Statistical Machine Learning Final Report

Alexis Athens, ID #1

Predicting the Spatial Distribution of an Invasive Apple Snail

Abstract

With an invasive apple snail dispersing quickly throughout the southeastern US, it is important to understand the extent of its range in order to predict the distribution and success of the endangered Florida snail kite population. Machine learning methods are ideal for modeling species distributions since they are relatively easy to implement and adept at capturing complex spatial and temporal dependencies within the data. In this paper, a logistic regression predicted the relationship between environmental variables and the habitat of the apple snail most accurately. Furthermore, predictors associated with precipitation levels best explained distribution variability.

Introduction

My objectives for this project were to identify which environmental variables best predicted the spread of the invasive apple snail (*Pomacea maculata*) and to build a predictive model of its distribution. This research question is pertinent since the endangered Florida snail kite selects highly for the apple snail. Historically, only the native Florida apple snail (*Pomacea paludosa*) inhabited the snail kite range, and the distribution and survival of the snail kite population was highly dependent on the location and abundance of the Florida apple snail.

Since then, the invasive species was introduced to southeast Florida in the early 1990's from South America and has quickly spread to the southeastern United States. The snail kite population has increased and has extended further north than their historical range, roughly following the distribution of the invasive apple snail. Thus, this project aims to capture the dynamics of the invasive apple snail population so that predictions about the distribution and success of the endangered Florida snail kite can also be inferred.

Background and Data

The data for this project was pulled from the USGS Nonindigenous Aquatic Species database. There were a total of 725 observations reported by citizens which included the coordinates of the sighting. In order to yield more accurate predictions, an equal number of pseudo-absences was generated to supplement the presence-only dataset (Fig.1). Seventeen bioclimatic variables along with the longitude and latitude of the observations constituted the p=19 predictors considered in the analysis. The methods employed predicted a binary categorical response of whether the snail was present (1) or absent (0) at the given coordinates.

Methods

K-fold cross-validation

In order to account for the spatial dependence between data points, a spatial k-fold cross-validation was implemented. Rather than having the observations randomly assigned to k folds, the n=1450 observations were first sorted into approximately 100 localized blocks (Fig. 2) so that the methods were guaranteed to predict the presence of snails in a totally new geographic area. These 116 blocks were then randomly assigned to k=10 folds for the analysis.

Logistic regression

For the baseline method, a simple logistic regression with all the bioclimatic predictors was fit to the data. The accuracy of this method was measured by the area under the ROC curve (AUC) and the corresponding average misclassification rate, which was 0.871 and 11.6%, respectively. The average misclassification rate was obtained by averaging the proportion of misclassified observations across the k folds.

Forward selection

Next, forward feature selection was implemented with a logistic regression to determine which variables had the highest predictive power. This was calculated by ordering the predictors from highest to lowest explanatory power in each fold. A glm was then fit for each of the p predictors, beginning with the null model and then iteratively adding in the next best predictor. The validation error rate per number of predictors was then averaged across folds.

This method indicated that including the best 14 out of the 19 predictors yielded the most accurate model with a misclassification rate of 11.1%. The best predictor, precipitation of the warmest quarter, accounted for 48.9% of the variability in the distribution of the apple snail. The next best predictor, the mean temperature of the wettest quarter, explained an additional 5.6% of the variability. Using the best 14 predictors increased the R^2 value from 48.9% to 59.2%.

Lasso shrinkage

Next, the lasso shrinkage method was employed. This method also performs variable selection as it can force some coefficient estimates to zero. However, with the optimal lambda of 0.01, the coefficients needed little tuning. Nonetheless, the model performed decently with a misclassification rate of 12.2% and an AUC of 0.861.

GAM

Next, a slightly more flexible model was considered. Generalized additive models (GAMs) with natural splines were fit to the data. This model was chosen since the low-order polynomials fit to localized areas of the curve are adept to capturing non-linearities while still controlling for the complexity of the model. Three GAMs were considered with natural splines placed on the best 3, 5, and 8 predictors. This resulted in misclassification rates (AUC) of 12.2% (0.858), 11.94% (0.867), and 13.4% (0.847), respectively, indicating that allowing more flexibility to a limited subset of variables could improve predictive power.

