An ecologically based statistical probe into the longitudinal status of reef fish of Grand Cayman Island

Adam Bruce

Lawrence University

Research Paper
(This is not part of the actual paper. For instructor grading use only)

Abstract:

Fish are an integral part of coral reef communities, and consequently they serve as key indicators of overall reef health. So, maintaining an abundant community of fish is critically important to reef survival. In this report, data collected from 1998 to 2018 on Grand Cayman Island were analyzed to assess family-based and species-specific variations in reef fish abundance. Altogether, this assessment used hypothesis testing, model building, and machine learning techniques to gain interpretable ecological insights. Ultimately, decreases in familial abundance over time across all Grand Cayman reefs was observed. Additionally, evidence of increased invasive Lionfish predator abundance and decreased critical herbivorous species abundance was found.

Introduction

Ecological Significance:

Coral reefs support more species per unit area than any other marine environment in the world (Ogden and Lobel, 1978). Therefore, they are fundamental in maintaining healthy marine ecosystems. In addition, they are economically critical for many communities, most often through commercial fishing, which makes their long-term health a priority for humans. Since fish are an integral part of these communities, their abundance serves as a major indicator of overall reef health. Consequently, analyzing their trends over time is focal to many biological studies. In this study, the key abundance measurement of density was studied temporally in relation to potential impactors such as fish families and diets or whether a reef was inside a marine reserve.

Within the extensive trophic network of coral reefs, particularly important families of reef fishes stand out as temporal indicators of reef health. Collectively, these indicator families include predatory and herbivorous fish. At the top of the trophic networks, predator fish have been shown to maintain biodiversity and ecosystem functioning (Armsworth et al., 2007). However, their densities have greatly decreased in the last decade, with some studies citing decreases up to an order of magnitude, because of fishing activities which target high-level predators (Pauly et al., 1998; Jackson et al., 2001).

Additionally, human activities have led to declines in herbivorous family populations (Pattengill-Semmens and Semmens, 2003). Because herbivores are the primary grazers of algal growth, these declines have often resulted in severe reef “phase shifts”. This term describes the process of a reef moving from coral dominated to macroalgal dominated, through direct competition, because of a lack of algal grazing (Williams and Polunin, 2001; Williamson et al., 2014). In addition to the loss of both predators and herbivores on reefs, understanding the dieting preferences of reef fish can also serve as a density indicator.

When observing patterns of dieting, the primary ecological concern is dietary expansion, which is a decline in dietary preferences of species. Often, these declines occur either because of major disturbances, such as those brought on by climate change, or because of density-dependent resource expansion (Lobato et al., 2014; Semmler et al., 2022). In both cases, preferences for specific food resources weakens because consumers cannot afford to pass on low quality foods, such as sponges and coral, when access to their preferred food source becomes scarce. As a result, the two species’ niches overlap, and their chances of survival decrease.

Ultimately, niche overlap causes competitive exclusion, which is the loss of a species from a habitat due to direct competition with another species for a shared resource (Chesson, 2000). Eventually, the loss of these species has cascading effects on the entire reef food-web, which causes a decrease in density of many fish populations. Therefore, higher densities of low-quality food consumers indicate lower overall population densities across all reef fish. However, one factor has been shown, at times, to offset density effects of herbivore and carnivore loss in addition to dieting expansions, and that is the implementation of marine reserves.

Globally, a considerable amount of funding has been spent on the establishment of Marine Protected Areas (MPA’s) to offset the impacts of human disturbances on coral reef communities. Overall, when designed and maintained properly, these sites have been shown to greatly increase densities of reef fish by as much as 3.7 times compared to non-protected sites

(Mosquera et al., 2000). However, intense debate has existed regarding whether these reserves are truly efficient, and many studies have found little to no impact of sites which are deemed inefficient (Edgar et al., 2011; Mora et al., 2006). Collectively, five factors have been used to designate an efficient reserve, which include (1) degree of fishing permitted within an MPA; (2) level of enforcement of fishing limitations; (3) MPA age; (4) MPA size; and (5) presence of continuous habitat for unconstrained movement of fish across an MPA (Edgar et al., 2014). By utilizing these five factors, efficiency of marine reserves on Grand Cayman could be examined, which provides the opportunity to identify their ecological impact on fish densities.

Collectively, previous studies relied on one common factor to examine the impacts of reef fish families, diets, and marine reserves temporally on fish densities. This was the implementation of statistical techniques to draw meaningful conclusions. Because this study aimed to draw similar conclusions for changes in reef fish densities on Grand Cayman Island, these techniques were the fundamental component of analysis. Specifically, this included hypothesis testing, model building, and machine learning, which at their core utilize mathematical algorithms and equations to draw inferences from patterns in data.

Island Background & Data Collection:

The Cayman Islands are the most remote islands of the Caribbean. Together, Grand Cayman, Cayman Brac, and Little Cayman make up this three-island archipelago. Grand Cayman is the largest of the three and is located on the Cayman Ridge at the southern end of the North American Plate (Jones, 1994). This plate features several faults with high tectonic activity, which has produced a plethora of coral reef habitats surrounding the island. Collectively, these reefs provide rich marine environments for a diverse range of species, which has made Grand Cayman a prominent site for marine biology related research.

From 1998 to 2018, several marine biology studies were conducted by the Lawrence University Marine Program (LUMP), which undertook a biyearly trip to Grand Cayman Island. On each trip, students used Reef Environmental Education Foundation (REEF) roving diver techniques to collect data on reef fish across eleven sites. These sites included from southeast clockwise to the northeast: Beach Bay (BB), Smith's Cove (SC), Sunset House (SH), Sea View (SV), Casuarina Point (CP), Devil's Grotto (DG), Eden's Rock (ER), Cemetery Reef (CR), Turtle Farm (TF), Spanish Bay (SB), and Coconut Harbor (CH) (Figure 1).

Overall, the roving diver techniques implemented at each site involved using underwater paper and pencils to record the number of individual species present. Readings lasted approximately 20 to 50 minutes depending on diver air consumption and non-decompression limits. After each trip, the data were used to quantify species abundances according to REEF guidelines (Timpe, 2018). In this study, the data from LUMP was analyzed in R software by implementing the statistical techniques of hypothesis testing, model building, and machine learning to draw conclusions about changes in reef fish densities on Grand Cayman Island over time.

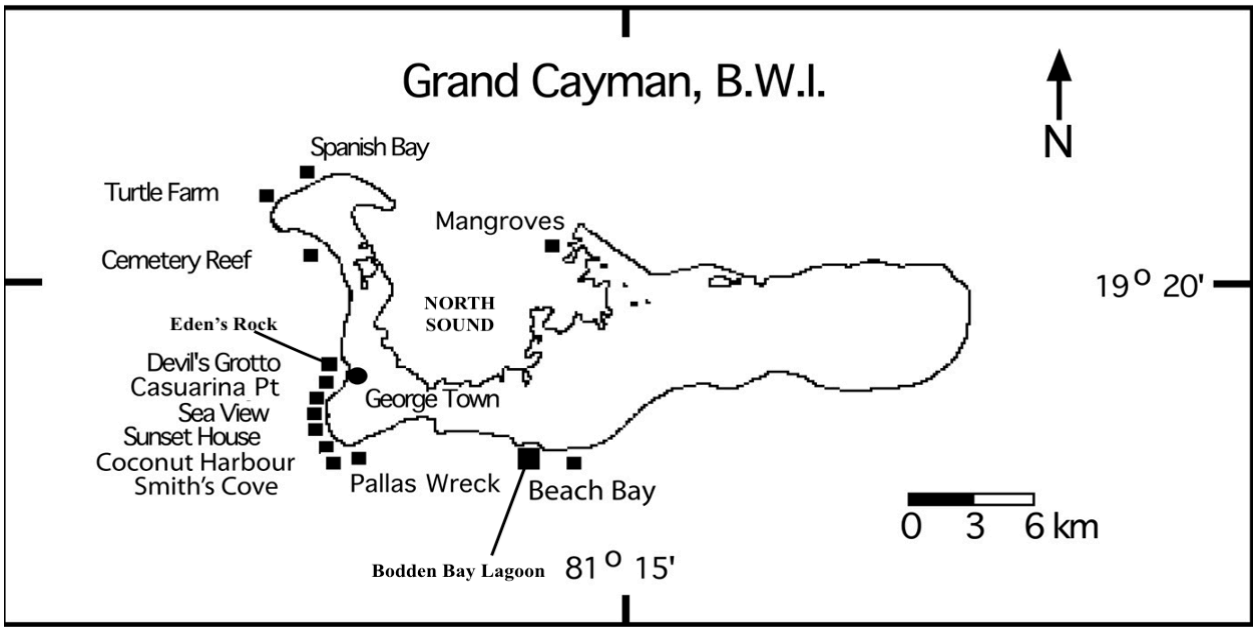


Figure 1: Visualization of Grand Cayman Island with survey sites. Both Coconut Harbour and Pallas Wreck were not used as dive sites for data collections. All other sites were surveyed.

Methods

Data Analysis:

For data analysis, a .csv file was obtained from the head of LUMP, Professor Bart De Stasio of Lawrence University, and imported into R Studio (Appendix A). REEF values collected by LUMP, which were primarily integer categories from 0 to 4, were represented by "DENSITY_INDEX" in the dataset. Overall, "0" referenced no individuals recorded, "1" was a single recorded individual, "2" meant two to ten individuals were recorded, "3" meant eleven to one hundred individuals were recorded, and "4" indicated over one hundred individuals were recorded for any given species. However, some sites were surveyed by multiple divers each year, so their density measurements were averaged, which resulted in some non-integer values. Ultimately, an analytic approach that would provide the most significant ecological insights from "DENSITY_INDEX" also had to be identified.

Numerous studies have found more significant ecological interpretations when statistical analysis was performed at an aggregate level than at the species level (Polunin and Roberts, 1993; Jennings and Polunin 1996). Therefore, analysis was primarily performed using the broader "SCIENTIFIC_FAMILY" variable in this study. Utilizing this familial-level approach, an Exploratory Data Analysis was performed to help identify patterns, relationships, and basic mathematical tendencies in the dataset (Appendix B). Collectively, this aided in identifying hypothesis testing, model building, and machine learning statistical techniques for analysis. Specifically, this included Levene's Test, ANOVA, and T-Tests for hypothesis testing, Two-Part Modeling for model building, and K-Means Clustering for machine learning.

Hypothesis Testing:

For investigating assumptions of parametric statistics and comparing means between groups, Levene's Test, ANOVA, and T-Tests were performed on the categorical variables of the

Grand Cayman data. Ultimately, an alpha statistic of 0.05 was used to determine significance at 95% confidence for all tests. There were three primary assumptions to test for parametric statistics, which included independence of samples, normality of the response distribution, and homogeneity of variance. Because data were collected on different individual fish across many years and sites, it was reasonable to assume independence of samples. Therefore, this assumption was not formally tested. As for normality, the large sample size (n = 22,052) allowed for the application of the Central Limit Theorem to the distribution of "DENSITY_INDEX".

Previously, it was described that the distribution of "DENSITY_INDEX" was not a singular bell-like curve in both the zero-inclusive and non-zero data (Appendix D). Most statistical models are built assuming the response variable has this bell shape, which is known formally as a Gaussian Distribution. Therefore, the distributions of "DENSITY_INDEX" violated this parametric assumption. Typically, violations require the use of alternative statistical tests, known as non-parametric tests, but the Central Limit Theorem allowed "DENSITY_INDEX to be tested as though it was a Gaussian distribution. Essentially, the mathematics behind this theorem prove that data samples with observations greater than thirty approximate Gaussian Distributions (Kwak and Kim, 2017). Thus, the large sample sizes for both the zero-inclusive and non-zero data allowed normality to be tested parametrically.

Finally, the homogeneity of variance assumption was assessed using Levene's Test (Levene, 1960). This test determined if the variance of *k* explanatory subcategories were equal. After each test, the results were used to determine what comparison of subcategory means would be performed. Ultimately, a standard ANOVA and T-Test was used for data with homogeneity of variance, and Welch's ANOVA and T-Test were used for data of unequal variance. After hypothesis testing was complete, the next task was to develop a model for interpretation.

Interpretive Modeling:

In this study, modeling techniques were used to produce ecological interpretations of abundance trends over time. Initially, ordinary least squares linear regression models of the form:

$$Y_i = B_0 + B_1 X_i \ \ldots \ B_p X_{ip} + \ \epsilon_i,$$

were considered for the zero-inclusive Grand Cayman data because of their relative simplicity for interpretation. However, when this model was fit, the resulting residual plot, which is a plot showing how values in the dataset compare to model requirements, contained violations. Therefore, other models had to be considered. Fortunately, a "two-part" model was identified as viable for the Grand Cayman Data.

The two-part modeling approach models responses of mixed discrete-continuous variables. Primarily, this approach is used when the outcome Y_i, which in this case Y_i represented the "DENSITY_INDEX" value for the *ith* fish surveyed, has the statistical features $Y_i > 0$ or $Y_i = 0$ (Belotti et al., 2015). This was the predominant structure of the Grand Cayman data, as the true zero observations fell under $Y_i = 0$ and the non-zero observations fell under $Y_i > 0$. The first part of this approach focused on the condition $Y_i = 0$, which involves a model with binary outcomes. This can be thought of as outcomes which are "yes" or "no", or in this case, yes, a fish was observed or no, it was not. A logistic regression model deals with these

binary outcomes, and they are relatively easy to interpret, so this model was used for part one of the two-part approach

Overall, the binomial logistic regression model was fit using only "SCIENTIFIC_FAMILY" as an explanatory variable because knowing the likelihood of observing a family was focal to ecological questions in this study. Also, this increased accuracy in the model as including all individual species would have increased the number of coefficients from 12 family estimates to 256 species estimates. Coefficients, or simply model estimates, are the values associated with each explanatory variable of a model. As the number of coefficients in a model increase, each associated estimate becomes inflated, which decreases the accuracy of interpretations. Additionally, a mathematical equation for the response is used to define coefficients. For binomial logistic regression, this equation is

$$E(Y_i) = P(Y_i \neq 0) = \pi_i = \frac{e^{B0+B1X_{i1}+\cdots B_pX_{ip}}}{1 + e^{B0+B1X_{i1}+\cdots B_pX_{ip}}},$$

where π_i represented the probability that the "DESNITY_INDEX" value for the *ith* family was non-zero. Ultimately, these equations are where model interpretations come from.

