Statistical Analysis of Sea Turtle Hatch and Emergence Success and Human Census Population on Coastal North Carolina for 2021

Authors
Dan Crownover, Reese Letts, Townes Ellum, Allison Fry

Abstract

This project utilized a combination of traditional statistical analysis, data wrangling techniques and choropleth mapping to produce a refined geographical analysis of sea turtle hatch and emergence success in coastal North Carolina for the year of 2021. .R studio and software was used for all analysis and geospatial comparisons. All package and library citations may be found in the original markdown document. Alternative Hypothesis: There is a correlation between sea turtle hatch and emergence success and coastal U.S census population, the null is that there is no statistical correlation between the confounding variables. Primary results concluded a statistical failure to reject the null hypothesis (p = 0.12) and a lack of correlation between sea turtle hatch success and total census population in NC for the year of 2021.

1. *Introduction and Literature Review*

For decades, sea turtle monitoring projects have taken a surge in order to gain better understanding of sea turtle behavior, ecology and conservation on a global and local scale (Hutchinson, B. et al, SWOT, 2020). Today, we have a better understanding of sea turtle populations than ever before, and monitoring is gaining traction from ecologists worldwide. For the past three decades, the North Carolina Wildlife Commission has been monitoring sea turtle populations on an extensive scale, while initially starting to measure only density, chronology and distribution, there is now an extensive data set with just over 30 measurements and varying environmental parameters (*North Carolina Wildlife Commission, 2018*).

It is well known that human populations have harmful impacts on sea turtles' overall nest success (NOAA Report, 2017). Anthropogenic coastal development, pollution and nest destruction are some of the key drivers that decrease overall sea turtle hatch and nest success (Oceana Europe, 2022). It is apt to concur that human population levels could correspond to the overall success of sea turtles.

According to the data collected by the U.S. Census Bureau and an extensive demographic analysis conducted by the Carolina Population Center, the population has been growing at high rates in the last decade with the highest growth in counties along NC's coastal areas (*UNC Population Center*. Carolina Demography, 2023). With this increased risk for nesting subpopulations, this geographic and statistical analysis aimed to identify a statistically significant correlation ($p < 0.05$) between the human census population and average hatch and emergence success for the year of 2021.

Our research question was : how does the U.S census population, clutch count, and beach protection status influence sea turtle hatch and emergence success along the North Carolina coast? This investigation could prove beneficial due to the fact that nearly all species of sea turtles are threatened or endangered. Thus, understanding what specific variables can increase success in hatching and emerging can help biologists/conservationists assist sea turtles to increase their population. Other studies have shown that in 2023, there were a record number of sea turtle hatches and emergence on the NC Coast (Graff 2024). Unfortunately, with the majority of the studies showing trends of sea turtles population decreasing, it is possible that this could lead to future extinction (Southeast Fisheries Science Center (SEFSC)). We were looking to add to the understanding of sea turtle populations by examining the variables that impact the success of hatch and emergence in sea turtles.

We designed the study to replicate our interest in observing how census population, clutch count, and beach protection status affected hatch and emergence success. Our hypotheses were that with more human population, hatch/emergence success would decrease. As for clutch count, we predicted that with increased clutch count, solely from probability there should be an increase in both hatch and emergence success. Lastly, we thought that if a beach was protected, the hatch and emergence success would increase. We also wanted to look at if the array of predictor variables impact our two response variables differently or not. We were interested to see if there were specific variables that helped only one (hatch or emergence) or if they were similar for both variables.

We had data on 24 North Carolina beaches, with 21 of them unprotected and 3 protected. The response variables used were average hatch success (percent of sea turtles that successfully break out of their shells in the nest) and average emergence success (percent of sea turtles that successfully make it out of the nest/sand). The predictor variables used were census population, average clutch count, beach protection status, and interaction effects. This data was obtained from sources that include North Carolina State Wildlife Commission, Duke University, SWOT (State of the World's Sea Turtles), and United States Census Bureau. The data was collected by volunteers under the watch of Dr. Godfrey at Duke University at multiple North Carolina beach sites in 2021. Volunteers excavate, count and calculate hatch and emergence success on site and enter data into a physical database which is later digitalized.

A brief overview of our analysis, we began running a multiple linear regression, as our data looked to be linear and the assumptions passed - with the data not being statistically significant. We then ran a cross validation with the variables to see how to make the best model. In doing so, there were some significant variables, but the overall models were not statistically significant. Since no models were significant, we investigated machine learning models to best predict the model best fit. We did a few different modeling techniques (simple decision tree, bagged decision tree, random forest, boosted decision tree, and multiple linear regression), and we checked the training/ testing for both variables. We determined the best modeling technique for both in-sample (training) and out-of-sample (testing data), but due to time constraints, we did no further investigation into these techniques.

2. *Methods*

The process commenced with the importation of the biological dataset into R Studio. A sequence of code was then employed to isolate the specific columns and remove N/A values required for analysis, namely hatch success, latitude and longitude, clutch count, and emergence success.

An API key was utilized to extract U.S census data from the U.S Census website. Another line of code was then used to isolate the necessary columns: estimate, latitude, longitude, and geometry. Using the geometry column from the U.S census data frame, which contains the individual census block lat and longitude, with the lat long data for each nest site in the biological data frame, an st_intersect operation was run between the data sets. Following the assignment of this newly formed dataframe, which contained the intersected data from both biological and census datasets, all isolated data from the biological dataframe were now assigned to a specific census tract geometry, where the nest location could be plotted into these specific census tracts (*see html for visualization*).

With many nest locations occurring on the same beach, but within different census tracts, we wanted to aggregate the data into a new data frame to observe the average hatch success of turtles by beach (which was assigned to one or more tracts), then compare this average success to the aggregated census population estimates from the tracts that lie within these beaches. To do this, the data was grouped by beach and piped through a mean operator to determine average hatch success and average census estimate for those tracts.This information was assigned to a new dataframe, which provided the average nest success per beach for all listed beaches in North Carolina.

Further data wrangling involved the creation of a new column in the final dataset, the omission of NA values and code to isolate specific columns for analysis. A new column was created which compared unprotected areas vs. protected areas. Beach protection status was determined based on the "beach status" being either a national seashore/state-protected area (SP or NS), or an unprotected area. This was achieved using an if-else statement, where protected beaches were assigned the character “P” and unprotected beaches the character “U” .

In order to build a more accurate linear model, a traditional cross validation was run on with the data. A 10 repeat and 3-fold was chosen and performed using a fully saturated model for average hatch success with the three predictor variables of interest and all of the possible combinations of interactions between the three predictor variables with the following formula:

[Average.Hatch.Success ~ Census_Pop + BeachProtection_status + Average.Clutch.Count + Census_Pop*BeachProtection_status + Census_Pop*Average.Clutch.Count + BeachProtection_status*Average.Clutch.Count]

A traditional linear model was run using these variables. All assumptions were tested using the gvlma function in R. This is a powerful tool that evaluates the assumptions of linear regression models by assessing four key assumptions: linearity, homoscedasticity (constant variance of errors), normality of residuals, and independence of errors. A traditional shapiro-wilks test was also run on the data to ensure normality of residuals.

