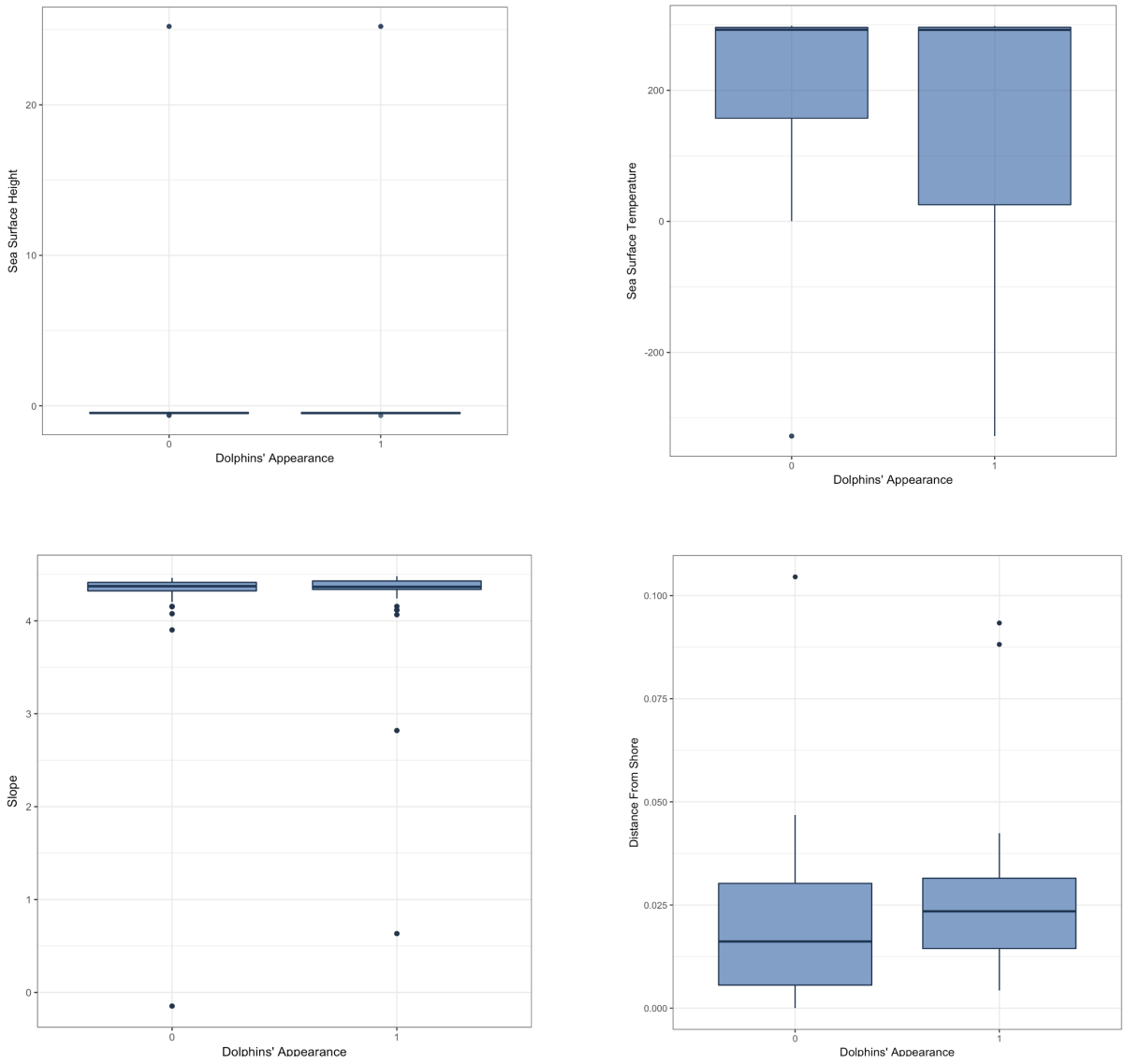


Figure 1 Box plots of 6 variables

3. Data Set Validation

- Negative and Positive data ratio validation

In our model, we have 1 factor data, which is P_A, dolphins' appearance. And all other variables are continuous data. However, we have a very unbalanced dataset because we have 43 positive data with possible 76000 negative data. We must manually adjust the proportion of negative and positive data because dolphin observation is a very random activity, observations of dolphins may differ depends on boat survey time and weather condition. From statistical accuracy consideration, 1:1 ratio between negative and positive data will produce more satisfying results.