Next, the second part of this approach needed to find a model for "DENSITY_INDEX" values of $Y_i > 0$, which meant the non-zero dataset was used. Ultimately, multi-level longitudinal modeling was used for this part because the data were observed temporally. This approach is built on the foundations of linear mixed-effects models, which use fixed and random effects to investigate similarities and differences between variables over time. Fixed effects are variables which are of focal importance to a study, while random effects are variables which are not focally important but play a role in accounting for variability and correlation in a dataset (Roback and Legler, 2021). Longitudinal versions of these models use multiple levels and consider the inclusion of random slope terms. Importantly, the interpretations in this approach mirror ordinary least squares regression models, such that a one unit increase in an explanatory variable will yield an increase in "DENSITY_INDEX" equal to the coefficient estimate of the model.

Level one variables are those whose observations change over time. Therefore, the variables "YEAR", "LOCATION", and "MARINE_RESERVE" were considered plausible for this level. Alternatively, level two variables are those whose observations remain constant over time, which include "SCIENTIFIC_FAMILY" and "DIET". Because of this multi-level structure, comparisons of variability within and between families could be made. Regarding random slope terms, they allow the rate of change for the response to vary according to an explanatory variable. An example of a plausible random slope term that was investigated was "YEAR1998", which accounted for the fact that families "DENSITY_INDEX" values may have changed at different rates over time (Figure 2). Overall, these features made multi-level longitudinal modeling ideal for the second piece of the two-part approach.

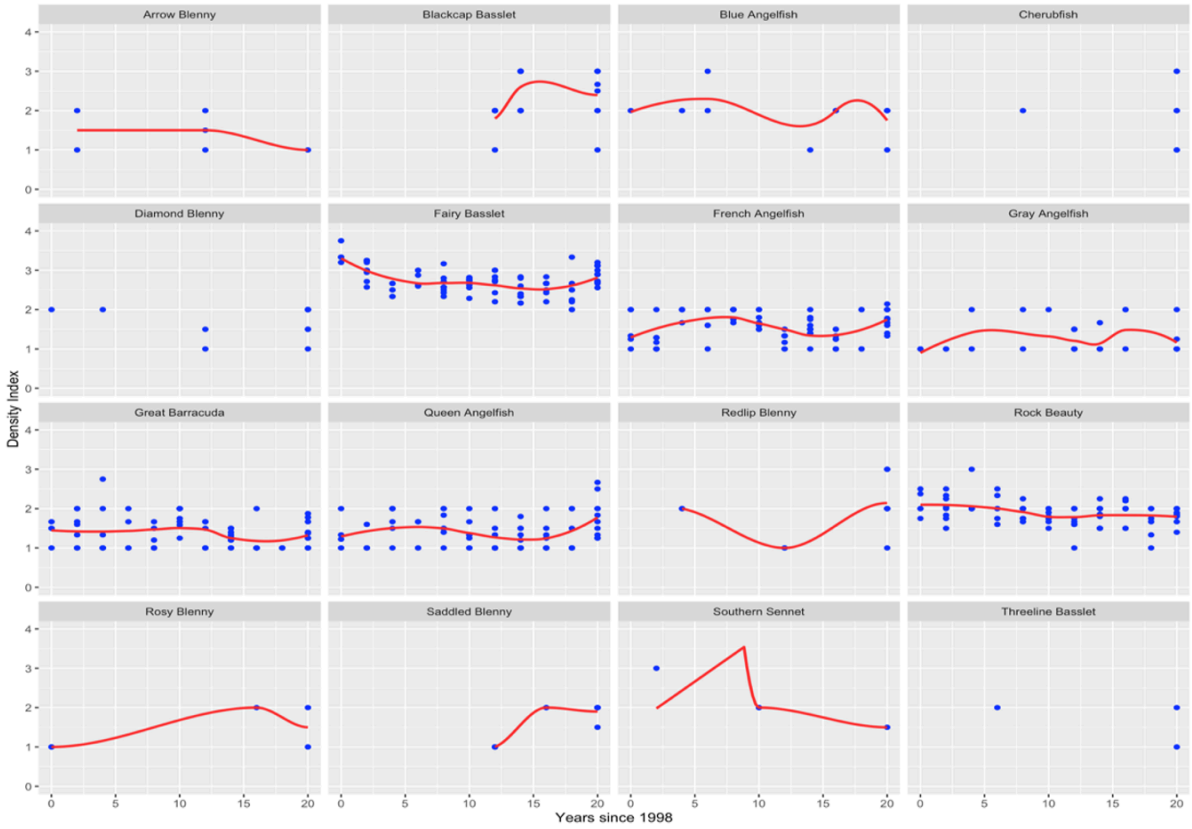


Figure 2: Species' "DENSITY_INDEX" values change at different rates over time. Scatter plots with lines of best fit, shown in red, are pictured for the first sixteen non-zero Grand Cayman dataset species. This investigation aimed to evaluate whether the rate of change in "DENSITY_INDEX" was not the same for species over time and provided support for the use of a random slope term for "YEAR". For Cherubfish, Diamond Blenny, and Threeline Basslet, lines of best fit could not be determined.

When building these models, a structure starting with a model with no explanatory variables which is built upon to reach a final model is used. Conventionally, the most basic model is termed an Unconditional Means Model. Importantly, this model produces an intraclass correlation coefficient (ICC), which measures variability within and between families from the equation $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma^2}$. In this equation, σ^2 represents the total variance within datapoints for a single fish family over time, while σ_u^2 represents the total variance between datapoints of all families over time. Following this model, the secondary model in this structure, an Unconditional Growth Model was built. Unconditional Growth Models are built for investigating the temporal nature of the response. This involves including only time as an explanatory variable, which in this dataset was the variable "YEAR". Collectively, these models provided valuable ecological insights and formed the foundations for finding a final model for interpretation using Maximum Likelihood-Ratio Tests.

When testing explanatory variables for their possible addition to an initial model, those that help capture variability in the response will be deemed necessary to add. From these added variables, additional coefficients will be produced and used for interpretation.

Ultimately, the statistical test known as Maximum Likelihood-Ratios was implemented to test variable addition. Fundamentally, this test compares the efficiency of a nested model against the model with the added variable based on a Chi-Square distribution (Etz, 2018). However, it cannot evaluate random slope terms, and thus that addition was evaluated using a similar variability-based technique known as Akaike's Information Criterion (AIC) (Kuha, 2004).

The significance level in all Maximum Likelihood-Ratio tests was $\alpha = 0.05$, and in addition to single variable additions, tests were performed to investigate the inclusion of interaction variables between "YEAR" and "DIET" in addition to "YEAR" and "SCIENTIFIC_FAMILY". These interactions, which are cases in which one explanatory variable's impact on the response depends on the value of another explanatory variable, were investigated because graphical evidence suggested they may exist. Once the optimal Maximum Likelihood-Ratio model was found, AIC was used to test for the addition of random slopes. Unfortunately, there exists a tradeoff between interpretability and fit in statistics. This means that though some AIC and Maximum Likelihood-Ratio were indicative of better models, the interpretability of these models could be lost due to their complexity. Therefore, the model that had the most variable additions based on Maximum Likelihood-Ratio, produced the lowest AIC value, and was most interpretable was selected as the final model. Ultimately, this piece of the two-part approach achieved the goal of creating an ecologically interpretable model.

Machine Learning:

While the primary goal of this study involved analyzing fish at the family level, looking at specific indicator species can also be beneficial. Several studies have shown that these species, such as the predator Lionfish and algae feeding Rainbow Parrotfish and Surgeonfish, can be critical indicators of reef health (Benkwitt, 2002; Pattengill-Semmens and Semmens, 2003; Williams and Polunin, 2001). Therefore, to perform this species level analysis, an unsupervised machine learning technique known as K-Means Clustering was used. This technique groups numeric observations together based on similarity, which relies on a mathematical distance measurement (Steinley, 2010). For this study, the distance measure used was Euclidean Distance, which measures the distance of a line segment between two data points. For implementation, the Grand Cayman dataset first had to be broken down by years.

Ultimately, "SPECIES_NAME" was used as a marker for datapoints in each year. Because of this, the number of observations, known as the the sample size, for each species also indicated the number of sites they were observed at each year. So, K-Means Clustering ultimately grouped species based on both the density index and location numeric measurements. In the results, this location grouping was named "N_SITES". However, before implementation, the datasets variables by year were scaled because scaling is required in K-Means Clustering. Essentially, scaling variables ensures their observations all occur within an equal numeric range, because no overlap would occur between variables during grouping if they were not.

Next, it was necessary to determine how many groups, the so-called *K* in this technique, would be optimal. Optimization of *K* occurs when the greatest drop in total sum of squares occurs, which can simply be determined by finding an elbow-like bend in `fviz_nbclust` plots in R. For each of the ten years of data collection, the plots indicated the optimal value of *K* was

four (Appendix K). Using this optimal grouping value, K-Means Clustering was performed on data for each year.

Once complete, visualizations of the four optimal K-Means Clusters each year could be visualized using the `fviz_cluster` function in R (Appendix L). To further analyze the clustering results, summary tables were created which contained the centers and sample size based on "DENSITY_INDEX" and "N_SITES". However, scaling the data resulted in centers which were uninterpretable, so they had to be unscaled for reliable interpretation (Appendix M). Additionally, these tables also showed the clustering in which an individual species fell for a given year using the "SPECIES_NAME" variable. This assignment allowed for indicator species to be tracked temporally for drawing ecological conclusions. Overall, this would specifically involve the cluster location and center based on "DENSITY_INDEX" values in 1998, 2008, 2014, and 2018.

Results

Hypothesis testing was carried out using both the zero-inclusive and non-zero data. Levene Tests on the zero-inclusive data for the variables "DIET", "SCIENTIFIC_FAMILY", "LOCATION", and "MARINE_RESERVE" all yielded significant p-values based on an alpha of 0.05, which meant the variances between the subcategories of these variables were unequal (Appendix N). In the Levene-Tests for non-zero data, the same variables were investigated. Overall, significant p-values were observed for "DIET" and "SCIENTIFIC_FAMILY", while insignificant p-values were observed for "LOCATION" and "MARINE_RESERVE" (Appendix N). Again, the variables with significant p-values had evidence supporting unequal variance, while those with insignificant p-value lacked this evidence, meaning they had equal variance. Ultimately, these findings determined the correct testing procedure for comparisons of means.

Because the zero-inclusive categorical variables all had unequal variance, comparison of means tests had to be performed using either Welch's ANOVA or Welch's T-Test based on the total number of subcategories in each variable. For "DIET", "SCIENTIFIC_FAMILY", and "LOCATION", which all had more than two subcategories, Welch's ANOVA was used. Alternatively, Welch's T-Test was used for "MARINE_RESERVE" because it had only two subcategories. Overall, the p-values from each test were all significant ($\alpha = 0.05$), which provided evidence that the mean differences in "DENSITY_INDEX" for their respective subcategories were not zero (Appendix O).

Regarding the non-zero data, standard ANOVA and an equal variance t-test were used to compare subcategory means for the "LOCATION" and "MARINE_RESERVE" variables respectively. For "DIET and "SCIENTIFIC_FAMILY", the comparisons were conducted using Welch's ANOVA. Overall, significant p-values ($\alpha = 0.05$) were obtained in the tests for "DIET and "SCIENTIFIC_FAMILY", but not for "LOCATION" and "MARINE_RESERVE" (Appendix O). As before, the significant p-values for "DIET" and "SCIENTIFIC_FAMILY" indicated mean differences in "DENSITY_INDEX" for their respective subcategories were not zero. However, for "LOCATION" and "MARINE_RESERVE" the tests provided evidence that their true mean subcategory "DENSITY_INDEX" differences could plausibly be zero.

After hypothesis testing, model building for interpretation was carried out based on the two-part approach. Ultimately, a logistic regression model was fit to the zero-inclusive data where coefficients estimated increases or decreases in the probability of a family's

"DENSITY_INDEX" being zero or non-zero. The baseline for estimates from this model, β_0, was measured for the family Apogonidae since these models create coefficients in alphabetical order. Overall, of the eleven additional model coefficients estimates, eight achieved significant p-values ($\alpha = 0.05$; Figure 3). Based on these outputs, the probability that a "DENSITY_INDEX" observation of a certain family was non-zero could be found using the binomial logistic regression equation previously provided (Table 1). If the probability that a family's observations were non-zero was high enough, typically greater than 20 percent, then further interpretation for that family was made using the longitudinal multi-level linear mixed effect models. Overall, the families Holocentridae, Labridae, Lutjanidae, Pomacentridae, Scaridae, Serranidae, and Other met this probability mark.

```
Call:
glm(formula = DENSITY_INDEX ~ SCIENTIFIC_FAMILY, family = binomial(link = "logit"),
    data = CAYMAN_LOGISTIC_DATA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1950  -0.7387  -0.6628  -0.4241   2.2151

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.17682    0.14227 -15.301  < 2e-16 ***
SCIENTIFIC_FAMILYCarangidae -0.07192    0.17599  -0.409 0.682793
SCIENTIFIC_FAMILYGobiidae   -0.18648    0.16765  -1.112 0.266011
SCIENTIFIC_FAMILYHaemulidae  0.58481    0.15873   3.684 0.000229 ***
SCIENTIFIC_FAMILYHolocentridae  2.21810    0.16658  13.316  < 2e-16 ***
SCIENTIFIC_FAMILYLabridae    1.15606    0.15512   7.453 9.14e-14 ***
SCIENTIFIC_FAMILYLutjanidae  1.38316    0.15952   8.671  < 2e-16 ***
SCIENTIFIC_FAMILYOther       1.01748    0.14443   7.045 1.86e-12 ***
SCIENTIFIC_FAMILYPomacentridae  2.17520    0.15320  14.199  < 2e-16 ***
SCIENTIFIC_FAMILYScaridae    1.58743    0.15485  10.251  < 2e-16 ***
SCIENTIFIC_FAMILYSerranidae  0.77282    0.15027   5.143 2.71e-07 ***
SCIENTIFIC_FAMILYSparidae    0.17534    0.18808   0.932 0.351224
```

Figure 3: R output for the Logistic Regression Model fit to the zero-inclusive dataset. Model equation, coefficient estimates, and corresponding p-values are depicted. Asterisks (***) indicate significant p-value estimates at p<0.01. Intercept estimate corresponds to observations of the Apogonidae family.