The addition of 3 added variable plots were created to accurately assess the relationship between a predictor variable and the response variable while controlling for the effects of other predictors in the model. These plots help in understanding the individual contribution of each predictor variable to the overall relationship with the response variable by plotting the residuals from a regression model of the predictor of interest against the residuals from a model containing all other predictors.

Exploratory analysis included both a natural breaks choropleth map (*figure 1)* and summary statistics (*figure 2*). For purposes of this paper four statistical plots were generated using the ggplot package. Two box plots delineating average hatch/emergence success versus beach protection status (*figures 3 and 4*) were included in order to investigate the general relationships of interest before statistical analysis.

For a genuine interest in how machine learning techniques could play a role in delineating best predictor variables, we constructed five traditional machine learning models 1) simple decision tree, 2) bagged decision tree 3) random forest 4) boosted decision tree and 5) multiple linear regression. We then compared the training RMSE and the testing RMSE for each of these models to determine the lowest RMSE.

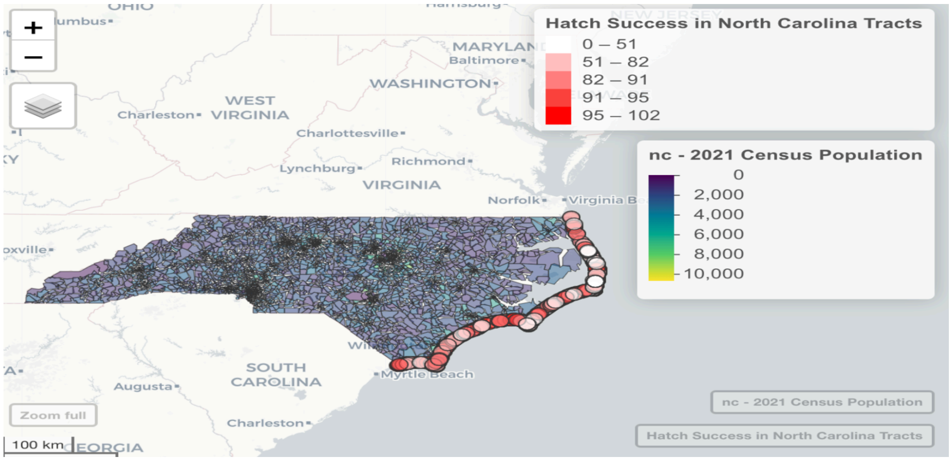
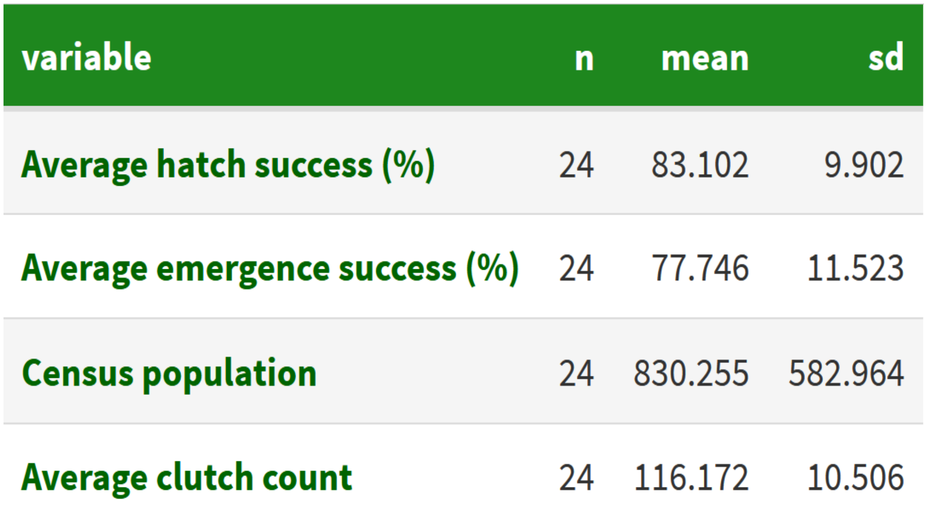


Figure 1. *Above shows a natural breaks choropleth map of the U.S census population and nest site with red being higher percentage avg. hatch success and white being low hatch success*



variable	n	mean	sd
Average hatch success (%)	24	83.102	9.902
Average emergence success (%)	24	77.746	11.523
Census population	24	830.255	582.964
Average clutch count	24	116.172	10.506

Figure 2. *Above shows summary statistics for average hatch/emergence success, clutch count, and census population across all beaches.*

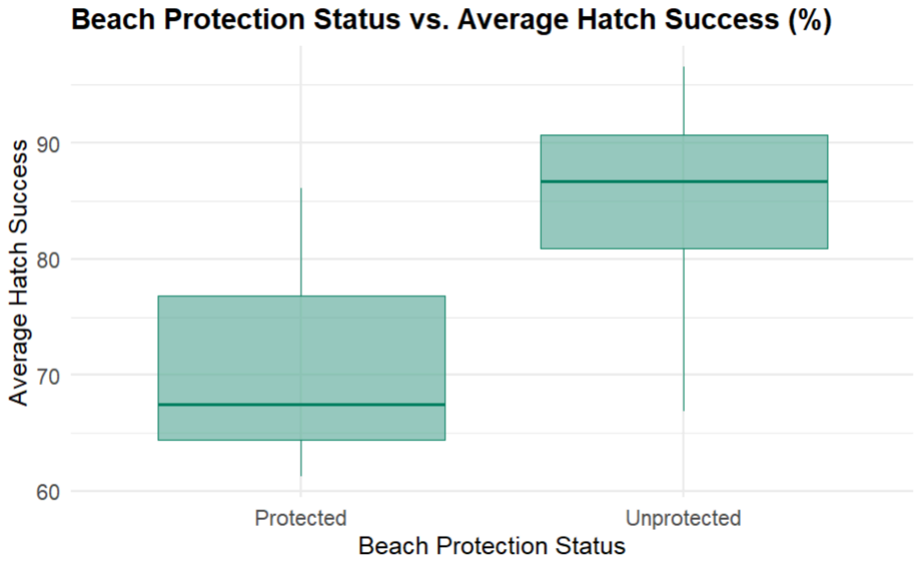


Figure 3*: Boxplots that demonstrate the average hatch success for protected beaches and unprotected beaches*

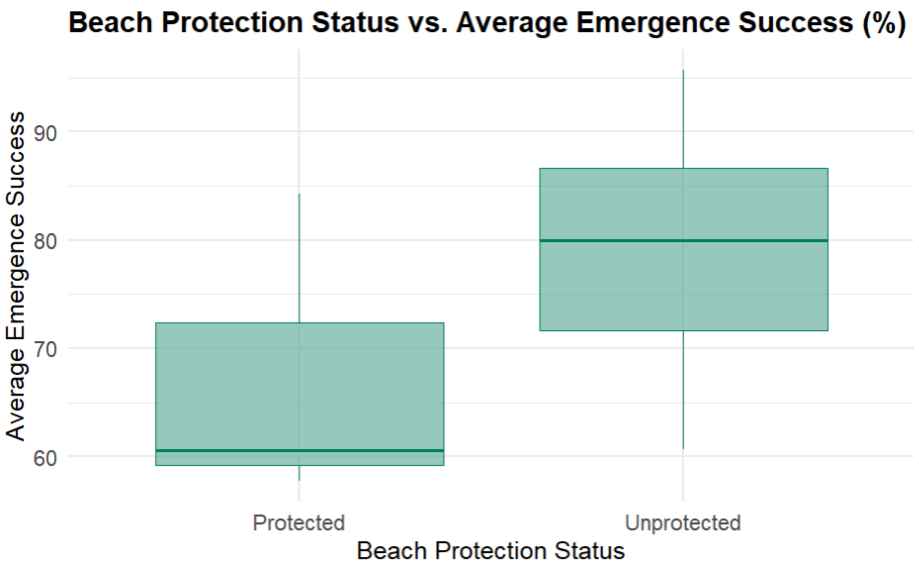


Figure 4: *Boxplots that demonstrate the average emergence success for protected beaches and unprotected beaches*