- Boat Survey Routes Validation

In order to go further to predict dolphin appearance probability with selected variables, we need to verify the eligibility of our boat survey transect from 2016 and 2017. Due to our limitation of our time, funding and professional facilities and with take concerns of safety issues, we usually travel two main routes, one is horizontal around samos island and other one is vertical route where we head into Lipsi island. (Please reference to the black line in Figure 2)

In order to test the two main routes feasibility, we made yearly average maps of each variables and compare average maps in order to test whether our limited boat survey routes cover variables' ranges in our research area.

As we can see from Figure2(a), our route has covered most data ranges of Chlorophyll data around samos island, we cannot travel to the right side of samos island because it's near Turkey's control. The same situation happens in terms of Sea Surface Height and Sea Surface Temperature, although our boat overlooked areas with temperatures from -0.44 to -0.417, we covered most of the data and were not able to travel into Turkey's territory. While, from Figure2(b), we can see most of the spots cover salinity from 29.14—39.25 and areas with 38.99-39.13 are ignored.

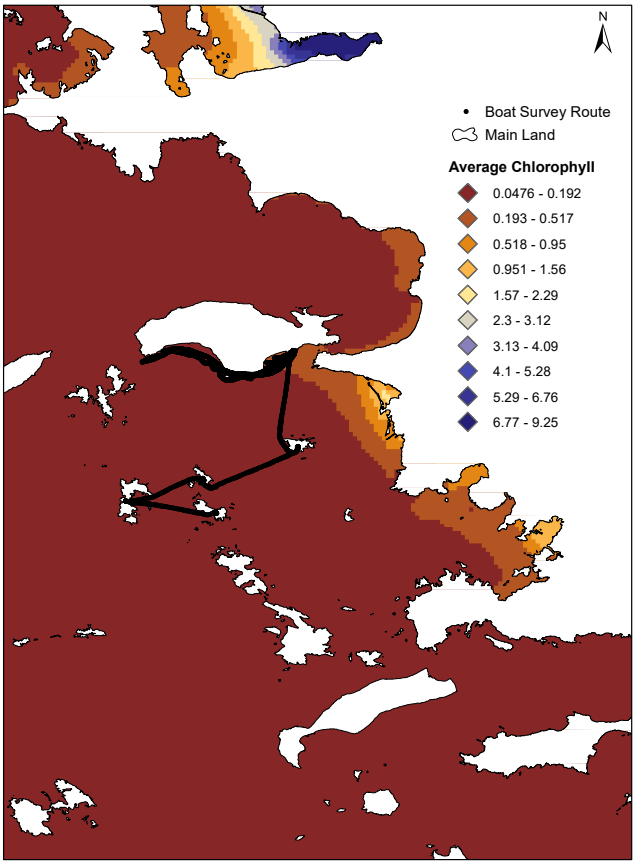


Figure 2 (a) Chlorophyll Average Map

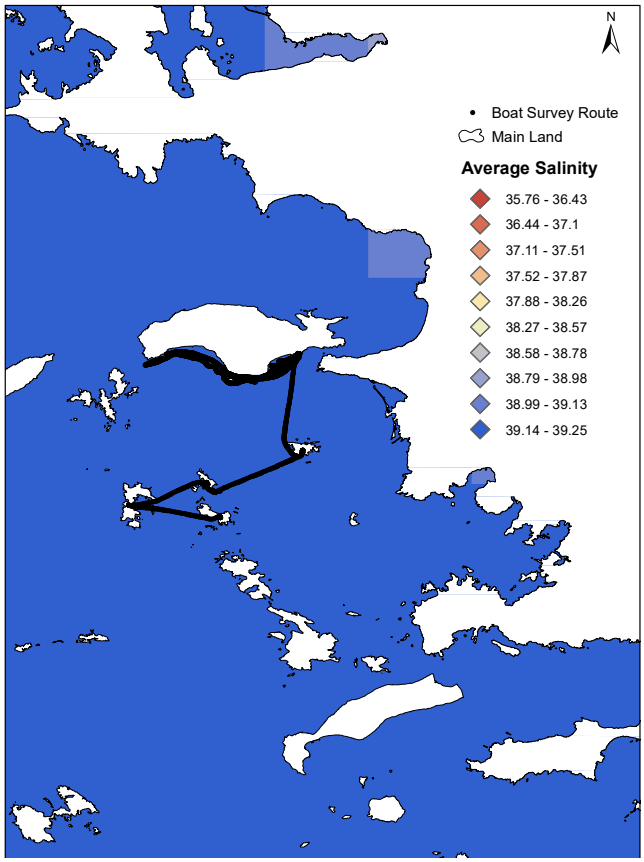


Figure 2 (b) Salinity Average Map

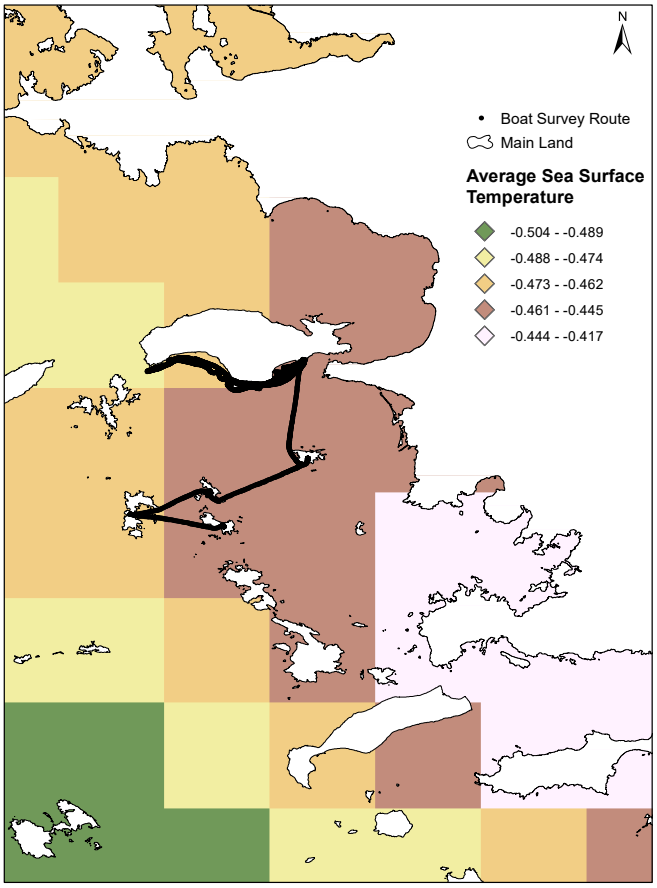


Figure 2 (c) Sea Surface Temperature Average Map

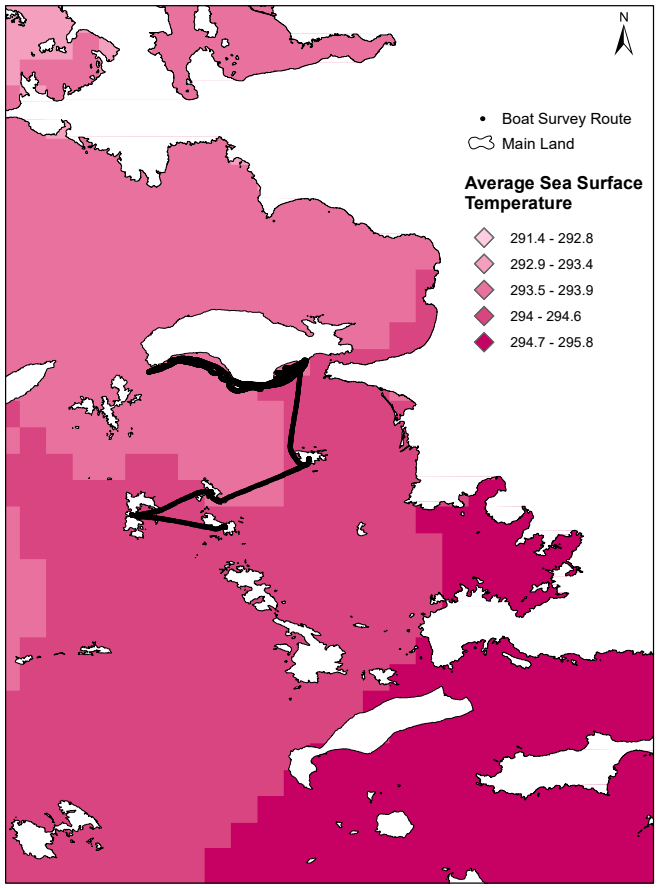


Figure 2 (d) Sea Surface Height Average Map