Logistic Regression Probabilities of Non-Zero Observations By Families

Family	Model P-Value	Probability Non-Zero
Apogonidae	2e-16	0.102 (10.2%)
Carangidae	0.683	0.102 (10.2%)
Gobiidae	0.266	0.102 (10.2%)
Sparidae	0.351	0.102 (10.2%)
Haemulidae	0.000229	0.169 (16.9%)
Holocentridae	2e-16	0.510 (51.0%)
Labridae	9.14e-14	0.264 (26.4%)
Lutjanidae	2e-16	0.311 (31.1%)
Pomacentridae	1.86e-12	0.499 (49.9%)
Scaridae	2e-16	0.356 (35.6%)
Serranidae	2e-16	0.197 (19.7%)
Other	2.71e-7	0.238 (23.8%)

Table 1: Resulting probabilities that an observation from a particular family was non-zero. Probabilities were calculated using the logistic regression equation associated with the model. Corresponding p-values from the model are provided for each family. Findings indicate some families are more likely to be observed than others.

In the non-zero data, the initial Unconditional Means Model did not include explanatory variables for "DENSITY_INDEX". This model was of the form:

$$Y_{ij} = \alpha_0 + u_i + \epsilon_{ij}, \text{ with } u_i \sim \mathcal{N}(0, \sigma_u^2) \; and \; \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

where Y_{ij} defined the "DENSITY_INDEX" value for species i in year j. Overall, this model produced a single fixed effect coefficient, α_0, which was 1.74. This represented the estimated average "DENSITY_INDEX" value across all species and years (Appendix P). The intraclass correlation coefficient (ICC) was also found in this model based on the equation:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2},$$

where it was calculated at 0.423. This indicated that 42.3 percent of the variation in "DENSITY_INDEX" occurred between observations of families, while 57.7 percent of the variation occured within a single family's observations. Additionally, this model produced estimates for standard deviations of random effects. These included $\sigma = 0.5151$ and $\sigma_u = 0.4413$, where σ represented the standard deviation in "DENSITY_INDEX" values for a single fish family each year and σ_u was the standard deviation in "DENSITY_INDEX" values between families averaged across all twenty years (1998 to 2018).

The second model for the non-zero data was the Unconditional Growth Model, which introduced time, indicated by "YEAR", as a level one variable (Appendix P). However, this model

did not include any predictors at level two. Additionally, this model included a random slope term for "YEAR". Overall, it was of the form:

$$Y_{ij} = B_0 + B_1 YEAR_{ij} + u_i + v_i YEAR_{ij} + \epsilon_{ij}, \text{ with}$$
$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & 1 \\ \rho_{uv}\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix}\right) \text{ and } \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

This structure is rather complex, but it can be simply interpreted as Y_{ij} defines the "DENSITY_INDEX" value for species i in year j. The fixed-effects summary coefficients in the output for this model provided estimates of 1.84 for B_0 and -0.008 for B_1, both of which obtaining significant p-values ($\alpha = 0.05$). Overall, B_0 represented the average "DENSITY_INDEX" value across all species in 1998, while B_1 indicated that each year the average decrease in "DENSITY_INDEX" was -0.008. In the random effects table, estimates of $\sigma_u, \sigma_v, and\ \sigma$ were obtained at 0.508, 0.019, and 0.504 respectively.

In this model, σ represented the standard deviation in "DENSITY_INDEX" values for an individual family, σ_u represented the standard deviation in "DENSITY_INDEX" values between all families in 1998, and σ_v represented the standard deviation in rates of change in "DENSITY_INDEX" values during the twenty-year observation period. Finally, the most meaningful estimate of this model was ρ_{uv} at -0.52, which measured correlation between families 1998 "DENSITY_INDEX" values and their rates of change in "DENSITY_INDEX" from 1998 to 2018. Since both B_1 and ρ_{uv} were less than zero, species with higher "DENSITY_INDEX" values in 1998 had larger decreases in densities between 1998 and 2018 than species with lower 1998 values. Essentially, this meant the higher a species' 1998 "DENSITY_INDEX" value was, the faster and steeper their decrease in "DENSITY_INDEX" was over time.

Finally, Maximum Likelihood-Ratios, AIC, and interpretability were used to produce an optimal linear mixed effects model. Overall, an explanatory variable improved this model if it yielded lower AIC values and higher Maximum Likelihood-Ratio values. First, the categorical variables "DIET", "SCIENTIFIC_FAMILY", "LOCATION", and "MARINE_RESERVE" were tested as additive effects. In the results, AIC decreased, and Maximum Likelihood-Ratio values, denoted "LRT", increased with the addition of "DIET" and "SCIENTIFIC_FAMILY" but not for "LOCATION" or "MARINE_RSERVE". Thus, "LOCATION" and "MARINE_RESERVE" were excluded from the final model, and "DIET" and "SCIENTIFIC_FAMILY" were added to it.

Next, the interaction effect of "YEAR" and "DIET" and "YEAR" and "SCIENTIFIC_FAMILY" were tested. Overall, AIC decreased for the interaction between "YEAR" and "SCIENTIFIC_FAMILY", but not for "YEAR" and "DIET". Additionally, the Maximum Likelihood-Ratio test value increased for "YEAR" and "SCIENTIFIC_FAMILY" while the value for "YEAR" and "DIET" did not (Appendix Q). Therefore, adding the interaction between "YEAR" and "SCIENTIFIC_FAMILY" was considered an improvement for the final model. However, this addition would have increased complexity of the model, and greatly decreased its interpretability.

Increasing complexity by including more coefficient estimates can decrease accuracy of interpretations from a model. This loss of significance occurred when adding the interaction between "YEAR" and "SCIENTIFIC_FAMILY", and therefore, the interaction was excluded from the final model (Appendix R). Notedly, had the goal of this study been creating an efficient

prediction model, the interaction would have been kept because interpretability is not important for prediction. Finally, the addition of a random slope term for "YEAR" resulted in a significant drop in AIC (Appendix Q). This random slope term was easily interpretable, added only a single coefficient, and increased accuracy, so it was added to the final model.

Altogether, the final model included the variables "YEAR", "DIET", and "SCIENTIFIC_FAMILY" and included a random slope variable for "YEAR". However, before interpretations were made from this model, diagnostics were generated to check the assumptions of normality, independence, and linearity. These assumptions mirror those for parametric statistics, which were explained previously, besides the assumption of linearity. Essentially, linearity in this case meant "DENSITY_INDEX" increased linearly with any increases in model explanatory variables. The assumptions for normality and linearity were checked using a QQ plot and residuals versus fitted plot respectively. In the QQ plot, when points match closely with the line, normality is met. Meanwhile, residuals versus fitted plots verify linearity when patterns do not occur in the graphs data points. Fortunately, both plots verified their assumption was met for this model, which meant interpretations from the model were valid (Appendix S)

Collectively, the final Multi-Level Linear Mixed Effects model produced twenty coefficient estimates, where ten were significant at an alpha of 0.10 (Figure 4). The intercept coefficient, B_0, was 1.916, which estimated the expected "DENSITY_INDEX" value in 1998 for the family Apogonidae with a diet of algae. Because the estimated coefficients for diets of coral and crustaceans in addition to families of Carangidae, Haemulidae, Labridae, Lutjanidae, Other, Scaridae, Serranidae, and Sparidae fish were all insignificant in this model, their estimated "DENSITY_INDEX" values in 1998 were also 1.916. All other coefficients had significant p-values and could thus be interpretated. Here, the interpretations for the significant estimates are included in the context of the model

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: DENSITY_INDEX ~ YEAR1998 + DIET + SCIENTIFIC_FAMILY + (YEAR1998 |
    SPECIES_NAME)
   Data: CAYMAN_NO_ZEROS

REML criterion at convergence: 8318.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.3376 -0.6208 -0.0392  0.6299  4.0417

Random effects:
 Groups       Name        Variance  Std.Dev. Corr
 SPECIES_NAME (Intercept) 0.2127850 0.46129
              YEAR1998    0.0003699 0.01923  -0.62
 Residual                 0.2541073 0.50409
Number of obs: 5208, groups:  SPECIES_NAME, 231