3. *Results*

Initially, a multiple linear regression was used to model average hatch success (%) using the three predictor variables of interest: census population, beach protection status, and average clutch count. At the 95% confidence level, the multiple linear regression model was not determined to be statistically significant (Adjusted $R^2 = 0.15$, F = 2.37 (20), p = 0.10). It was concluded that neither census population (p = 0.33) nor average clutch count (p = 0.44) were statistically significant predictor variables for average hatch success. It was determined that at the 90% confidence level, beach protection status could be considered a statistically significant predictor variable for average hatch success with a p-value of 0.08. Alternatively, a multiple linear regression was used to model average emergence success (%) using the same three predictor variables: census population, beach protection status, and average clutch count. Similarly to the model used for average hatch success, it was concluded that the multiple linear regression model was not statistically significant at the 95% confidence level (Adjusted $R^2 = 0.00$, F = 0.98 (20), p = 0.42). The model for average emergence success provided even less information than the model for average hatch success, as none of the predictor variables: census population (p = 0.84), beach protection status (p = 0.13), or average clutch count (0.65) were determined to be statistically significant. Although there was not a lot of information about the influence of the given predictor variables on average hatch/emergence success derived from the initial multiple linear regression models, the statistical analysis was expanded for the use of repeated, k-fold cross validation to source alternative models that may provide more insight on the dataset of interest.

The assumptions for the multiple linear regression of average hatch success vs. census population, beach protection status, and average clutch count were validated using the gvlma function. The global stat parameter was used to verify that the relationship between the average hatch success and the predictor variables was linear (p = 0.90). The skewness (p = 0.58) and kurtosis (p = 0.41) parameters demonstrate that the residuals of the model correspond to a normal distribution. The link function parameter was used to validate that average hatch success is a continuous variable (p = 0.84). Finally, the heteroskedasticity parameter verified that the error variance is equally random, demonstrating that the model is homoskedastic (p = 0.93). Given that none of the p-values were concerningly low, false acceptance of the assumptions for the multiple linear regression model for average hatch success was not a concern. In order to assess the multicollinearity assumption, the Variance Inflation Factor (VIF) for each of the independent variables was measured. The VIFs for census population, beach protection status, and average clutch count were reported to be 1.101, 1.103, and 1.007, respectively; and given that none of the VIFs exceeded 10, it was concluded that there is no violation multicollinearity assumption within this model. The same methods were used to test the assumptions for the multiple linear regression of average emergence success vs. census population, beach protection status, and average clutch count. The gvlma parameters: global stat (p = 0.60), skewness (p = 0.85), kurtosis (p = 0.31), link function (p = 0.21), and heteroskedasticity (p = 0.79), all validated that the required assumptions were accepted for this model. In addition, the VIFs for census population, beach protection status, and average clutch count were measured to be the same values as were reported for the multiple linear regression model for average hatch success; and given that none of the VIFs exceeded 10, it was concluded that there is no violation of the multicollinearity assumption.

Added-Variable Plots for both the multiple linear regression models of average hatch success and average emergence success are provided in Figures 5 and 6, respectively. Added-Variable Plots provide a visual demonstration of the association between the response variable and a given predictor variable, while holding the remaining predictor variables in the model constant. For average hatch success (Figure 5), it was determined that for every one unit increase in census population, average hatch success increases by 0.34% while holding beach protection status and average clutch count constant. (p = 0.32, df = 20). For every beach that is protected, average hatch success decreases by 1,102% while holding census population and average clutch count constant (p = 0.08, df = 20). Finally, for every one unit increase in average clutch count, average hatch success increases by 14.22% while holding census population and average clutch count constant (p = 0.78, df = 20). For average hatch success, neither census population nor average clutch count were determined to be statistically significant within the model, but beach protection status was interpreted to be statistically significant at the 90% confidence level. For average emergence success (Figure 6), it was determined that for every one unit increase in census population, average emergence success decreases by 0.089% while holding beach protection status and average clutch count constant. (p = 0.84, df = 20). For every beach that is protected, average emergence success decreases by 1,193% while holding census population and average clutch count constant (p = 0.13, df = 20). Finally, for every one unit increase in average clutch count, average emergence success decreases by 10.44% while holding census population and average clutch count constant (p = 0.65, df = 20). For average emergence success, none of the predictor variables were determined to be statistically significant within the model.