Overall, chlorophyll, Sea Surface Temperature, Sea Surface Height and Slope values are acceptable. Although Salinity and Distance variables' ranges are close to the average maps, future study should collect more data to better cover those circumstance in our research area. All in all, the boat survey transect is qualified and we can use those data as negative dataset resources.

- Test & Train data

I used 60% of raw data as train data and 40% as test data, because we have a small sample size. Then I used prop.table to test the response and key parameters' proportion in each train and test data set to ensure they have acceptable distance from each other. From the table below, we the P_A distribution in train and test dataset is reasonable.

P_A	0	1
Train	0.4693878	0.5306122
Test	0.4848485	0.5151515

Table 2. Test and Train data validation

- Cross-Validation

We used 10 fold cross-validation to evaluate and validate our train and test dataset.

IV. Model Selection

- Logistic Model

Logistic model is an elementary model for binary classifiers. Similar to linear regression model, logistic model trying to conclude relationship between the mean of dependent variables with independent variables. The difference if that the generalized linear model requires a link function which enables researchers to convert categorical (2-dimensional here) variables into a continuous/ numerical data and then use the same principles of linear regression to fit the model, as indicated below.

$$g(u) = \beta 0 + \beta 1x1 + \ldots + \beta pxp(xT\beta)$$

The link function, g, describes how the mean response, EY =μ, is linked to the covariates through the linear predictor.