Fixed effects:
                                Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                     1.915148   0.241492 258.333586   7.930 6.60e-14 ***
YEAR1998                       -0.008237   0.002016 114.681163  -4.085 8.19e-05 ***
DIETCoral                      -0.004197   0.200018 157.523733  -0.021   0.9833
DIETCrustaceans                -0.149682   0.134693 177.401252  -1.111   0.2679
DIETEchinoids                  -0.483278   0.243851 215.826263  -1.982   0.0488 *
DIETFish                       -0.298701   0.143444 180.335327  -2.082   0.0387 *
DIETSponges                    -0.351591   0.144600 171.116407  -2.431   0.0161 *
DIETUnidentified               -0.287394   0.134712 191.527127  -2.133   0.0342 *
DIETZooplankton                 0.257617   0.152187 183.787536   1.693   0.0922 .
SCIENTIFIC_FAMILYCarangidae     0.267566   0.273778 255.661065   0.977   0.3293
SCIENTIFIC_FAMILYGobiidae       0.416462   0.236033 262.306885   1.764   0.0788 .
SCIENTIFIC_FAMILYHaemulidae     0.134195   0.247645 263.791814   0.542   0.5884
SCIENTIFIC_FAMILYHolocentridae  0.492670   0.273139 219.684135   1.804   0.0726 .
SCIENTIFIC_FAMILYLabridae       0.076888   0.243233 246.974398   0.316   0.7522
SCIENTIFIC_FAMILYLutjanidae     0.164011   0.262733 241.374976   0.624   0.5331
SCIENTIFIC_FAMILYOther          0.039381   0.217654 267.493359   0.181   0.8566
SCIENTIFIC_FAMILYPomacentridae  0.499964   0.247085 259.102060   2.023   0.0441 *
SCIENTIFIC_FAMILYScaridae       0.040123   0.263446 236.895675   0.152   0.8791
SCIENTIFIC_FAMILYSerranidae    -0.075167   0.233803 254.849451  -0.321   0.7481
SCIENTIFIC_FAMILYSparidae       0.032084   0.294316 242.317112   0.109   0.9133
```

Figure 4: R output for the final longitudinal linear mixed-effects model from the non-zero data. Model equation, coefficient estimates, and corresponding p-values are depicted. Table for random effects estimating $\sigma_u, \sigma_v, \sigma, and\ \rho_{uv}$ is also shown. Asterisks (*) represent significance at p<0.05, while periods (.) represent significance at p<0.10.

The coefficient for "YEAR", B_1, was -0.00823, which indicated the average yearly decrease in "DENSITY_INDEX" was -0.00823. The coefficient for a diet of echinoids, B_4, was -0.483, which was the expected decrease in 1998 "DENSITY_INDEX" values for fish feeding on echinoids compared to fish feeding on algae. For diets of fish, B_5 was -0.299, which was the expected decrease in 1998 "DENSITY_INDEX" values for fish feeding on other fish (Carnivorous) compared to fish feeding on algae. Regarding a diet of sponges, B_6 was -0.352, which was the expected decrease in 1998 "DENSITY_INDEX" values for fish feeding on sponges compared to fish feeding on algae. The coefficient for unidentified diets, B_7, was -0.287, which was the expected decrease in 1998 "DENSITY_INDEX" values for fish with unidentified diets compared to fish feeding on algae. Finally, the coefficient for a diet of zooplankton, B_8, was 0.258, which was the expected increase in 1998 "DENSITY_INDEX" values for fish feeding on zooplankton compared to fish feeding on algae. In addition to these diet differences, three significant differences occurred in estimates for families

The coefficient for the family Gobiidae, B_{10}, was 0.416, which was the expected increase in 1998 "DENSITY_INDEX" values for species of the family Gobiidae compared to species of Apogonidae. The coefficient for the family Holocentridae, B_{12}, was 0.493, which was the expected increase in 1998 "DENSITY_INDEX" values for species of the family Holocentridae compared to species of Apogonidae. Finally, the coefficient for the family Pomacentridae, B_{12}, was 0.500, which was the expected increase in 1998 "DENSITY_INDEX" values for species of the family Pomacentridae compared to species of Apogonidae. Notably, the coefficient for Gobiidae was not very insightful, as Gobiidae did not meet the probability threshold for significance in the binomial logistic regression model.

For the random effects table of this model, estimates of $\sigma_u, \sigma_v, and\ \sigma$ were obtained at 0.461, 0.019, and 0.504 respectively. Additionally, an estimate of ρ_{uv} was obtained at -0.62. The interpretations for these estimates are identical to the Unconditional Growth Model. However, it is worth noting that ρ_{uv} was more negative in this model, -0.62 compared to -0.52, which indicated species with higher 1998 "DENSITY_INDEX" values had even faster density declines temporally than was observed in the Unconditional Growth Model.

Using K-Means Clustering, the range of maximum "DENSITY_INDEX" centers from 1998 to 2008 was found to be between 2.18 and 3.25, while the range of minimum centers was between 0.039 and 0.11. Additionally, the range of maximum "DENSITY_INDEX" centers from 2010 to 2018 was between 2.04 and 2.20, while the range of minimum centers was between 0 and 0.15. For the indicator Lionfish species in the years 1998, 2008, 2014, and 2018, its cluster centers for "DENSITY_INDEX" were 0.109, 0.045, 0.086, and 0.926 respectively (Appendix T). Additionally, for the Rainbow Parrotfish in the years 1998, 2008, 2014, and 2018, its cluster centers for "DENSITY_INDEX" were 1.57, 1.05, 2.20, and 0.926 respectively. Finally, for the Ocean Surgeonfish in the years 1998, 2008, 2014, and 2018, its cluster centers for "DENSITY_INDEX" were 1.57, 0.045, 2.20, and 2.14 respectively.

Discussion & Conclusions

This study aimed to understand how species abundance has varied on Grand Cayman Island from an ecological perspective. For the variables "LOCATION" and "MARINE_RESERVE", hypothesis tests did not yield significant p-values for differences in means of their subcategories. This indicates that fish on Grand Cayman Island do not differ in their densities across locations within MPA's or outside MPA's. Overall, this finding is alarming, as it has previously been shown that variations in density should exist between MPA and non-MPA sites if MPA's are efficient in their design (Mosquera et al., 2000). Therefore, it is likely that a design flaw exists for MPA's on Grand Cayman.

Of the five factors contributing to MPA ineffectiveness, it is likely that the presence of continuous habitat allowing unconstrained movement of fish is hurting the MPA's of Grand Cayman. This is because previous studies have noted that extremely limited fishing is allowed at these sites, and the enforcement of these limitations is prioritized by Grand Cayman authorities (Williams and Polunin, 2001). Additionally, the MPA's of the island were introduced in 1986 and they cover over 48% of the surrounding coastline, which has given them ample time to become effective. Therefore, it is likely many of the coral reefs are not completely within an MPA, which restricts the habitat protection of many species as they migrate throughout each site. Commercial fishing vessels could access parts of the habitat, which these fish are likely traveling

through daily, leading to their decline regardless of the perceived protections of the MPA. Therefore, future government initiatives should focus on redesigning the areas which the MPA's of the island cover, with a prioritization on habitat overlap.

In the final interpretation model, it was observed that overall densities of all fish families on Grand Cayman are decreasing yearly. However, this observation could only be applied to families which had a probability of being observed over 20% in the binomial logistic regression model, which included Holocentridae, Labridae, Lutjanidae, Pomacentridae, Scaridae, Serranidae, and Other. Of these seven families, three are major predator families, Lutjanidae, Serranidae, and Holocentridae, and two contain the primary herbivorous reef fish, Scarridae and Pomacentridae. This means over 71% of the families shown to be decreasing yearly are major predators and herbivores. Overall, this indicates that these major indicator families are following the same trends observed at other reefs across the Caribbean (Armsworth et al., 2007; Pattengill-Semmens and Semmens, 2003).