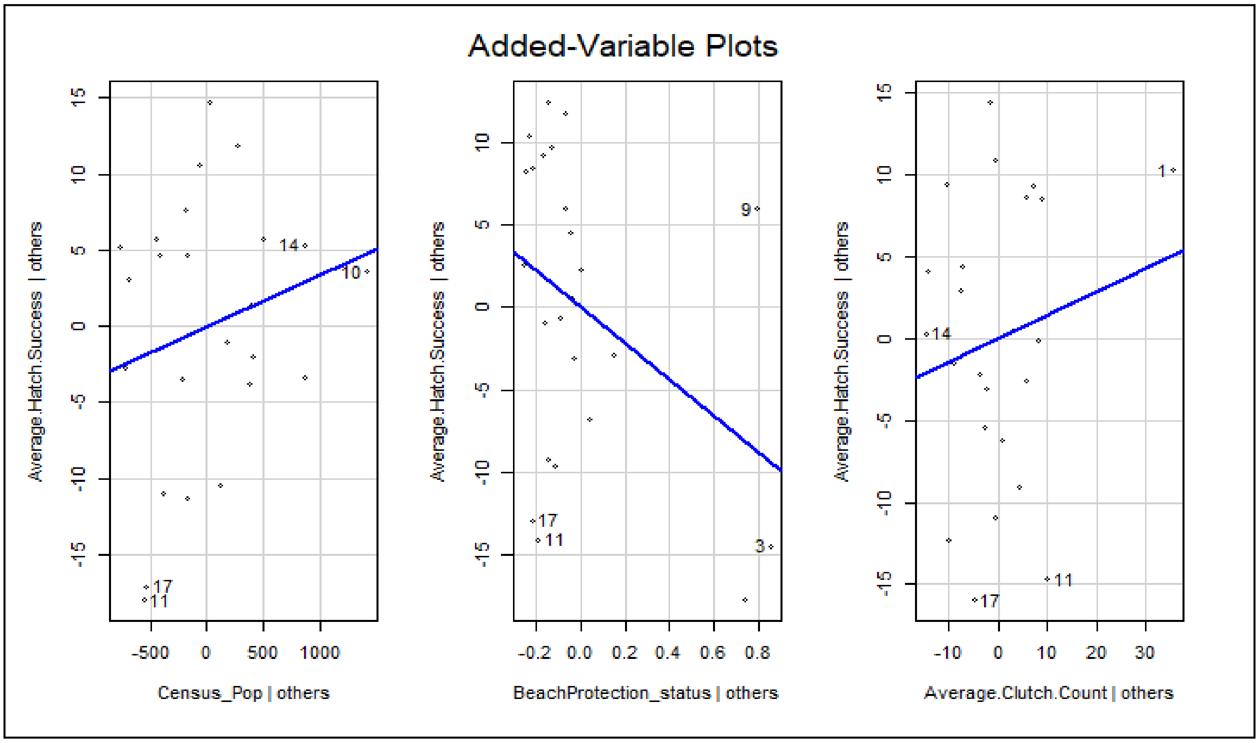


Figure 5: *Added Variable Plots from the Multiple Linear Regression Model of Average Hatch Success vs. Census Population, Beach Protection Status, and Average Clutch Count*

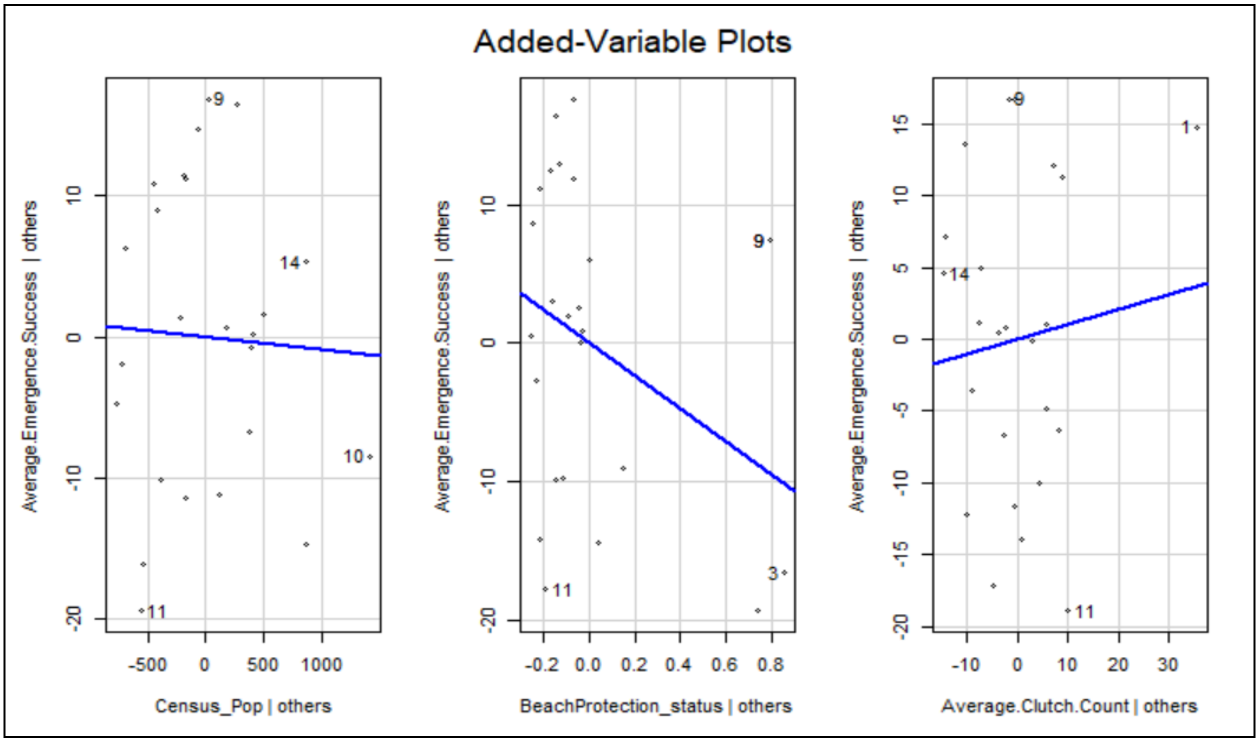


Figure 6*: Added Variable Plots from the Multiple Linear Regression Model of Average Emergence Success vs. Census Population, Beach Protection Status, and Average Clutch Count*

From the 3-fold repeated cross validation, the model with the lowest RMSE value, i.e. the model with the lowest error rate associated with predication of average hatch success, was determined to be a simple linear regression between average hatch success and census population (RMSE = 9.53). Alternatively, the model with the lowest R^2 value, i.e. the model that provides the most information about variability within the dataset, was determined to be Average.Hatch.Success ~ BeachProtection_status + Average.Clutch.Count + Census_Pop*BeachProtection_status + BeachProtection_status*Average.Clutch.Count (R^2 = 0.28).

The same cross validation procedure was executed as described above with the substitution of average emergence success in place of average hatch success, and similar results were observed. The model with the lowest RMSE value for the prediction of average emergence success was determined to be a simple linear regression between average emergence success and census population (RMSE =12.22). The model with the lowest R^2 value was determined to be Average.Emergence.Success ~ Census_Pop + BeachProtection_status + Average.Clutch.Count + Census_Pop*BeachProtection_status + BeachProtection_status*Average.Clutch.Count (R^2 = 0.29). Although there wasn't an outstanding difference between the model with the lowest R^2 value for average hatch success and that for average emergence success, it was noted that the model with the lowest R^2 value for average hatch success only contained two of the three predictor variables and two of the three possible interactions, while the model with the lowest R^2 value for average emergence success contained all three of the independent, predictor variables and the same two interactions included in the model for average hatch success. Thus, it was perceived that census population as an independent variable may play a role in the variation observed for average emergence success within the dataset.

The simple linear regression model for average hatch success vs. census population concluded that the slope coefficient for the model is 0.00550 with a standard deviation of 0.00343. The y-intercept coefficient was determined to be 78.5 with a standard deviation of 3.45. From the linear regression, it can be concluded that average hatch success does not have a statistically significant relationship with census population at the 95% confidence level. The residual standard error for this model was determined to be 9.58 on 22 degrees of freedom. It was concluded that 10.49% of the variation in average hatch success is explained by the linear relationship with census population. In conclusion, the simple linear regression model predicts that every additional unit of census population will lead to a 0.00550 increase in average hatch success. Given that the p-value of the model is 0.12, it was concluded that the simple linear regression model for average hatch success vs. census population is not statistically significant. The results for this simple linear regression model are visualized in Figure 7.