So here, our response is the probability of common dolphins appearance. And the model is listed below. We selected all 6 variables as independent variables I used quasi binomial function because it deals more properly for data with high variance, compared to binomial method.

model_glm<- glm(as.numeric(P_A)~ CHL+SAL+SSH+SST+NEAR_DIST+Slope, data=re_train)

We also choose gam model as supplementary model because of its' higher accuracy.

model_gam<- gam(as.numeric(P_A)~ CHL+SAL+SSH+SST+Slope +NEAR_DIST,data=re_train,family = quasibinomial)

- Random Forest

Compare to decision trees, random forest can improve models' accuracy by fitting many trees and each split with fit each one to a bootstrap sample of our data. In addition, with advantage of using cross validation, random forest itself can be resistant to overfitting scenario.

Instead of using common random forest model, I used caret package because it is faster and provide opportunity for us to tune the model by adding tune length, tune grid and separated cross validation method. We used tune length equal to 10 because is more accurate than smaller tune length, and I used 10 fold cross validation method. After defining the tune grid and customizing tune grid, I found when mtry equal to 6, the model reaches best performance. Mtry is important in Random forest because it is the number of randomly select parameters at each split.

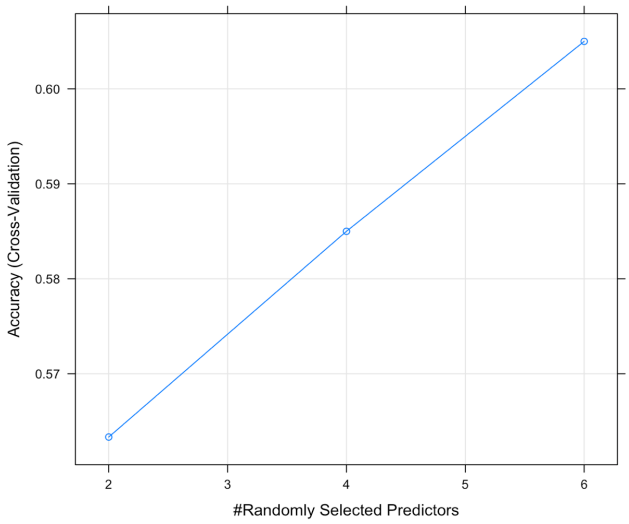


Figure 3 Random Forest Model Predictors Numbers Selection Chart

1. Key factor analysis

From Table, we can see that Distance from shore(NEAR_DIST) is the most important indicator of dolphins' appearance. Part of the reason is that dolphins will seek food near the port because of tourists.