Because high predator densities increase biodiversity, it is likely that the reef ecosystems of Grand Cayman are becoming more homogeneous due to their decline. Homogeneity of an environment is an indicator of declining richness, which collectively with abundance measurements like density are two primary indicators of ecosystem health in ecology. Therefore, this finding provides evidence of major health declines of coral reefs on Grand Cayman.

In addition, the declines in herbivory found in the model provide further evidence for health declines. One of these herbivore families, Pomacentridae, was shown to have a higher 1998 "DENSITY_INDEX" in the final model than all other families. This indicates that this family saw the most drastic declines in density by 2018 because of the interpretation of the negative intraclass correlation coefficient of the model. Unfortunately, it is well documented that the loss of herbivorous species produces overgrowth of macroalgae on coral reefs (Williams and Polunin, 2001; Williamson et al., 2014). Therefore, this major decline indicates it is likely the reefs on Grand Cayman are in the process of a "phase-shift" from coral to macroalgae dominated. Future studies should focus on investigating the extent to which total coral area has declined on Grand Cayman to begin combating this likely phenomenon.

Primarily, diets of coral and sponges serve as key indicators of the ecologically harmful practice of dietary expansion. This means densities of fish families feeding on these diets should be higher than those of other diets if it were occurring. Alternatively, lower densities of families with these diets would provide evidence of healthy dieting for these fish. However, no clear evidence to support or deny the occurrence of dietary expansion on Grand Cayman was found in the final model. Overall, the density of coral feeders was higher than most other diets, while the density of sponge feeders was much lower than most other diets. Therefore, further analysis of dieting practices must be carried out to make future conclusions about dietary expansion on Grand Cayman. While model building provided the opportunity to look more broadly at abundance trends over time, the use of K-Means Clustering allowed for the investigation into trends for specific species.

When investigating the invasive Lionfish species, it was observed that the species was grouped in clusters containing a higher center by the year 2018 than in 1998, 2008, or 2014. This indicates that abundance of the Lionfish has increased over time. A previous found that increasing Lionfish densities negatively affects both the abundance and biomass of native reef

fish (Benkwitt, 2015). Thus, the indication of increasing lionfish abundance in K-Means Clustering could help explain the general decay in "DENSITY_INDEX" values observed while modeling.

Two other species that were investigated using K-Means Clustering were the Rainbow Parrotfish and Ocean Surgeonfish, which are primary herbivores. Previous researchers have observed that increasing abundance of Parrotfish and Surgeonfish macroalgae feeders decreases macroalgae cover as expected (Pattengill-Semmens and Semmens, 2003). Using the K-Means Clusters for the Ocean Surgeonfish and Rainbow Parrotfish, this finding was assessed for Grand Cayman. Overall, both the Rainbow Parrotfish and Ocean Surgeonfish were grouped with clusters of lower "DENSITY_INDEX" centers when analyzed first from 1998 to 2008 and then from 2014 to 2018. This provides even more evidence of herbivore loss on Grand Cayman, and because these species specifically have been shown to control macroalgae, it is almost certain that "phase-shifts" are occurring on the island.

Collectively, the findings in this report indicate that species abundance on Grand Cayman Island has declined significantly between 1998 and 2018. Ultimately, the understanding of these declines produced through a statistical lens here are beneficial for identifying target families and species for recovery initiatives. Nevertheless, the marine ecosystems of the island should be considered unhealthy, and remediation must be prioritized. Therefore, future research and government initiatives should focus on restructuring MPA boundaries, quantifying the extent of macroalgae overgrowth, implementing further commercial fishing limitations for major predator and herbivore fish, and developing a community wide healthy reefs education program for both permanent residence and recreational tourists.

Acknowledgements

I would like to thank Professor Abhishek Chakraborty of Lawrence University for his time and guidance as my supervisor throughout this report. Additionally, I would like to thank Professor Andrew Sage of Lawrence University for his teachings from STAT 455 on modeling longitudinal data. I would also like to thank Professor Bart De Stasio of Lawrence University for providing the data used in this report and for his advisory role throughout the project. Finally, I want to thank Professor Brian Piasecki of Lawrence University for his feedback and insights as my section leader throughout my senior capstone course in the winter of 2023.

Project Repository

The R markdown document containing the code used in this report can be accessed at my Github repository: https://github.com/lu2021adam/Senior_Experience_Project.git

References

Alvarez-Filip, L., Paddack, M. J., Collen, B., Robertson, D. R., and Côté, I. M. (2015). Simplification of Caribbean reef-fish assemblages over decades of coral reef degredation. *Public Library of Science ONE. 10*(4), 1-14.

Armsworth, P. R., Chan, K. M. A., Daily, G. C., Ehrlich, P. R., Kremen, C., Ricketts, T. H., and Sanjayan, M. A. (2007). Ecosystem-service science and the way forward for conservation. *Conservation Biology.* 21, 1383–1384.

Belotti, F., Deb, P., Norton, E. C. and Manning, W. G. (2015). Twopm: Two-part models. *The Stata Journal. 15*(1), 3-20.

Benkwitt, C. E. (2015). Non-linear effects of invasive lionfish density on native coral-reef fish communities. *Biological Invasions.* 17, 1383-1395.

Chesson, P. (2000). Mechanisms and maintenance of species diversity. *Annual Review of Ecology and Systematics*. 31, 343–366.

Edgar, G. J. (2011). Does the global network of marine protected areas provide an adequate safety net for marine biodiversity? *Aquatic Conservation.* 21, 313–316.

Edgar, G. J., Stuart-Smith, R. D., Willis, T. J., Kininmonth, S., Baker, S. C., Banks, S., Barrett, N. S., Becerro, M. A., et al. (2014). Global conservation outcomes depend on marine protected areas with five key features. *Nature.* 506, 216-220.

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances In Methods and Practices in Psychological Science. 1*(1), 60-69.

Hintz, J. L., and Nelson, R. D. (1998). Violin plots: a box plot density trace synergism. *The American Statistician. 52*(2), 181-184.

Jackson, J. B. C., Kirby, M. X., Berger, W. H., Bjorndal, K. A., et al. (2001) Historical overfishing and the recent collapse of coastal ecosystems. *Science* 293, 629–638

Jennings, S., and Polunin, N. V. C. (1996). Effects of fishing effort and catch rate upon the structure and biomass of Fijian reef fish communities. *Journal of Applied Ecology.* 33, 400-412.

Jones, B. (1994). Geology of the Cayman Islands. In *The Cayman Islands: Natural History and Biogeography*, pp. 13-49. Berlin, Germany: Springer Science+Business Media.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research. 33*(2), 188-229.

Kwak, S. G., and Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology. 70*(2), 144-156.

Levene, H. (1960). Robust testes for equality of variances. In *Contributions to Probability and Statistics (I. Olkin, ed.)* 278– 292. Stanford Univ. Press, Palo Alto, CA.

Lobato, F. L., Barneche, D. R., Siqueira, A. C., Liedke, A. M. R., Linder, A., Pie, M. R., Bellwood, D. R., and Floeter, S. R. (2014). Diet and diversification in the evolution of coral reef fishes. *Public Library of Science ONE. 9(*7), e102094.