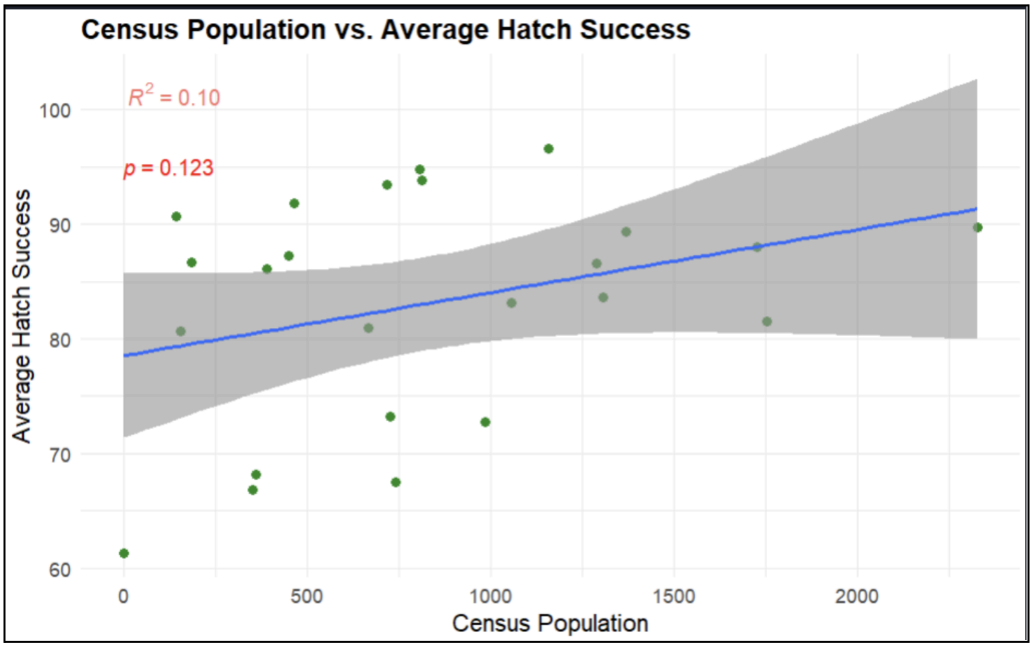


Figure 7: Visualization of the Simple Linear Regression between Average Hatch Success and Census Population

The simple linear regression model for average emergence success vs. census population concluded that the slope coefficient for the model is 0.00129 with a standard deviation of 0.00421. The y-intercept coefficient was determined to be 76.7 with a standard deviation of 4.24. From the linear regression, it can be concluded that average emergence success does not have a statistically significant relationship with census population at the 95% confidence level. The residual standard error for this model was determined to be 11.76 on 22 degrees of freedom. It was concluded that 0.43% of the variation in average emergence success is explained by the linear relationship with census population. In conclusion, the simple linear regression model predicts that every additional unit of census population will lead to a 0.00129 increase in average emergence success. Given that the p-value of the model is 0.76, it was concluded that the simple linear regression model for average emergence success vs. census population is not statistically significant. The results for this simple linear regression model are visualized in Figure 8.

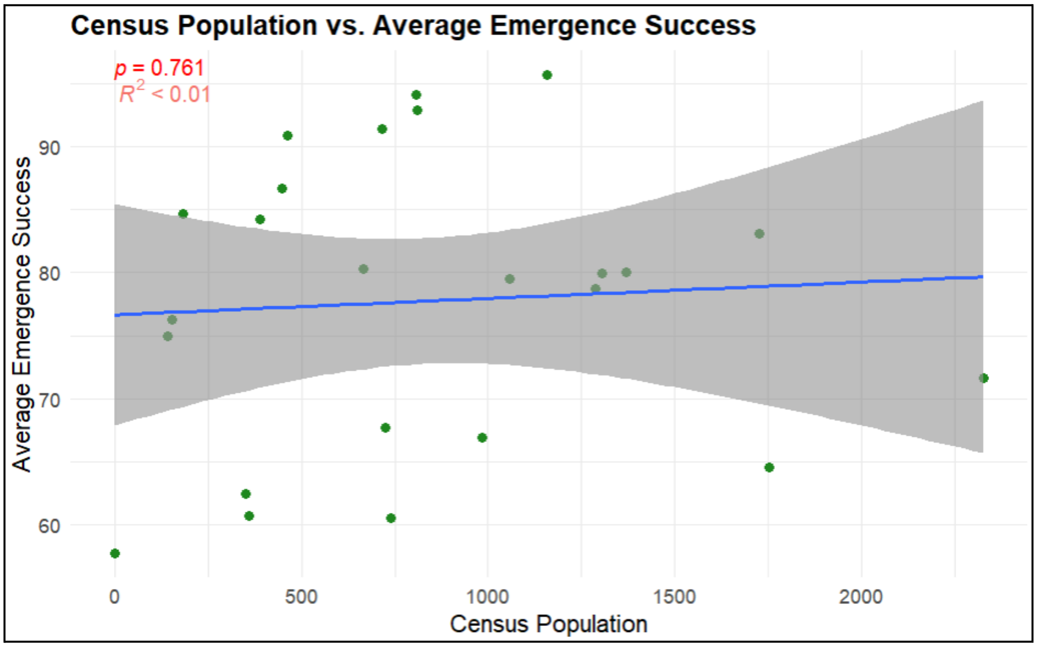


Figure 8: Visualization of the Simple Linear Regression between Average Emergence Success and Census Population

The gvlma function was used to assess the assumptions for the multiple linear regression model of Average Hatch Success vs Beach Protection Status, Average Clutch Count, Census Population*Beach Protection Status, and Beach Protection Status*Average Clutch Count. The gvlma parameters: global stat (p = 0.65), skewness (p = 0.20), kurtosis (p = 0.79), link function (p = 0.67), and heteroskedasticity (p = 0.44), all validated that the required assumptions were accepted for this model. Given that this model contains interaction effects between different combinations of the predictor variables, the VIF measurements are skewed and are not an accurate reflection of the multicollinearity associated with the independent variables of interest. Thus, the VIF measurements from the simpler multiple linear regression model for average hatch success that were reported previously can be referenced to ensure that independent variables do not violate the multicollinearity assumption. The same process was repeated to test the assumptions for the multiple linear regression model of Average Emergence Success vs Census Population, Beach Protection Status, Average Clutch Count, Census Population *Beach Protection Status, and Beach Protection Status*Average Clutch Count. The gvlma parameters: global stat (p = 0.57), skewness (p = 0.55), kurtosis (p = 0.56), link function (p = 0.14), and heteroskedasticity ($p = 0.82$), all validated that the required assumptions were accepted for this model. Again, since this model considers various interaction effects, the VIF measurements from the simpler multiple linear regression model for average emergence success that were reported previously should be referenced to verify that independent variables do not violate the multicollinearity assumption.

At the 95% confidence level, the multiple linear regression model selected via repeated 3-fold cross validation for the prediction of average hatch success was not determined to be statistically significant (Adjusted $R^2 = 0.25$, $F = 2.51$ (18), $p = 0.07$). For average hatch success, it was determined that for every beach that is protected, average hatch success increases by 108,600% while holding the other independent variables in the model constant ($p = 0.05$, $df = 18$). The influence of beach protection status on average hatch success was determined to be statistically significant at the 90% confidence level. The association of average hatch success with beach protection status contradicts the conclusion of the original multiple linear regression model that did not consider interaction effects between the independent variables. It was concluded that for every one unit increase in average clutch count, average hatch success increases by 16.41% while holding the other independent variables in the model constant ($p = 0.35$, $df = 18$). The influence of average clutch count on average hatch success was not concluded to be statistically significant. For every one unit increase in census population, it was concluded that average hatch success increases by 0.31% while holding the other independent variables in the model constant ($p = 0.36$, $df = 18$). The association between census population and average hatch success was not concluded to be statistically significant. The interaction between beach protection status and census population was concluded to have a negative association of 7.66% with average hatch success ($p = 0.10$, $df = 18$). The interaction between beach protection status and average clutch count was also concluded to have a negative association of 935.7% with average hatch success ($p = 0.05$, $df = 18$). Both of the interaction terms were determined to be statistically significant at the 90% confidence level, which indicates that the interaction terms are providing significant additional explanatory power over the multiple linear regression model that did not consider interaction effects. The added-variable plots that visually represent the statistical results of this model are provided in Figure 9.