Random forest

The final method considered was the random forest. A random forest is created by building a large number of decision trees off of bootstrapped training samples. Each individual tree is shallow with only a few internal splits and is decorrelated from the other trees by only utilizing $m \approx \sqrt{p}$ predictors. Averaging the results across the trees reduces the overall variance of the model. The random forest model had a misclassification rate of 12.0% and AUC of 0.865 using the default settings.

Analysis

After analyzing the ROC curves (Fig. 3) and comparing misclassification rates, it is apparent that the logistic regression offered the highest level of predictive power. Though all models resulted in relatively low error rates, less flexible models like the logistic regression and the GAM performed best. This indicates that the truth is quite linear. Plotting the strongest predictor (Fig. 4), it is visible that the presence of apple snails decreases as the amount of precipitation decreases. Thus, given that all of the predictors are bioclimatic variables, it is perhaps not too surprising that lower-complexity models best predicted the apple snail's range.

However, the results from forward feature selection were quite unexpected. It has been well documented that the apple snail is highly sensitive to cold winter temperatures and was assumed to be the limiting factor of the population distribution. However, forward selection indicated that the strongest predictors are associated more with precipitation. Namely, "precipitation of warmest quarter", "mean temperature of wettest quarter [late spring]", and "precipitation of wettest quarter" were the strongest three predictors. In fact, the 9th best predictor was the first to even imply a spatial relationship with winter temperatures. Hence, the machine learning methods discovered key environmental predictors that were previously unidentified and generated a novel distribution prediction for the apple snail (Fig. 5).

Discussion

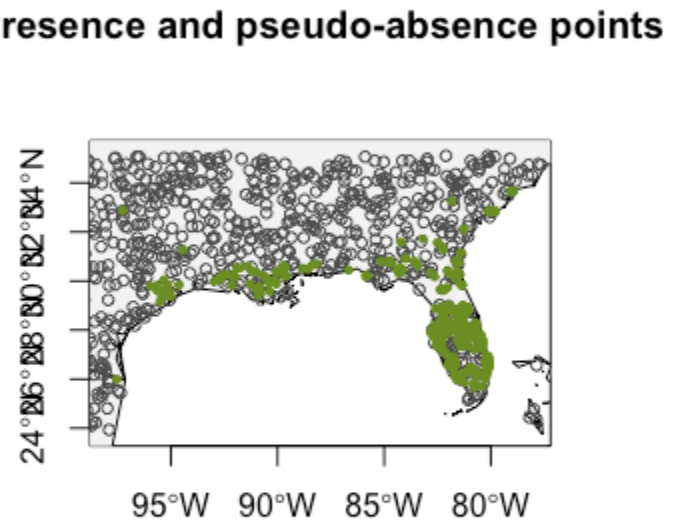
As previously noted, the collection of bioclimatic variables led to a somewhat linear relationship with apple snail presence. However, I think that including additional predictors such as water pH (a well-documented indicator), water levels, or soil composition across the southeastern US would have made for a more interesting and complex relationship. Unfortunately, I was not able

to access the raster files in time for the analysis, but I look forward to seeing how the results change with the new information.

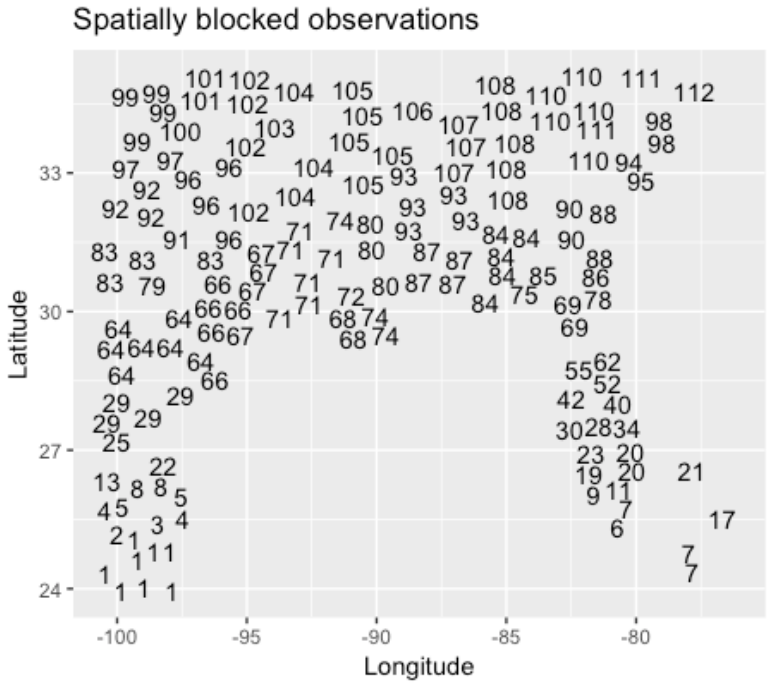
That being said, I also plan to improve this project by further refining the random forest model. Using the default settings, the random forest had an error classification rate of only 12.0% even with the linear truth. By further tuning the input parameters and adding the new raster files, I think the random forest could very well be the most accurate model. Furthermore, I could have improved model accuracy by redefining the folds and running k-fold cross-validation many times. For the sake of time and computational power, I decided to focus on learning to implement the methods by hand at the expense of prediction accuracy. In the future, I hope to be more confident in the accuracy of my results as well.