Mcfield, M. D., and Kramer, P. R. (2007). In *Healthy Reefs for Healthy People: A Guide to Indicators of Reef Health and Social Well-Being in the Mesoamerican Reef Region*, pp. 1-19. Franklin Trade Graphics, Miami, FL.

Mora, C., Andrèfouët, S., Costello, M. J., Kranenburg, C., et al. (2006). Coral reefs and the global network of Marine Protected Areas. *Science*. 312, 1750–1751.

Ogden, J. C., and Lobel, P. S. (1978). The role of herbivorous fishes and urchins in coral reef communities. *Environmental Biology of Fishes. 3*(1), 49-63.

Pattengill-Semmens, C. V., and Semmens, B. X. (2003). Status of coral reefs of Little Cayman and Grand Cayman, British West Indies, in 1999 (Part 2: Fishes). *Atoll Research Bulletin. 496*(12), 226-247.

Pauly, D., Christensen, V., Dalsgaard, J., Froese, R., and Torres, F. J. (1998). Fishing down marine food webs. *Science* 279, 860–863.

Polunin, N. V. C., and Roberts, C. M. (1993). Greater biomass and value of target coral-reef fishes in two small Caribbean marine reserves. *Marine Ecology Progressive Series.* 100, 167-176.

Robach, P., and Legler, J. (2021), Two-level Longitudinal Data. In *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R,* pp. 263-320. CRC Press, Boca Raton, FL.

Semmler, R. F., Sanders, N. J., CaraDonna, P. J., Baird, A. H., Jing, X., Robinson, J. P. W., Graham, N. A. J., and Keith, S. A. (2022). Reef fishes weaken dietary preferences after coral mortality, altering resource overlap. *The Journal of Animal Ecology. 91*(10), 2125-2134.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology. 59*(1), 1-34.

Timpe, A. (2018). A survey of Grand Cayman reefs. *Unpublished Manuscript.*

Williams, I. D., and Polunin, N. V. C. (2001). Large-scale associations between macroalgal cover and grazer biomass on mid-depth reefs in the Caribbean. *Coral Reefs.* 19, 356-358.

Williamson, D. H., Ceccarelli, D. M., Evans, R. D., Jones, G. P., and Russ, G. R. (2014). Habitat dynamics, marine reserve status, and the decline and recovery of coral reef fish communities. *Ecology and Evolution. 4*(4), 337-354.

Appendix A

Data Structuring: A breakdown of the LUMP dataset used in this study

Altogether, the LUMP .csv file contained data on “SPECIES_NAME”, “DIET”, “LOCATION”, “SPECIES_ID”, “DENSITY_INDEX”, and “YEAR”. In full, this made the original data structure six columns with 22,052 rows of observations for 246 unique fish species. Each column represented a unique variable, which are groups of characteristics or numeric quantities that can be measured or counted. Two additional variables were also added to the dataset, named “SCIENTIFIC_FAMILY” and “MARINE_RESERVE”, using ‘mutate’ and ‘case_when’ functions in R.

Firstly, “SCIENTIFIC_FAMILY” was created by specifying each of the fifty-six families represented by fish species in the “SPECIES_NAME” variable as a case for ‘case_when’ in R. Of these families, those with six or more species within them were then chosen as subcategories for “SCIENTIFIC_FAMILY” using ‘mutate’. All other families were then grouped into a subcategory called “Other”, which resulted in a total of twelve subcategories for the variable.

Finally, the variable “MARINE_RESERVE” was added to the dataset with the ‘mutate’ and ‘if_else’ functions in R. Before grouping with these functions, the Cayman Islands Government Department of Environment marine parks boundary map was used to determine if a site was in an MPA (Appendix C). In total, four of the eleven survey locations were found outside of marine reserves on the island. Therefore, these sites were represented by a “No” subcategory within the “MARINE_RESERVE” variable, and all other sites were indicated by a “Yes” subcategory using ‘if_else’.

Appendix B

Exploratory Data Analysis

Since their introduction in 1970 by statistician John Tukey, the techniques of exploratory data analysis have become a principal component in statistical analysis. Collectively, they help identify patterns, relationships, and basic mathematical tendencies, like those relating to measures of center or variance, for variables of interest. Additionally, they help identify statistical techniques that could be useful for data analysis. Because the ecological questions of this study focused on abundance, the variable "DENSITY_INDEX" was considered the response variable, Y, and was the focal point of this section.

Response Patterning:

Most statistical techniques make assumptions about the distribution of the response variable. Fundamentally, distributions are functions that specify all possible values for a variable and quantify the relative frequency of any single value. Behind the scenes, models, algorithms, and hypothesis tests use distributions to build probability-based expectations of a response in relation to predictors. For visualizing the distribution of "DENSITY_INDEX", a density plot was built in R using the `ggplot()` function. This resulted in a distribution with several bell-shaped curves centered around the values zero, one, two, and three (Appendix D). This did not resemble a common statistical distribution, but it was representative of the way the data were collected. Primarily, one diver surveyed a site each year, which resulted in most datapoints, 20,136 of 22,056, being whole number REEF integers. However, some datapoints, 1,916 of 22,056, had to be averaged because the site was dived more than once in a year. This created the minor variations away from the peaks of these bell-shaped curves.

Interestingly, the largest curve peaked at the value zero. In the dataset, zero represented a "true zero value", because when used it indicated no individuals of a particular species were observed during a dive. In total, 76.4% of observations were true zeros, which meant the response variable was what statisticians call zero-inflated. Immediately this raised questions about the distribution of "DENSITY_INDEX" without the true zeros. Therefore, another density plot was created for all non-zero observations, which showed higher observation frequencies around the value two with more of a singular bell-shaped curve (Appendix D). However, peaks still existed around the values one and three. Collectively, the shape of these distributions would be used later for model building, but the next task was to identify any relationships between the response and the other variables in the dataset.

Variable Relationships:

Often, it can be tempting to make assumptions about expectations of biological data based on prior knowledge. For instance, it might be assumed that, given extensive research on coral reef decay due human disturbances, "DENSITY_INDEX" values will decrease over time in the Grand Cayman dataset. However, the biological world is extremely unpredictable, which makes exploring graphical relationships between variables essential. In statistics, the variables that have influence on a response are known as explanatory variables. Though any variable has potential to be an explanatory variable, only those that have statistically significant influence on a response will end up in a final model.

Significance, in this context, describes a variable that produces a trend in "DENSITY_INDEX" that would not be found either in nature or by chance. Ultimately, through exploratory plots, insights as to which variables will most likely have significant influence can be found. Therefore, exploratory violin plots were used to explore categorical variable relationships, while lattice plots were built to explore numeric variable relationships to "DENSITY_INDEX". Importantly, they were built for both the zero-inflated data and the non-zero data to further investigate the impact of true zero values.

Violin plots are a hybrid mixture of box and whisker plots and kernel density plots (Hintz and Nelson, 1998). They combine local densities of observations and basic summary statistics, which makes them useful for interpreting categorical relationships. For this analysis, violin plots were generated with shapes representing mean "DENSITY_INDEX" values for "LOCATION", "SCIENTIFIC _FAMILY", "DIET", and "MARINE_RESERVE". The resulting plots for "LOCATION" and "MARINE_RESERVE", showed little variation in densities and means for the subcategories of these variables. Meanwhile, the violin plots for "SCIENTIFIC_FAMILY" and "DIET" showed notable variations in their subcategory densities and means (Appendix E & F).