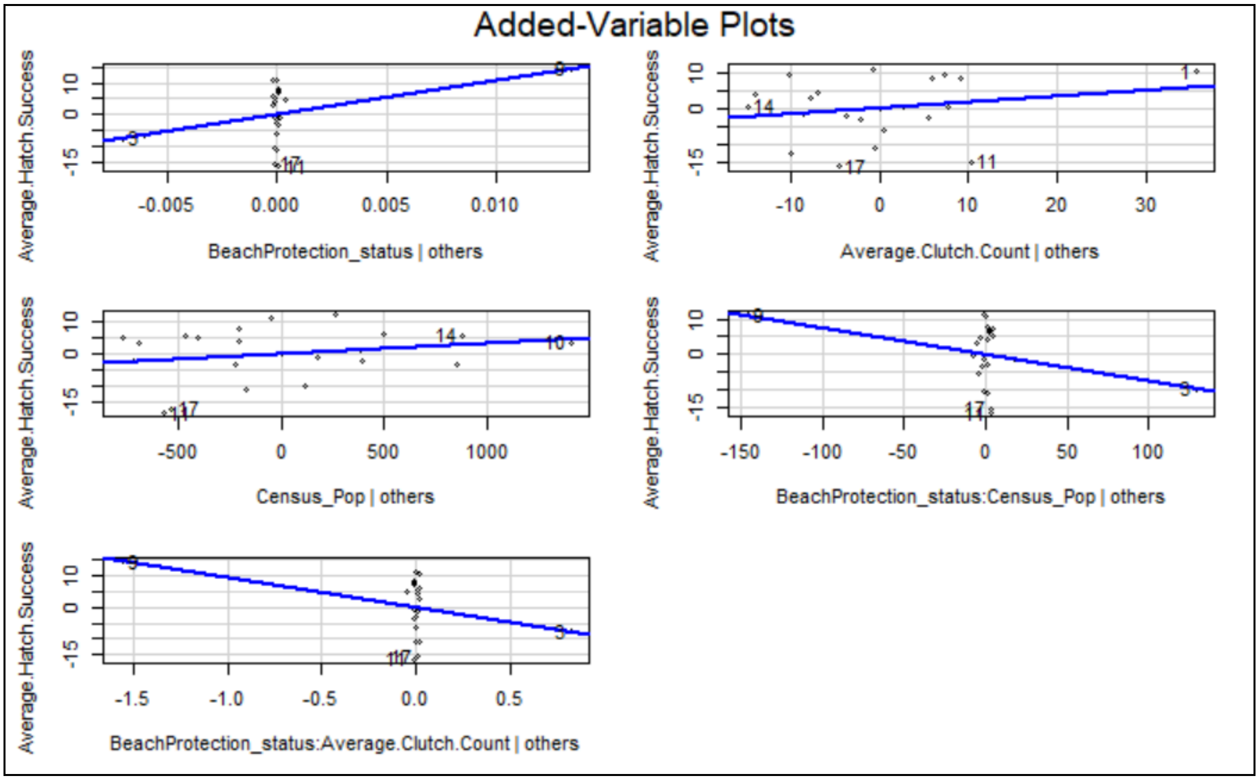


*Figure 9: Added Variable Plots from the Multiple Linear Regression Model of Average Hatch Success vs. Beach Protection Status, Average Clutch Count, Census Population, Beach Protection Status * Census Population, and Beach Protection Status * Average Clutch Count.*

At the 95% confidence level, the multiple linear regression model selected via repeated 3-fold cross validation for the prediction of average emergence success was not determined to be statistically significant (Adjusted R^2 = 0.07, F = 1.35 (18), p = 0.29). For average emergence success, it was concluded that for every one unit increase in census population, average emergence success decreases by -0.12% while holding the other independent variables in the model constant (p = 0.78, df = 18). The association between census population and average emergence success was not concluded to be statistically significant. It was determined that for every beach that is protected, average emergence success increases by 125,200% while holding the other independent variables in the model constant (p = 0.08, df = 18). The influence of beach protection status on average emergence success was determined to be statistically significant at the 90% confidence level. It was concluded that for every one unit increase in average clutch count, average emergence success increases by 12.79% while holding the other independent variables in the model constant (p = 0.57, df = 18). The influence of average clutch count on average emergence success was not concluded to be statistically significant. The interaction between beach protection status and census population was concluded to have a negative association of 8.98% with average emergence success (p = 0.13, df = 18). This interaction term was not concluded to be statistically significant, therefore it does not provide any additional significant explanatory power for this model. The interaction between beach protection status and average clutch count was also concluded to have a negative association of 1078% with average emergence success (p = 0.08, df = 18). In contradiction to the previous interaction term, the interaction between beach protection status and average clutch count was determined to be statistically significant at the 90% confidence level, which indicates this interaction provides significant additional explanatory power over the multiple linear regression model that did not consider interaction effects. The added-variable plots that visually represent the statistical results of this model are provided in Figure 10.