Figures

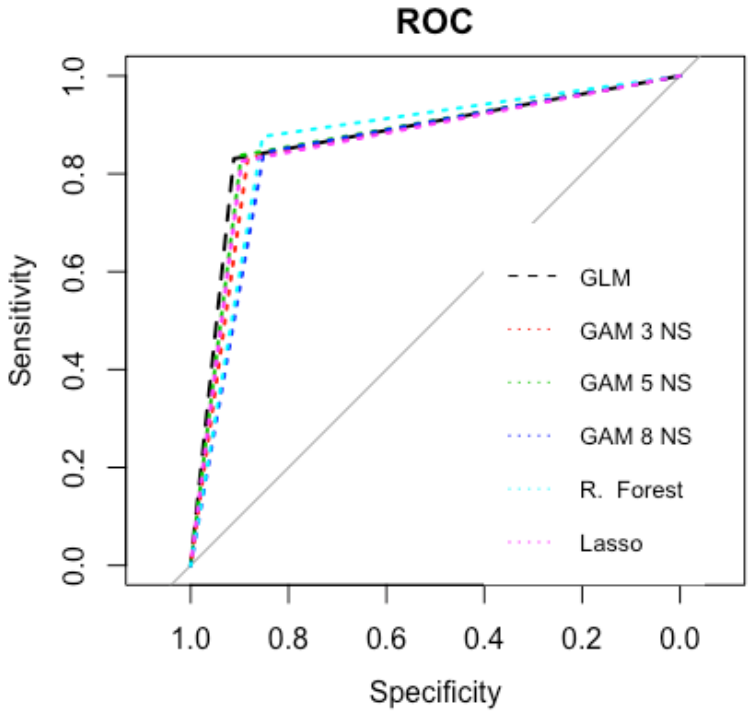
(Fig. 1) Presence and generated pseudo-absence points



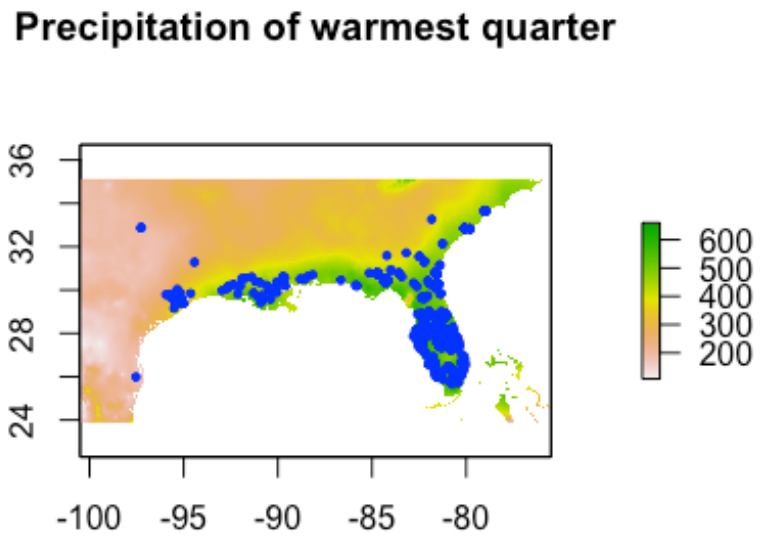
(Fig. 2) Data sorted into localized blocks in preparation for k-fold cross-validation



(Fig. 3) ROC curves



(Fig. 4) Precipitation of warmest quarter variable with observed presences



(Fig. 5) Projected range extent of invasive apple snail