Key Factor	Importance	Significance Level
CHL	3.765256	**
SSH	3.457194	*
SST	2.821028	
SAL	3.915013	**
Slope	2.790663	
NEAR_DIST	5.923566	***

Table 3 Key factor analysis

2. Model Evaluation

We used kappa value as our model evaluation standard. Kappa value ranges from 0 to 1, as we can referenced from table 4, two of our models are of poor agreement and only random forest model reaches kappa value of 0.3294.(table 5)

model_forest <- train(P_A ~ CHL+SSH+SST+SAL+Slope+NEAR_DIST, data = re_train, method = "ranger", trControl = trainControl(method = "cv", number = 10, verboseIter = TRUE))

Evaluation Results	Judgment Standards
Poor agreement	Less than 0.20
Fair agreement	0.20 to 0.40
Moderate agreement	0.40 to 0.60
Good agreement	0.60 to 0.80
Very good agreement	0.80 to 1.00

Table 4 Kappa evaluation standards

Overall, all of our models performance indicate weak relationship between CHL,SAL,SSH,SST, SLO and NEAR_DIST and common dolphins' appearance. One of the reasons is because of our limited sample size. Normally, we would expect to have much larger dataset to reveal relationship or trend in environmental field because of randomness in in-situ measurement. Another reason is our sampling mechanism. I used 1:1 ratio to sample dolphins' absence data based on our current sighting spots as we can see in figure 4. However, dolphins move with great

distance per day, it would be more reasonable if we can sample absence data by adding moving distance on the sighting spots although we will need to record and predict common dolphins moving range per day and there will be mathematical transformations involved. Currently, I suspect the inaccuracy of the models partly because the points I randomly selected as absence data by sampling algorithm are too close to the sighting spots that they have similar value.

Model/Evaluation	Accuracy	Kappa
GLM	0.3636	-0.2669
GAM	0.5758	0.1381
Random Forest(mtry=2)	0.6016667	0.2307692
Random Forest(mtry=4)	0.5933333	0.2141026
Random Forest(mtry=6)	0.6550000	0.3294872