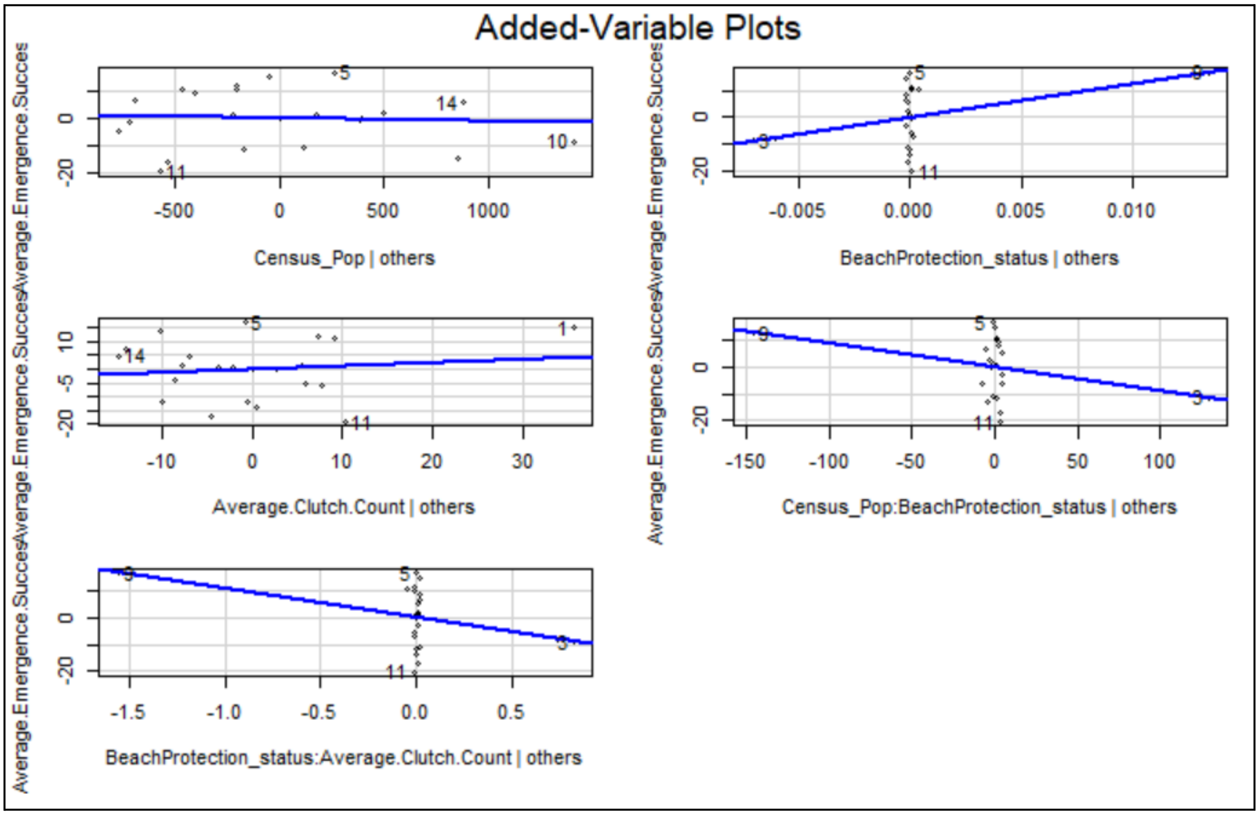


*Figure 10: Added Variable Plots from the Multiple Linear Regression Model of Average Emergence Success vs. Census Population, Beach Protection Status, Average Clutch Count, Census Population * Beach Protection Status, and Beach Protection Status * Average Clutch Count.*

The results for investigating which machine learning model would provide the best results for predicting average hatch success and average emergence success are summarized in Table 1 and Table 2, respectively. For the prediction of average hatch success, it was concluded that the boosted decision tree model (RMSE = 0.000980) performed the best during the in-sample period in which the training data is used for initial parameter estimation and model selection. Alternatively, the multiple linear regression model (RMSE = 7.64) performed the best during the out-of-sample period when the testing data is used to evaluate forecasting performance. Similar results were obtained when investigating machine learning methods for prediction of average emergence success. It was concluded that the boosted decision tree model (RMSE = 0.000895) performed the best during the in-sample period while the multiple linear regression model (RMSE = 10.1) performed the best during the out-of-sample period.

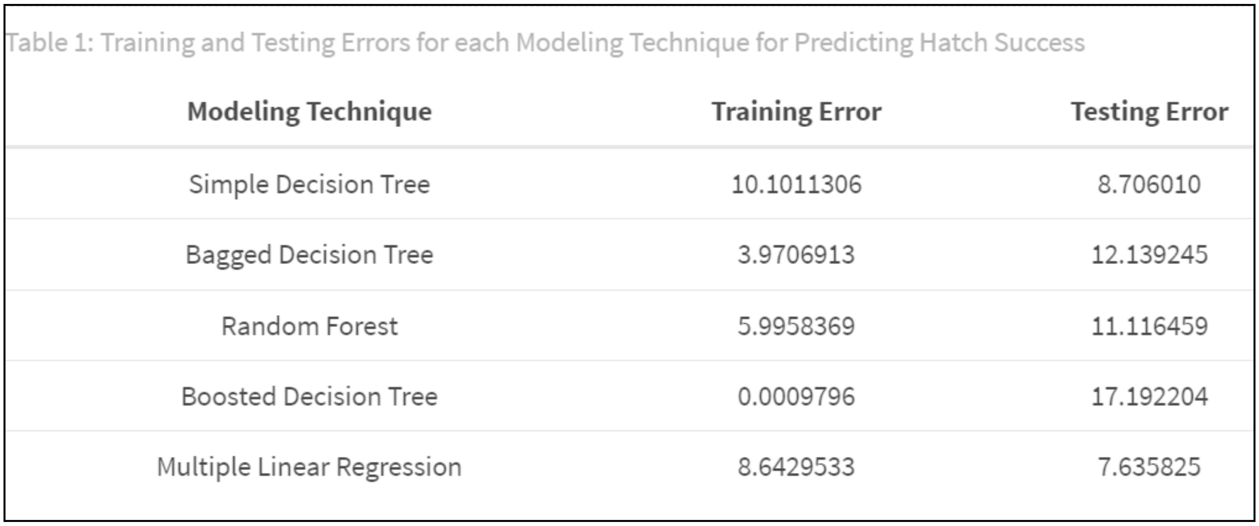


Table 1: Training and Testing Errors for each Modeling Technique for Predicting Hatch Success

Modeling Technique	Training Error	Testing Error
Simple Decision Tree	10.1011306	8.706010
Bagged Decision Tree	3.9706913	12.139245
Random Forest	5.9958369	11.116459
Boosted Decision Tree	0.0009796	17.192204
Multiple Linear Regression	8.6429533	7.635825

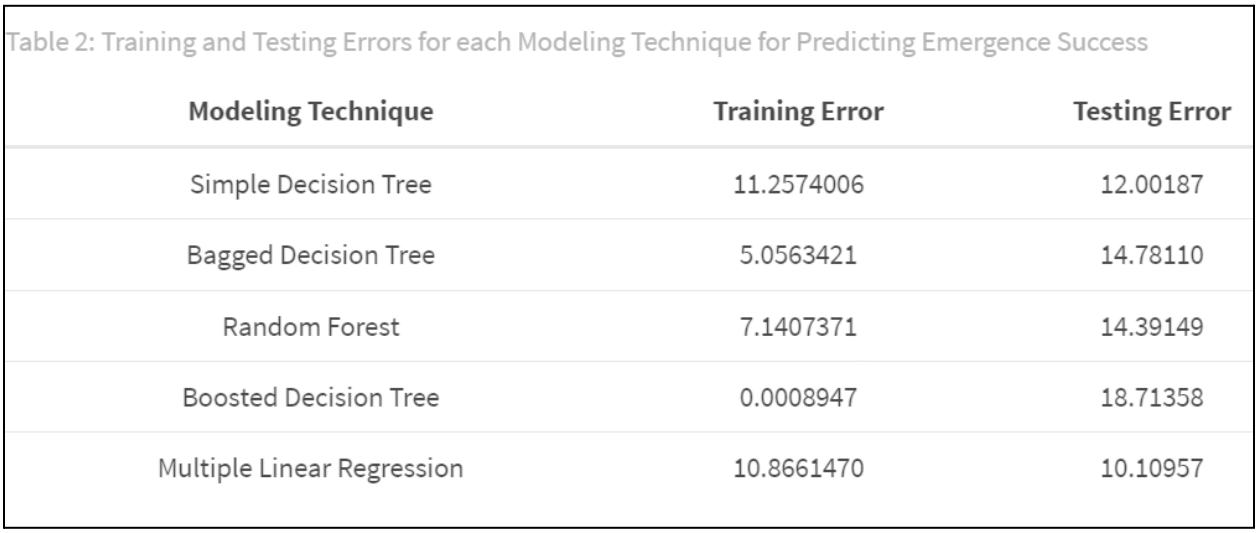


Table 2: Training and Testing Errors for each Modeling Technique for Predicting Emergence Success