For "SCIENTIFIC_FAMILY", the zero-inclusive violin plot contained a wider range of observed densities for the families Holocentridae and Pomacentridae. Additionally, they had significantly higher mean densities than all other families. This finding was also true in the non-zero plot, but Sparidae and Gobiidae also had notably higher mean density values in this group. Interestingly, these two families went from extremely low mean density families in the zero-inclusive plot, to high mean density families in the non-zero plot. This indicates that select species of these families are extremely rare, leading to many zero values, while other species are very common, leading to higher observations when the zeros are removed. Finally, it was observed in the non-zero plot that the family Serranidae had a particularly low average density index, with most observations occurring at the value one.

When looking at "DIET", the zero-inclusive plot showed species feeding on coral had the highest mean density index values followed by algae and zooplankton. Additionally, species with unidentified diets and those feeding on echinoids had extremely low mean density index values. However, in the plot for non-zero values, there was variation in these findings. Namely, the primary density observation was higher compared to the zero-inclusive plot. Additionally, zooplankton and algae feeders now had the highest mean densities, but sponge and fish feeders were now the lowest.

Perhaps the best visualization of numeric relationships is observed with a scatter plot. These plots include a point for every individual observation in the cartesian plane, commonly known as the xy plane, between the variables of interest. Additionally, they have the capability to work in conjuncture with categorical variables by using unique point shapes for different subcategories. In the Grand Cayman data, the primary numeric explanatory variable was "YEAR". For investigating its relationship to the response, a specialized scatter plot, known as a lattice plot, was used. Overall, lattice plots were made for "YEAR" in conjuncture with the categorical variables "LOCATION", "MARINE_RESERVE", "DIET", and "SCIENTIFIC_FAMILY". Additionally, a basic one-to-one plot of "DENSITY_INDEX" by "YEAR" was also made.

In the one-to-one lattice plot, a very intriguing trend was observed. For the zero-inclusive data, it appeared species abundance was increasing over time. However, when the zeros were removed, the trend was reversed, and abundance decreased over time (Appendix

G). For the conjunctive categorical plots, the same two abundance trends were primarily observed (Appendix H). However, in the non-zero lattices, a select few subcategories remained increasing or were stable over time instead of decreasing as expected. In total six plots contained this trend, four of which were “SCIENTIFIC_FAMILY” subcategories. This included three plots that remained increasing for the families Carrangidae, Gobiidae, and Haemulidae, and a constant plot for Pomacentridae. The other two plots, which were constant in the non-zero data, included Coral, and Sunset House of the “DIET”, and “LOCATION” variables respectively. Overall, these observations indicate certain families and diets are thriving with the loss of densities of other families, which is an important insight for the ecological goals of this report.

Mathematical Tendencies:

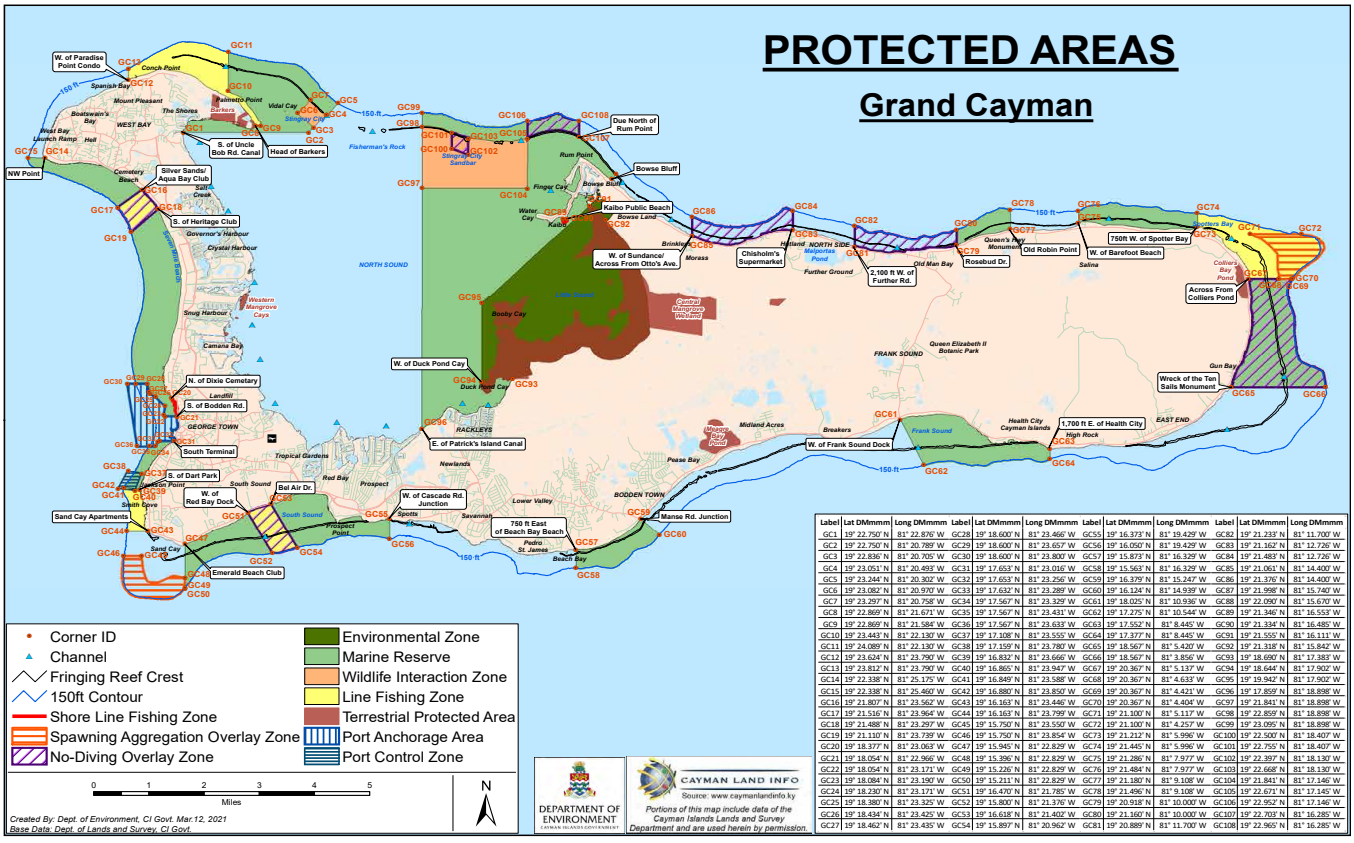
Mathematical tendencies are single values that describe key aspects of an entire set of observations. In data science, the focal tendencies are mean, median, mode, standard error, and standard deviation, which are typically portrayed using summary tables. Therefore, in the Grand Cayman data, summary tables were generated for the entire dataset, and for each categorical variable of interest (Appendix I & J). In these tables, the primary tendencies of interest were the means and standard deviations. The mean values provided an overall average “DENSITY_INDEX”, which were used to assess how subcategories varied and to find a density expectation for the entire dataset. Regarding standard deviation, these values indicate, on average, how far observations vary from the mean, which provides a measure variance in the data.

Typically, lower variability is preferred when making predictions and interpretations. Put simply, lower standard errors indicate that observations are very similar to their mean, which makes drawing conclusions from them more accurate. Interestingly, almost all standard deviations decreased from the zero-inclusive to non-zero summary tables. This indicates the zero-inclusive data has more variability than the non-zero data, which means our findings will be more accurate for the non-zero data. Thus, this supports using the non-zero data when modeling for primary ecological interpretations.

Appendix C

Cayman Island Government Department of Environment Marine Park Map

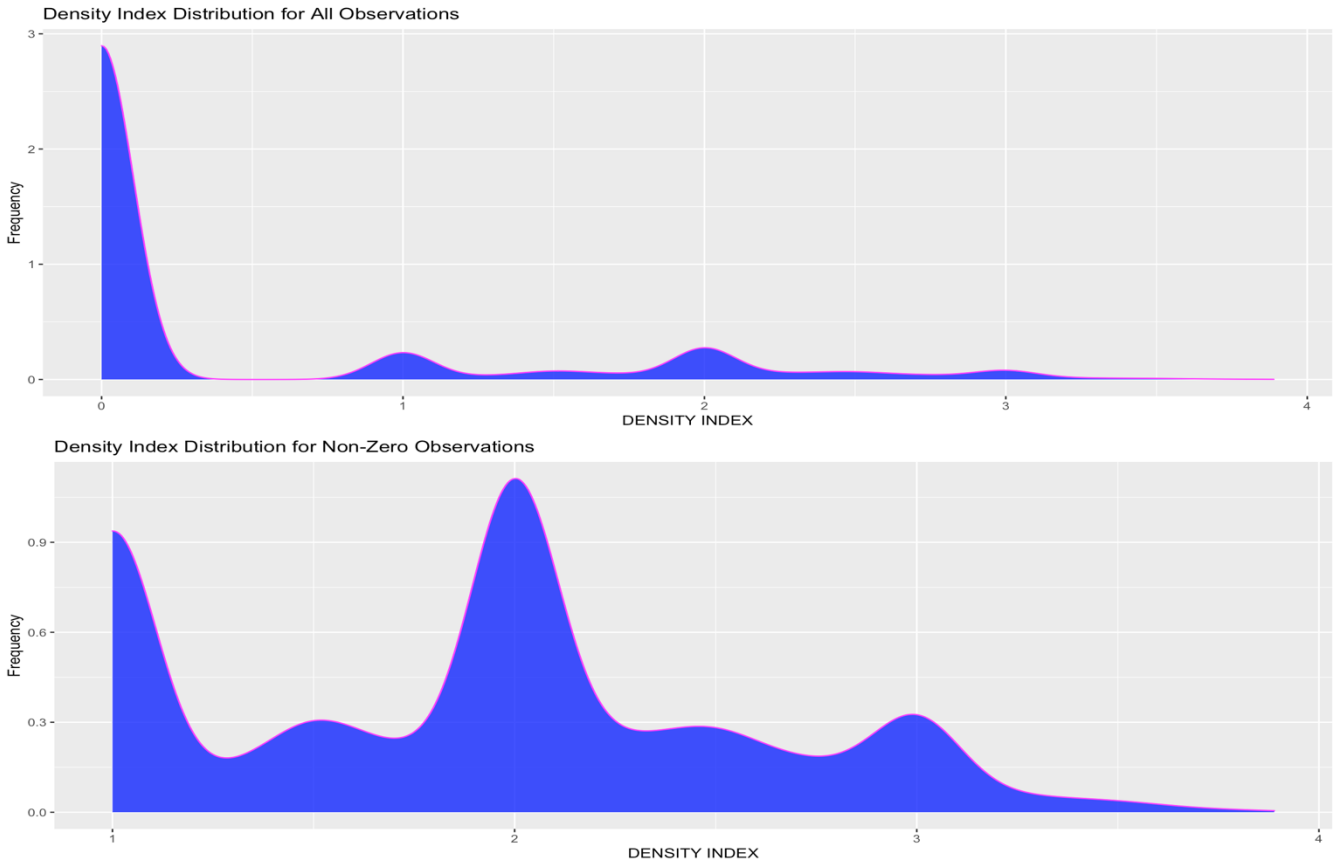
The following map was produced in 2021 and was used to analyze whether a dive site was contained within a marine reserve area or not.



Label	Lat DMmmm	Long DMmmm	Label	Lat DMmmm	Long DMmmm	Label	Lat DMmmm	Long DMmmm	Label	Lat DMmmm	Long DMmmm
GC1	19° 22.750' N	81° 22.876' W	GC28	19° 18.600' N	81° 23.466' W	GC55	19° 16.373' N	81° 19.429' W	GC82	19° 21.233' N	81° 11.700' W
GC2	19° 22.750' N	81° 20.789' W	GC29	19° 18.600' N	81° 23.657' W	GC56	19° 16.050' N	81° 19.429' W	GC83	19° 21.162' N	81° 12.726' W
GC3	19° 22.836' N	81° 20.705' W	GC30	19° 18.600' N	81° 23.800' W	GC57	19° 15.873' N	81° 16.329' W	GC84	19° 21.483' N	81° 12.726' W
GC4	19° 23.051' N	81° 20.493' W	GC31	19° 17.653' N	81° 23.016' W	GC58	19° 15.563' N	81° 16.329' W	GC85	19° 21.061' N	81° 14.400' W
GC5	19° 23.244' N	81° 20.302' W	GC32	19° 17.653' N	81° 23.256' W	GC59	19° 16.379' N	81° 15.247' W	GC86	19° 21.376' N	81° 14.400' W
GC6	19° 23.082' N	81° 20.970' W	GC33	19° 17.632' N	81° 23.289' W	GC60	19° 16.124' N	81° 14.938' W	GC87	19° 21.998' N	81° 15.740' W
GC7	19° 23.297' N	81° 20.758' W	GC34	19° 17.567' N	81° 23.329' W	GC61	19° 18.025' N	81° 10.936' W	GC88	19° 22.090' N	81° 15.670' W
GC8	19° 22.869' N	81° 21.671' W	GC35	19° 17.567' N	81° 23.431' W	GC62	19° 17.275' N	81° 10.544' W	GC89	19° 21.346' N	81° 16.553' W
GC9	19° 22.869' N	81° 21.584' W	GC36	19° 17.567' N	81° 23.633' W	GC63	19° 17.552' N	81° 8.445' W	GC90	19° 21.334' N	81° 16.485' W
GC10	19° 23.443' N	81° 22.130' W	GC37	19° 17.108' N	81° 23.555' W	GC64	19° 17.377' N	81° 8.445' W	GC91	19° 21.555' N	81° 16.111' W
GC11	19° 24.089' N	81° 22.130' W	GC38	19° 17.159' N	81° 23.780' W	GC65	19° 18.567' N	81° 5.420' W	GC92	19° 21.318' N	81° 15.842' W
GC12	19° 23.624' N	81° 23.790' W	GC39	19° 16.832' N	81° 23.666' W	GC66	19° 18.567' N	81° 3.856' W	GC93	19° 18.690' N	81° 17.383' W
GC13	19° 23.812' N	81° 23.790' W	GC40	19° 16.865' N	81° 23.947' W	GC67	19° 20.367' N	81° 5.137' W	GC94	19° 18.644' N	81° 17.902' W
GC14	19° 22.338' N	81° 25.175' W	GC41	19° 16.849' N	81° 23.588' W	GC68	19° 20.367' N	81° 4.633' W	GC95	19° 19.942' N	81° 17.902' W
GC15	19° 22.338' N	81° 25.460' W	GC42	19° 16.880' N	81° 23.850' W	GC69	19° 20.367' N	81° 4.421' W	GC96	19° 17.859' N	81° 18.898' W
GC16	19° 21.807' N	81° 23.562' W	GC43	19° 16.163' N	81° 23.440' W	GC70	19° 20.367' N	81° 4.404' W	GC97	19° 21.841' N	81° 18.898' W
GC17	19° 21.516' N	81° 23.964' W	GC44	19° 16.163' N	81° 23.799' W	GC71	19° 21.100' N	81° 5.117' W	GC98	19° 22.859' N	81° 18.898' W
GC18	19° 21.488' N	81° 23.297' W	GC45	19° 15.750' N	81° 23.550' W	GC72	19° 21.100' N	81° 4.257' W	GC99	19° 23.095' N	81° 18.898' W
GC19	19° 21.110' N	81° 23.739' W	GC46	19° 15.750' N	81° 23.854' W	GC73	19° 21.212' N	81° 5.996' W	GC100	19° 22.500' N	81° 18.407' W
GC20	19° 18.377' N	81° 23.063' W	GC47	19° 15.945' N	81° 22.829' W	GC74	19° 21.445' N	81° 5.996' W	GC101	19° 22.755' N	81° 18.407' W