Table 5 Quantitive Model evaluation tables

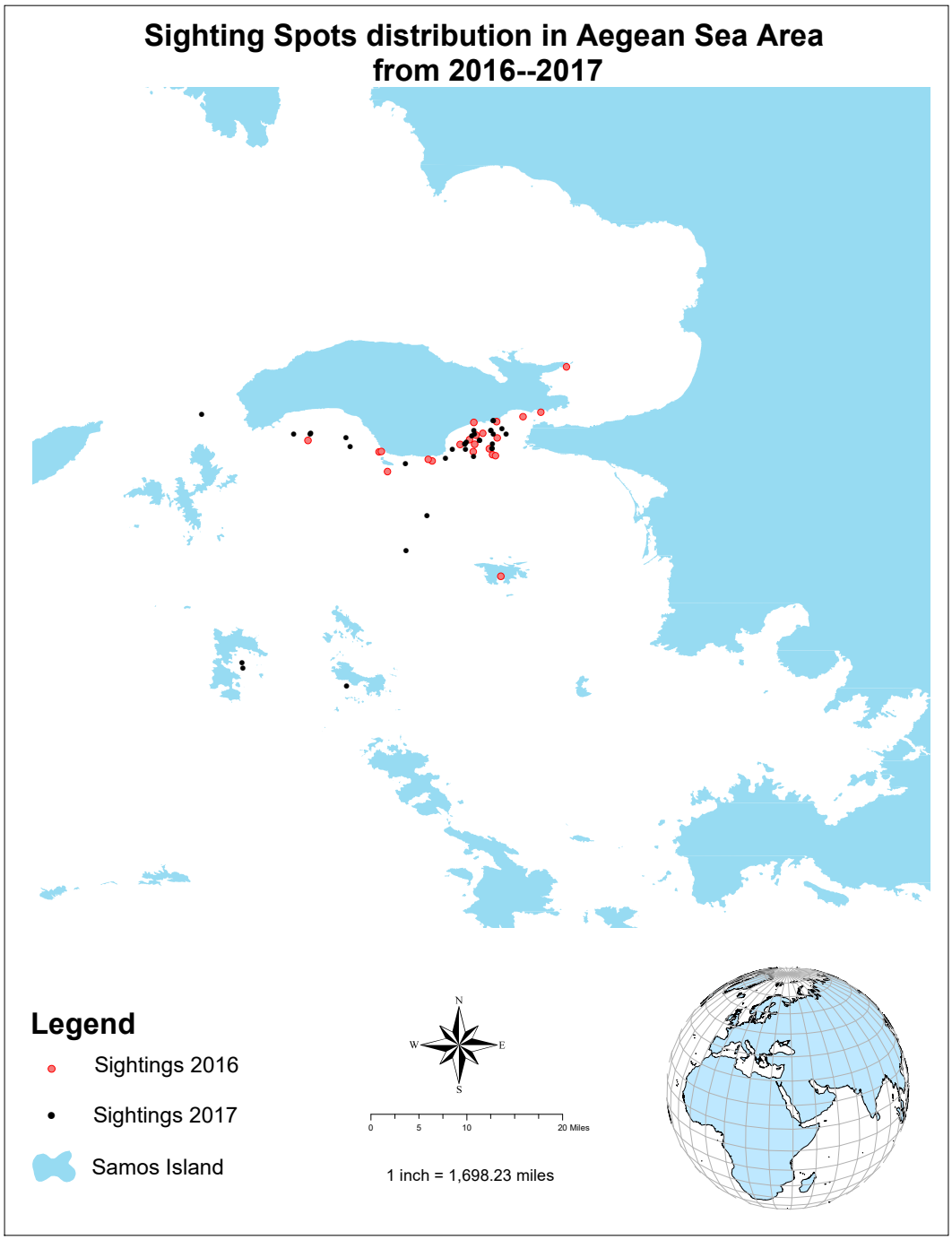


Figure 4 Sighting Spots(Dolphins Appearance Data) from 2016-2017

Yuqing Xia
Archipelagos, Institution of Marine Conservation

Discussion

The aim of this project is to predict Common Dolphin Delphinus delphis' distribution around samos island in order to help our researchers record dolphins' behaviors, habits and actives more efficiently. In this chapter, I will try to explain the reasons of our model evaluation results and try to present my own thoughts to better predict dolphins and collect data for future research.

This project is formerly conducted by Kasia and Julita where they believed that Chlorophyll, Salinity, Sea Surface Height, Sea Surface Temperature, Distance from shore and Slope will influence dolphins' habitat. However, the models we built proved there is no significant relationship between them. We do found that NEAR_DIST, SAL and CHL will influence dolphin locations by key factor analysis, however I believe the inner reason is more concerned with nutrients in the sea, while salinity and chlorophyll are indirect cause of dolphins show up probability.

Future researchers should add biomass, material flow, flow rate, eddy variable to dig deeper into the dolphins' feeding locations, and thus there we have higher probability to see dolphins. Like in this project, we concluded NEAR_DIST is the most significant variables to help us record dolphins' appearance and one of the reason is that there are food left by tourist near the shore.

We should also add error index into our model in order to better predict dolphins' distributions. As we discussed before, our future model should look like below:

$$P\_A = f(CHL+SAL+SLO+NEAR\_DIST+Biomass+Flow\_rate+material\_flow+eddy)+\varepsilon$$

Error index is extremely important in our model because observing dolphins are random and of high risk of missing dolphins based on the weather, cloud cover and other factors. Thus we can add cloud cover and dolphins' general moving ranges into the models as possible errors in order to increase model accuracy.

Final suggestion is to design new route according to variables average map per year and try to cover as many variable ranges as we can to test dolphins' distribution area.

Individual Efforts

- Yuqing Xia
Data pre-processing of 2017's data, data analyzing of both 2016 and 2017's data,and model building and evaluation.

- Kasia and Julita

Yuqing Xia
Archipelagos, Institution of Marine Conservation

Data collection and geographical data manipulation of 2015 and 2016's data.