Modeling Technique	Training Error	Testing Error
Simple Decision Tree	11.2574006	12.00187
Bagged Decision Tree	5.0563421	14.78110
Random Forest	7.1407371	14.39149
Boosted Decision Tree	0.0008947	18.71358
Multiple Linear Regression	10.8661470	10.10957

4. *Discussion*

Our statistical analysis comparing census tract population and sea turtle hatch/emergence success along the coast of North Carolina in 2021 determined that there is no significant linear correlation between census population and average hatch/emergence rates for this year. Therefore, the null hypothesis - that there is no significant correlation between census population and hatch/emergence success - cannot be rejected. While our 3-fold cross validation indicated that simple linear regressions between census population/hatch success and census population/emergence success have the lowest RMSE values when predicting the response variables, because these regressions are not statistically significant, we cannot draw any meaningful conclusions from these results. However, as discussed in our Results section, cross-validation did determine that the models with the lowest RMSE and R^2 values for both hatch and emergence success would include census population as a predictor variable. This suggests that census population may have some effect on the variation displayed within average hatch and emergence success in this dataset, although the direction and magnitude of this effect cannot be determined from our analysis.

Across all linear models, the only predictor variable that displayed statistical significance was beach protection status (protected vs. unprotected), with $p = 0.08$ as a predictor of hatch success. Simple box plots comparing hatch and emergence success to protection status (Figures 3 and 4) show visually that mean hatch and emergence rates are higher in unprotected beaches. However, there are limitations to the relevance of the results. Of the 24 beaches included in the dataset, only 3 are listed as protected, meaning that there may be an overrepresentation of hatch and emergence data for unprotected beaches as opposed to protected beaches and these data may exhibit kurtosis. However, despite only 3 beaches being listed as protected, nearly half (1321/2716, 48.6%) of the nest data points in our dataset were from these 3 beaches. This may have undetermined implications in the behavior of our data. Additionally, a two-way anova would need to be run in order to compare the mean success rates in protected/unprotected beaches and determine any significant difference in means. Due to time constraints, we did not do any further investigation into the effect of beach protection status on hatch/emergence success, but our results indicate that this relationship would be worth investigating in future studies.

Our investigation into machine learning yielded similar results for both hatch and emergence success. It was concluded that for predicting both hatch and emergence success, the boosted decision tree model performed the best on the in-sample training data. For the out-of-sample testing data, the multiple linear regression model performed the best for predicting both hatch and emergence success. Due to time constraints, we did not investigate these methods any further, but it would be useful to do so in further studies.

The scope of inference of our statistical analysis is broad in some regards, and limited in others. Geospatially, the results of our analysis can be broadly applied to sea turtle nesting sites along the entire coast of North Carolina. However, our analysis is very temporally limited. Because our tests and analyses were only run on data collected from a single year (2021), the results can not be used to make general statements about the effect of our predictor variables on hatch and emergence success. We have no way of knowing if 2021 was a "typical" year for nesting sea turtles, and therefore cannot draw any long-term conclusions about nest success. In order to improve the temporal scope of our analysis, hatch/emergence success data for prior years would need to be included, and the means between years compared.

Our results differed from our expectations in several ways. We expected to see higher average hatch and emergence rates in nests on beaches with lower census populations, as we assumed these nests would be in close proximity to more humans and therefore more affected by anthropogenic threats such as pollution, disturbance and nest destruction. However, our linear models comparing hatch/emergence success and census population (Figures 7 and 8) indicated the opposite was true; that there actually appears to be higher hatch success in beaches with higher census populations. However, this model has a non-significant p-value, so this trend does not tell us much. We did, however, come up with a possible explanation for this apparent, unforeseen trend. The vast majority of people who interact with North Carolina's coastal areas are not residents, but tourists/visitors. It was observed in 2022 that coastal North Carolina was ranked #6 in the U.S. in terms of domestic visitation for tourism, with approximately 11.2 million visitors traveling to the coastal Region, and 80% of these being overnight visitors (*North Carolina Regional Visitor Profile and U.S Census, 2022*). These visitors would not appear in census tract data for the region, but likely affect sea turtle nesting habitat, directly and indirectly, much more than permanent residents. Therefore, census data may not be as relevant as tourist data in predicting nest success. Another unexpected result we found was that mean hatch/emergence success appeared to be higher in unprotected rather than protected beaches. Again, this result is not statistically significant, and needs more tests in order to attach real importance to it. But it is an interesting phenomenon that would be worth investigating further. One possible explanation for this is that natural predator counts may be higher in protected beaches, leading to higher nest predation rates. However, human-associated predators (such as dogs) are likely more abundant in unprotected areas, so these effects may cancel out. Given the insignificance of our results, it's difficult to formulate any concrete hypotheses explaining the trends we observed.

Our analysis is limited mainly in that we only had access to nesting data from one year (2021). In order to run a more stout analysis, we would need to obtain data from multiple years, and compare trends across years. We also had a small number of predictor variables. Our analysis could benefit from the inclusion of additional predictor variables such as transient tourist fluctuations, annual utility power trends for households and hotels (as a proxy for light pollution), economic trends, annual migration patterns, and climate data. Additionally, given the time, we would likely conduct a more in-depth geospatial analysis, and examine more nuanced geospatial trends in nesting success.

5. *Conclusion*

With the coast of North Carolina experiencing significant population growth in recent years, as well as major increases in tourist visitation, anthropogenic threats to sea turtle nesting grounds in this region are an issue of ongoing and increasing concern for wildlife ecologists and sea turtle conservationists. Nestled at the crossroads of anthropogenic demand, domestic use, and marine wildlife populations, this area stands as a pivotal resource for the posterity of threatened sea turtle populations on a global scale. Going forward, it is crucial that we gain a better understanding of the factors affecting sea turtle nest success in this region, and adjust our own behavior, on an individual and community-level scale, accordingly Although this statistical analysis provided statistically non-significant correlations between census population and hatch/emergence success, it illustrates the need for further statistical analysis and a more in-depth dataset that can more comprehensively represent the anthropogenic interactions/threats (tourist flux, energy use, migration, climate change) that affect the success of sea turtle nests annually.

Citations

Graff, Frank. “A Record Number of Green Sea Turtle Nests in NC This Year.” *PBS North Carolina*, PBS North Carolina, 17 Nov. 2023, www.pbsnc.org/blogs/science/a-record-number-of-green-sea-turtle-nests-in-nc-this-year/#:~:text=The%20North%20Carolina%20Wildlife%20Resources,leatherback%20turtles%20(53%20each)

“Sea Turtle Population Study in the Coastal Waters of North Carolina from 1988-06-07 to 2015-09-22 (NCEI Accession 0162846).” *NOAA Fisheries*, 26 May 2017, www.fisheries.noaa.gov/inport/item/26466\