GC21	19° 18.054' N	81° 22.966' W	GC48	19° 15.396' N	81° 22.829' W	GC75	19° 21.286' N	81° 7.977' W	GC102	19° 22.397' N	81° 18.130' W
GC22	19° 18.054' N	81° 23.171' W	GC49	19° 15.226' N	81° 22.829' W	GC76	19° 21.484' N	81° 7.977' W	GC103	19° 22.668' N	81° 18.130' W
GC23	19° 18.084' N	81° 23.190' W	GC50	19° 15.211' N	81° 22.829' W	GC77	19° 21.180' N	81° 9.108' W	GC104	19° 21.841' N	81° 17.146' W
GC24	19° 18.230' N	81° 23.171' W	GC51	19° 16.470' N	81° 21.785' W	GC78	19° 21.496' N	81° 9.108' W	GC105	19° 22.671' N	81° 17.145' W
GC25	19° 18.380' N	81° 23.325' W	GC52	19° 15.800' N	81° 21.376' W	GC79	19° 20.918' N	81° 10.007' W	GC106	19° 22.952' N	81° 17.146' W
GC26	19° 18.434' N	81° 23.425' W	GC53	19° 16.618' N	81° 21.402' W	GC80	19° 21.160' N	81° 10.007' W	GC107	19° 22.703' N	81° 16.285' W
GC27	19° 18.462' N	81° 23.435' W	GC54	19° 15.897' N	81° 20.962' W	GC81	19° 20.889' N	81° 11.700' W	GC108	19° 22.965' N	81° 16.285' W

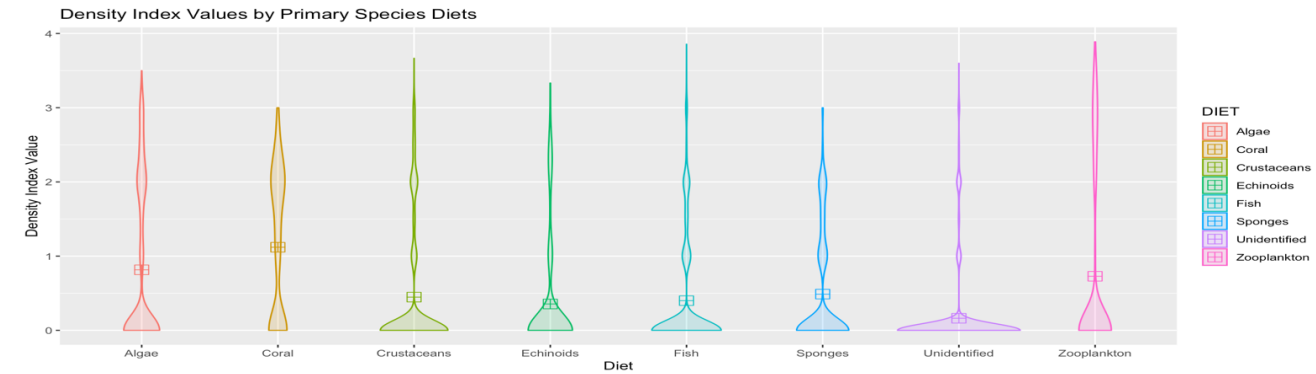
Appendix D
Density Plots for the Response Variable "DENSITY_INDEX"

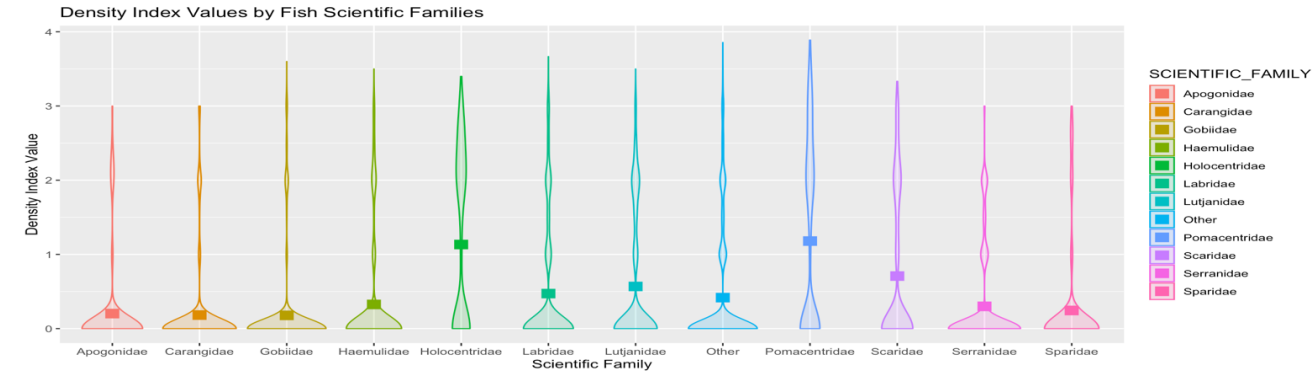
The complete and non-zero density distributions for the abundance measurement response variable density index. Heavy zero-inflation is observed in the complete distribution with minor Gaussian-like truncations centered around whole number integer responses. A more Gaussian-like shape is observed in the non-zero plot, but truncations remain at the extrema of the distribution.

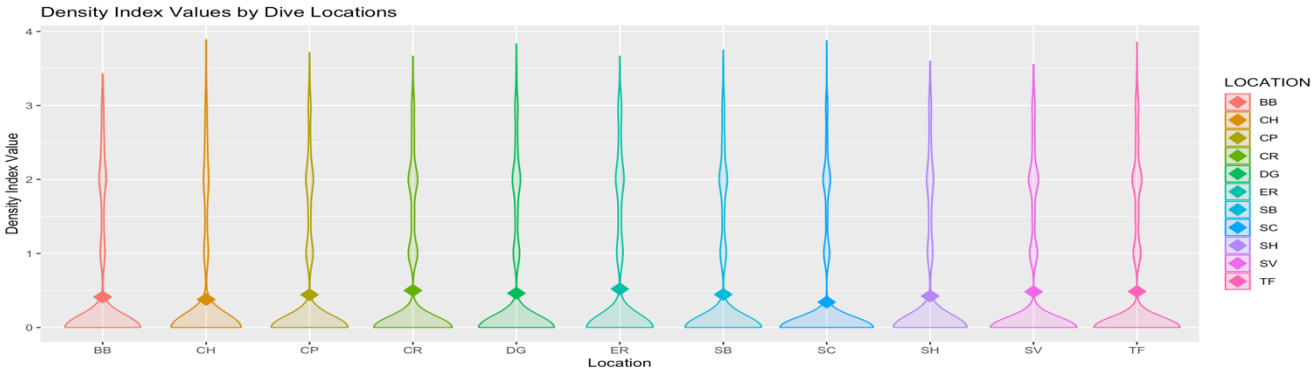


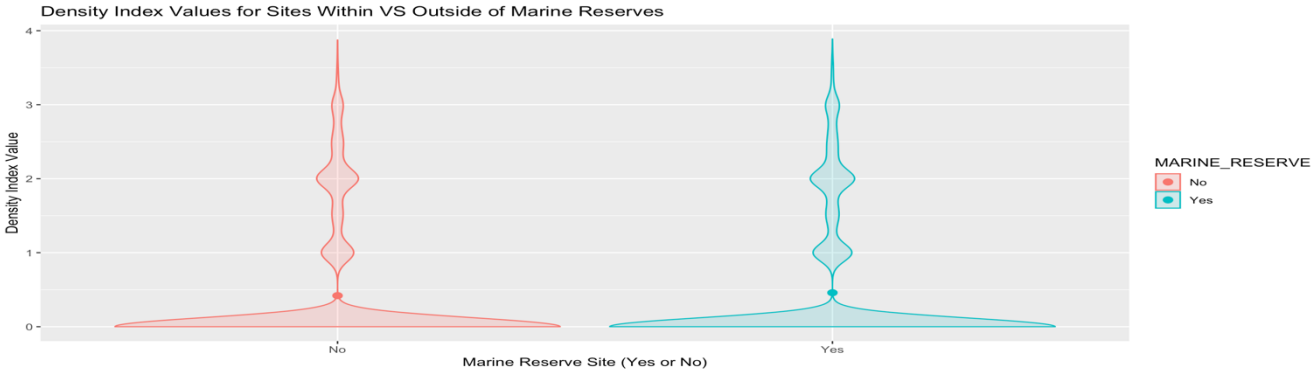
Appendix E
Zero-Inclusive Violin Plots of "DIET", "SCIENTIFIC_FAMILY", "LOCATION", and "MARINE_RESERVE"

Variation in observation density and subcategory means is observed by violin plots. Means are represented by shapes with colors denoting subcategories.





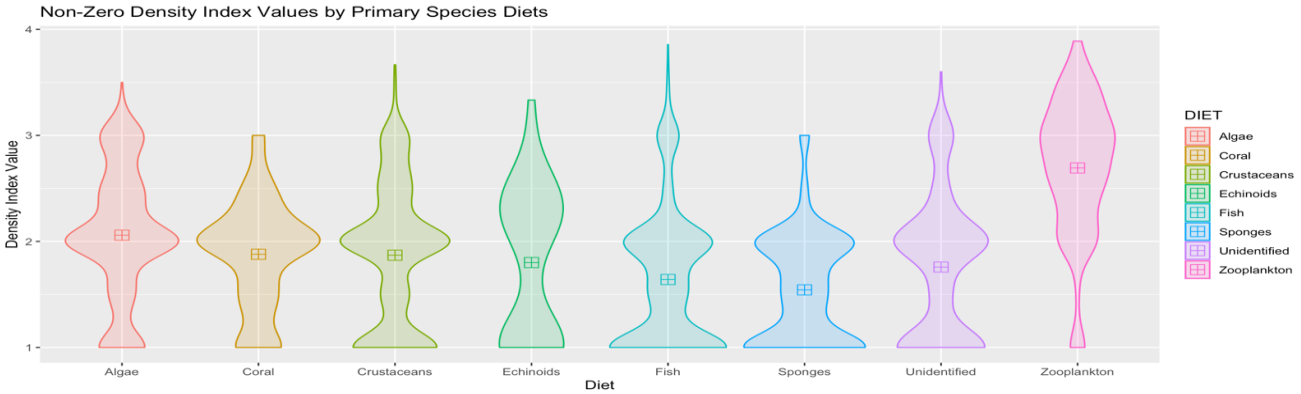


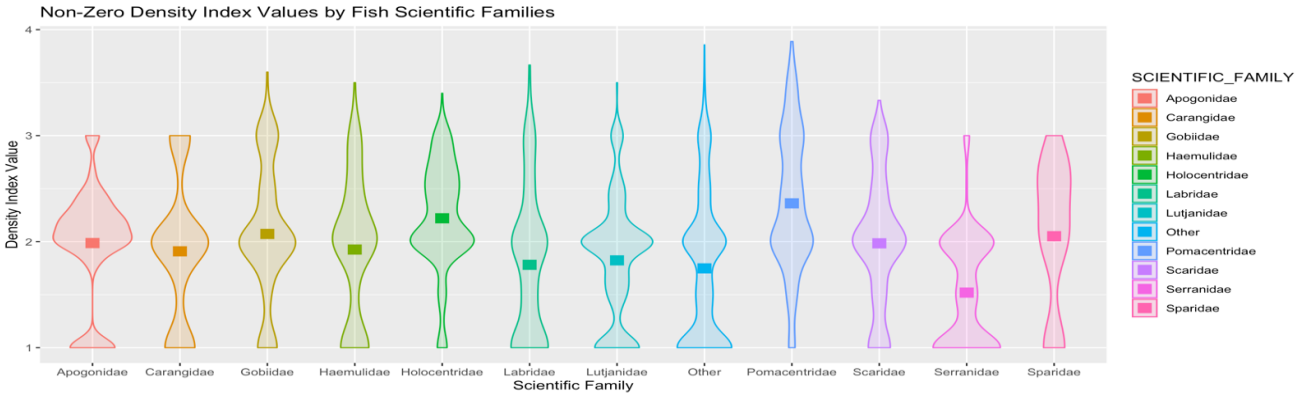


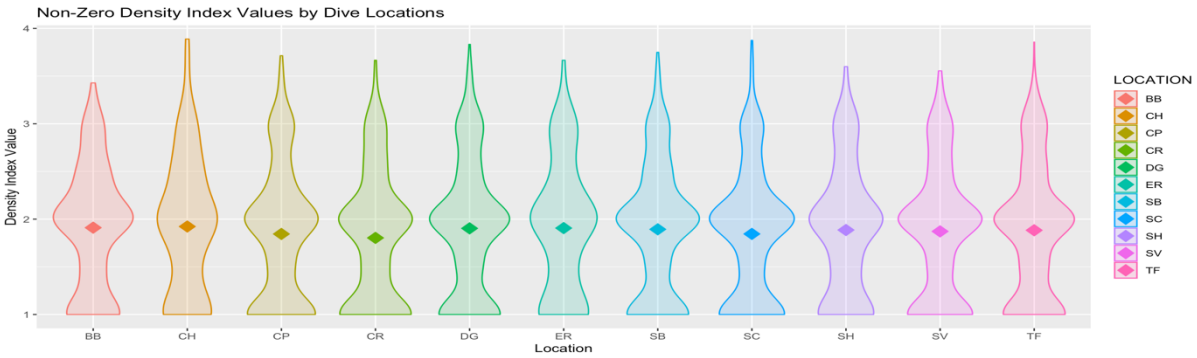
Appendix F

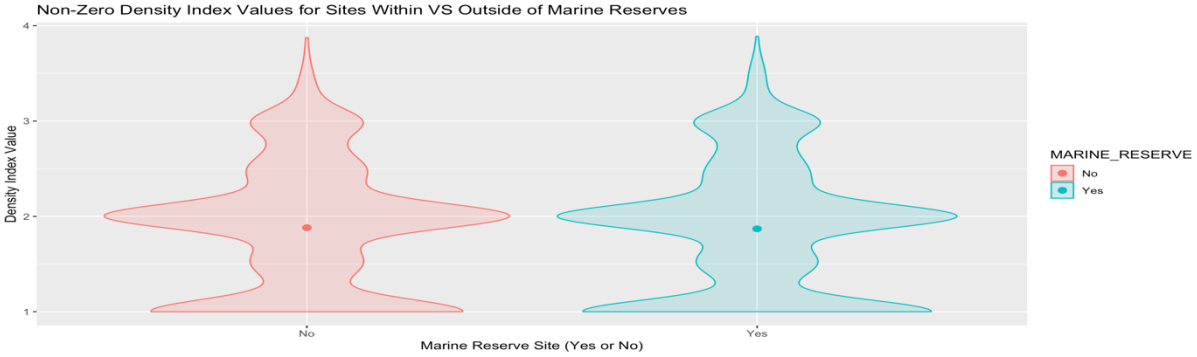
Non-Zero Violin Plots of "DIET", "SCIENTIFIC_FAMILY", "LOCATION", and "MARINE_RESERVE"

Variation in observation density and subcategory means is observed by violin plots. Means are represented by shapes with colors denoting subcategories





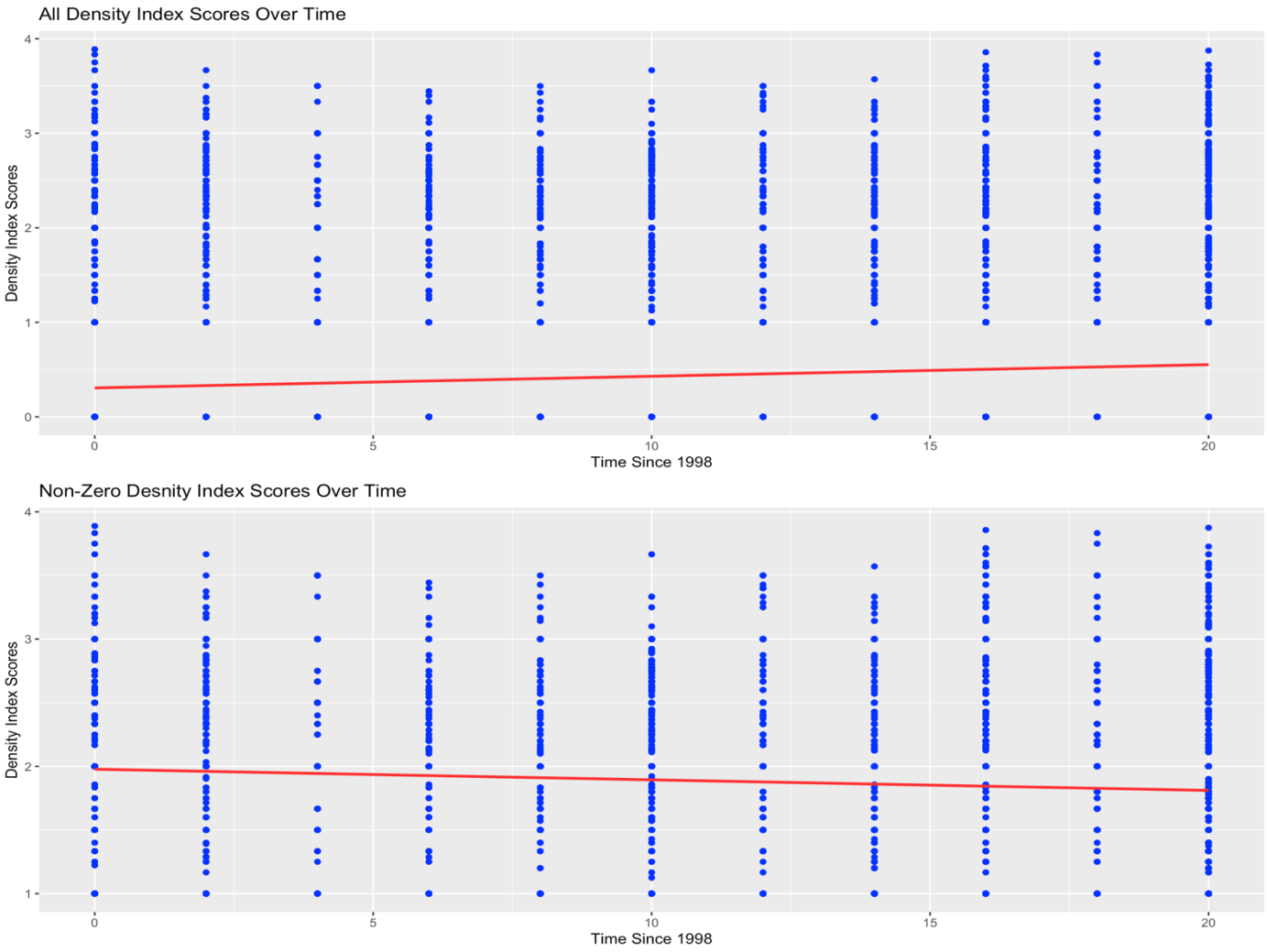




Appendix G

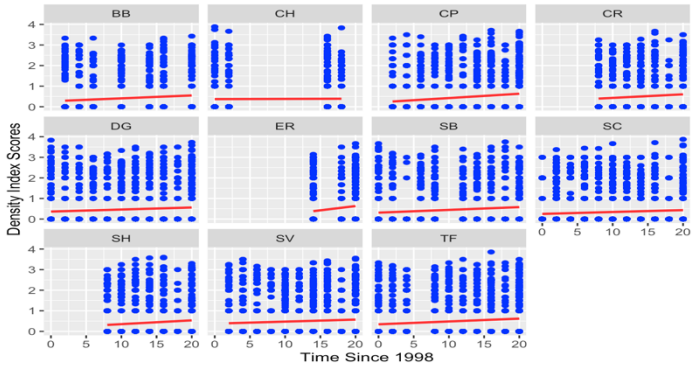
Zero-Inclusive and Non-Zero Lattice Plots of "DENSITY_INDEX" by "YEAR"

Individual datapoints are denoted in blue with the red linear line of best fit showing change over time. Linear line was fit using `stat_smooth()` command.

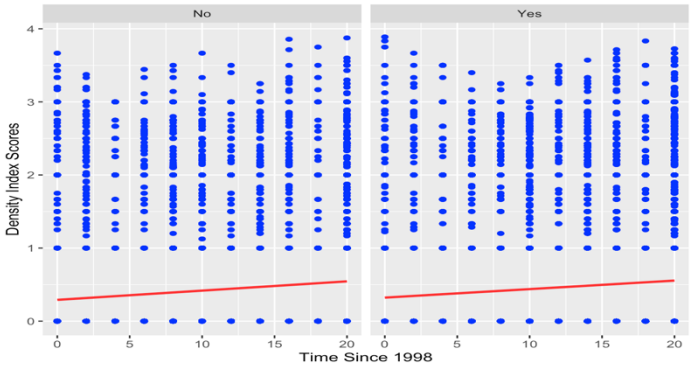


Appendix H
Zero-Inclusive and Non-Zero Lattice Plots in Conjuncture with Categorical Variables

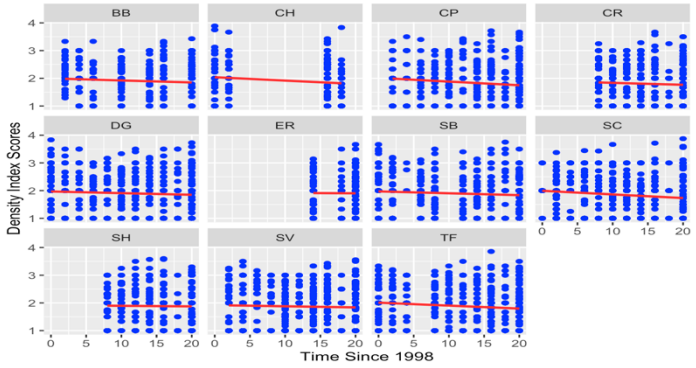
All Density Index Scores by Location



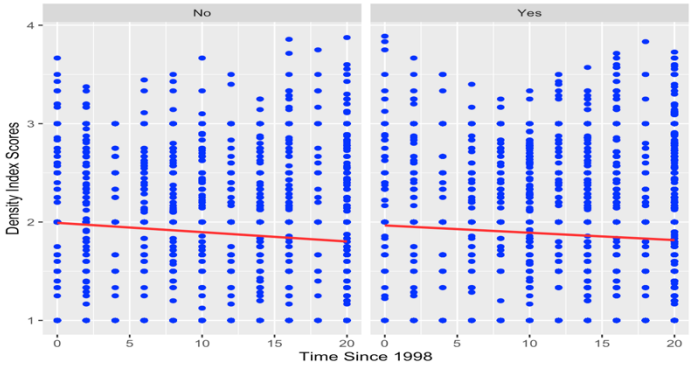
Non-Zero Density Index Scores for Marine Reserve Vs. Non-Reserve



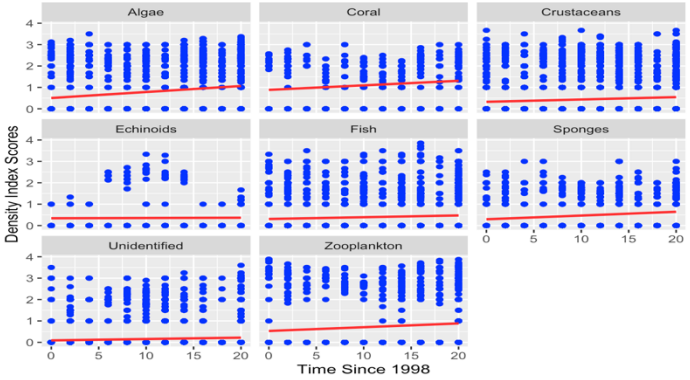
Non-Zero Density Index Scores by Location



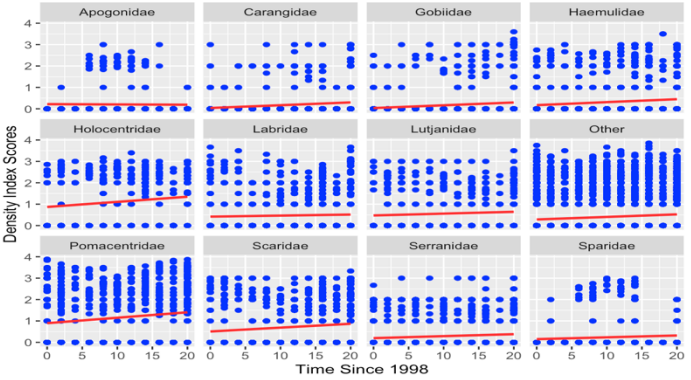
All Density Index Scores for Marine Reserve Vs. Non-Reserve Sites



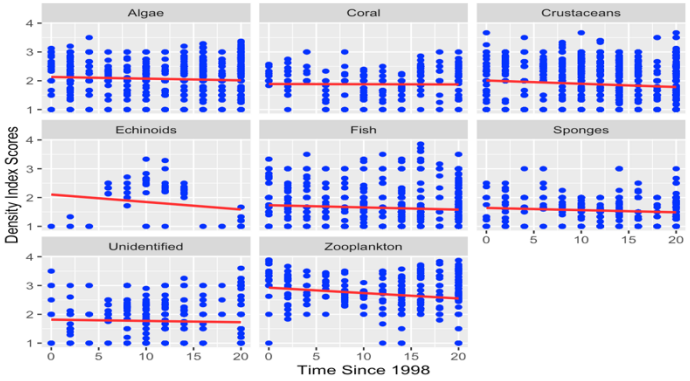
All Density Index Scores by Diet



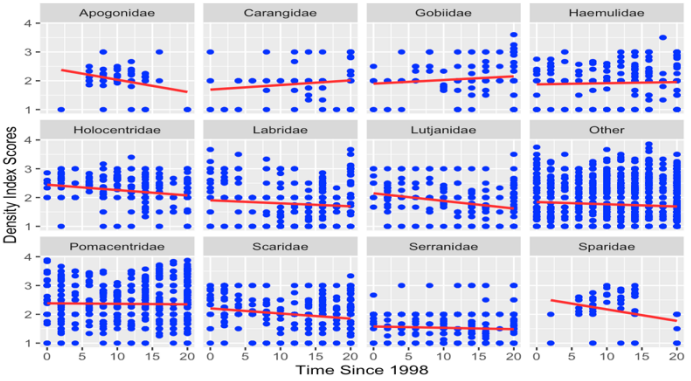
All Density Index Scores by Scientific Family



Non-Zero Density Index Scores by Diet



Non-Zero Density Index Scores by Scientific Family



Appendix I
Zero-Inclusive Summary Statistics by "DENSITY_INDEX"

Included are summaries for the overall zero-inclusive data with categorical variables "SCIENTIFIC_FAMILY" and "DIET".

1. Overall Zero-Inclusive Data Summary

Overall Summary Statistics of Density Index on Grand Cayman Island

Mean Density Index	Standard Deviation	Standard Error	Sample Size
0.4426821	0.8623694	0.0058072	22052

2. Zero-Inclusive "SCIENTIFIC_FAMILY"

Overall Summary Statistics Across Scientific Families on Grand Cayman Island

Family	Mean Density Index	Standard Deviation	Standard Error	Sample Size
Apogonidae	0.2023553	0.6248673	0.0268900	540
Carangidae	0.1822212	0.5963663	0.0181553	1079
Gobiidae	0.1782912	0.6100019	0.0151744	1616
Haemulidae	0.3256079	0.7707249	0.0203316	1437
Holocentridae	1.1331251	1.1666672	0.0505340	533
Labridae	0.4719138	0.8630698	0.0235422	1344
Lutjanidae	0.5678172	0.9033246	0.0301780	896
Other	0.4173957	0.8162026	0.0086644	8874
Pomacentridae	1.1796572	1.2673786	0.0360057	1239
Scaridae	0.7079198	1.0171146	0.0297866	1166
Serranidae	0.2995883	0.6482272	0.0124798	2698
Sparidae	0.2442227	0.7033578	0.0280224	630

3. Zero-Inclusive "DIET"

Overall Summary Statistics for Species Primary Diets on Grand Cayman Island

Diet	Mean Density Index	Standard Deviation	Standard Error	Sample Size
Algae	0.8157071	1.0791521	0.0211761	2597
Coral	1.1197525	1.0038735	0.0474286	448
Crustaceans	0.4475921	0.8598790	0.0108516	6279
Echinoids	0.3550889	0.7826984	0.0412518	360
Fish	0.3999007	0.7719094	0.0117715	4300
Sponges	0.4879182	0.7743653	0.0204134	1439
Unidentified	0.1628468	0.5479348	0.0074654	5387
Zooplankton	0.7284863	1.2434362	0.0352828	1242

Appendix J
Non-Zero Summary Statistics by "DENSITY_INDEX"

Included are summaries for the overall non-zero data with categorical variables "SCIENTIFIC_FAMILY" and "DIET".

1. Overall Non-Zero Data Summary

Non-Zero Summary Statistics of Density Index on Grand Cayman Island

Mean Density Index	Standard Deviation	Standard Error	Sample Size
1.874429	0.6820364	0.0094509	5208

2. Non-Zero "SCIENTIFIC_FAMILY"

Family	Mean Density Index	Standard Deviation	Standard Error	Sample Size
Apogonidae	1.986762	0.5352753	0.0721765	55
Carangidae	1.908900	0.6560420	0.0646417	103
Gobiidae	2.072795	0.6319028	0.0535973	139
Haemulidae	1.925508	0.6568684	0.0421381	243
Holocentridae	2.220425	0.4988204	0.0302454	272
Labridae	1.781607	0.6914175	0.0366451	356
Lutjanidae	1.823528	0.5735640	0.0343384	279
Other	1.747980	0.6811335	0.0147968	2119
Pomacentridae	2.361220	0.6505770	0.0261489	619
Scaridae	1.984217	0.6046368	0.0296448	416
Serranidae	1.519341	0.5268009	0.0228397	532
Sparidae	2.051471	0.6689056	0.0772386	75

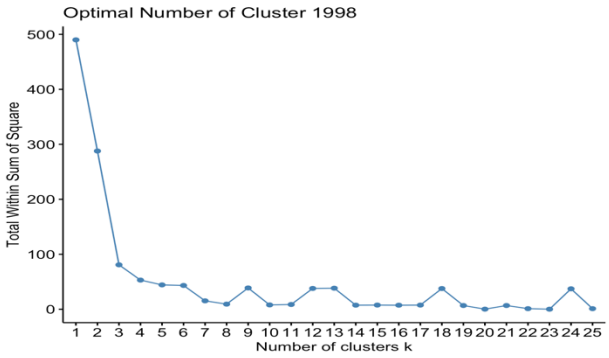
3. Non-Zero "DIET"

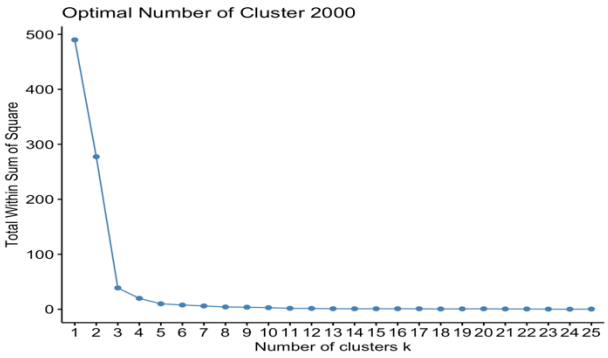
Non-Zero Summary Statistics for Species Primary Diets on Grand Cayman Island

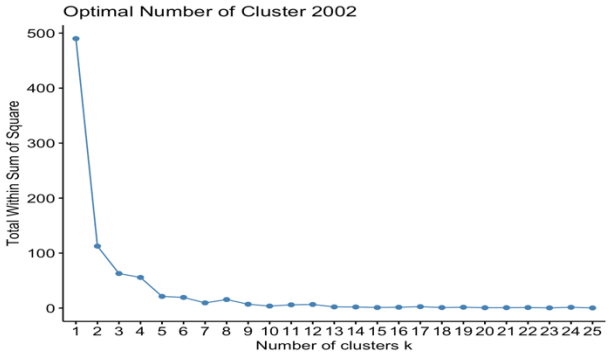
Diet	Mean Density Index	Standard Deviation	Standard Error	Sample Size
Algae	2.058689	0.6160146	0.0192036	1029
Coral	1.878836	0.5118014	0.0313217	267
Crustaceans	1.869881	0.6551331	0.0168986	1503
Echinoids	1.800451	0.7087813	0.0841169	71
Fish	1.640814	0.6391348	0.0197429	1048
Sponges	1.543108	0.5171518	0.0242445	455
Unidentified	1.758027	0.6611286	0.0295962	499
Zooplankton	2.692798	0.6498730	0.0354535	336

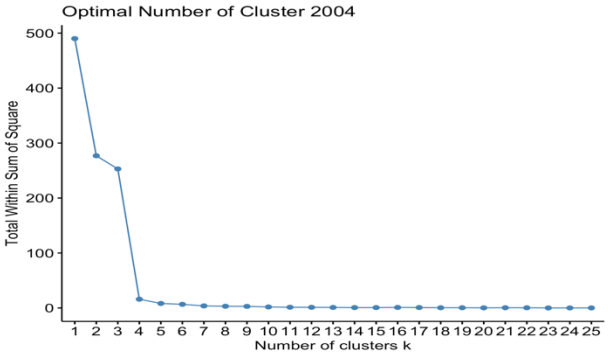
Appendix K
Optimal K Choice Plots for K-Means Clustering

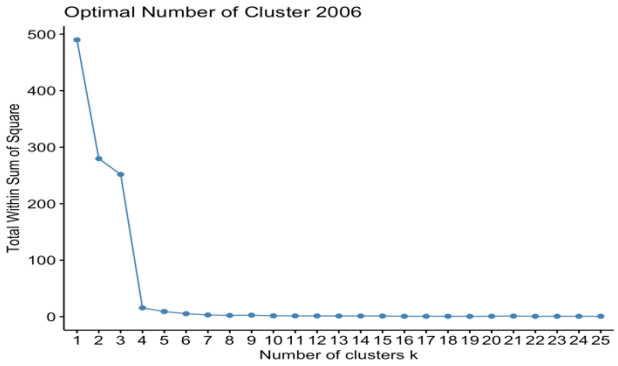
Optimization of the number of clusters, K, as determined by drop in total sum of squares is visualized in `fviz_nbclust()` plots. Optimization occurs at the elbow-like bends in the plots, which made K=4 a reasonable choice for clustering across all years.

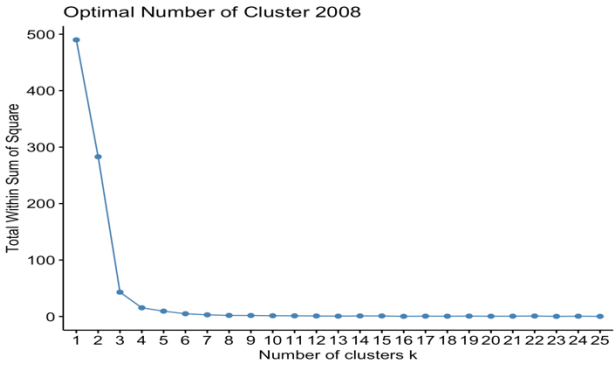


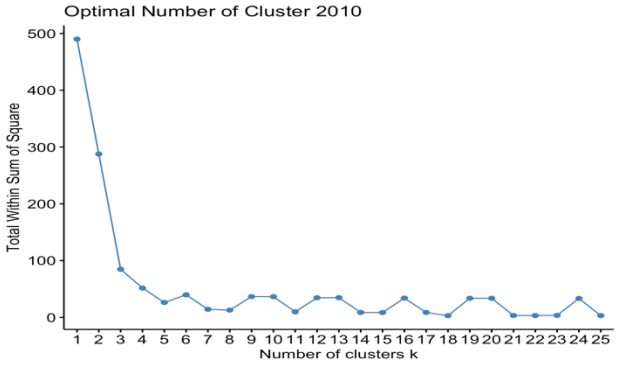


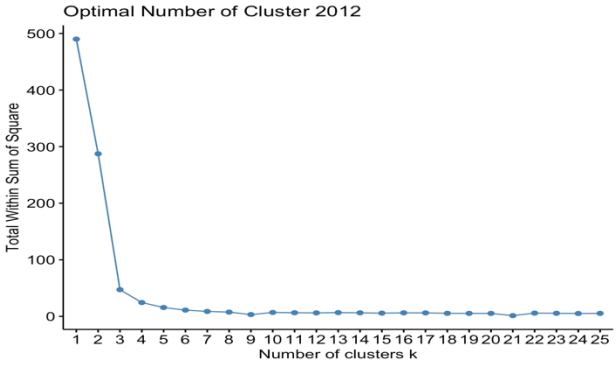






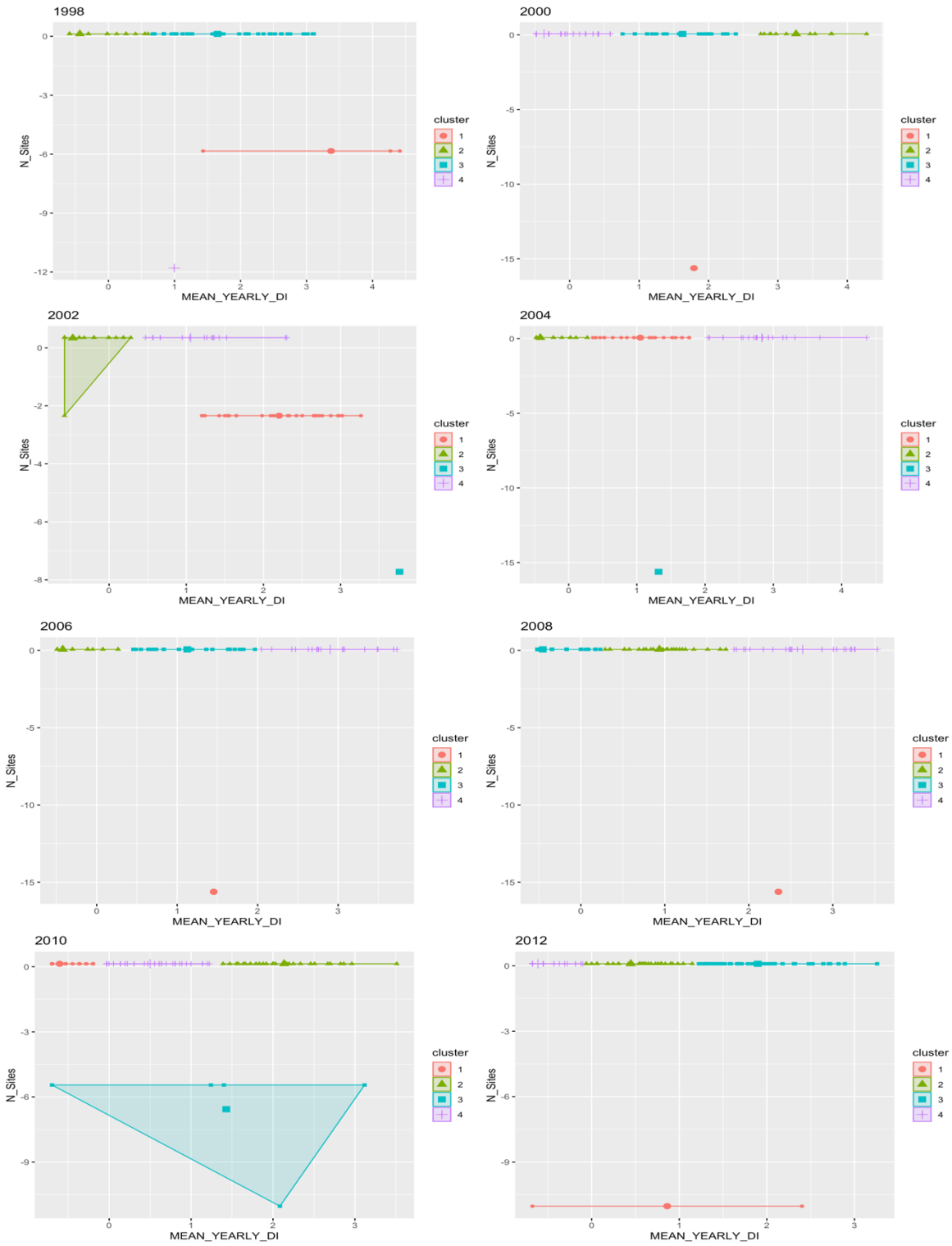






Appendix L
K-Means Clusters Visualization

Groupings of K-Means Clustering outputs from 1998 to 2012 in the Grand Cayman dataset are visualized using the `fviz_cluster` function. Similarity was based on scaled “DENSITY_INDEX” and “N_SITES”. Centers are depicted by bolded shapes with slightly larger size.



Appendix M
K-Means Cluster Centers

Unscaled values of “DENSITY_INDEX” and “N_SITES” are followed by their corresponding scaled values in the plots below. The results aid in visualizing the means of centers from Appendix K

1998 Centers

DI_CENTERS <dbl>	SITES_CENTERS <dbl>	MEAN_YEARLY_DI <dbl>	N_Sites <dbl>
2.7750663	4	3.3708185	-5.838370
0.1087986	5	-0.4338479	0.121128
1.5708361	5	1.6524260	0.121128
1.1111112	3	0.9964154	-11.797868

2000 Centers

DI_CENTERS <dbl>	SITES_CENTERS <dbl>	MEAN_YEARLY_DI <dbl>	N_Sites <dbl>
1.60000006	5	1.7907923	-15.62062947
2.63654775	8	3.2617558	0.06375767
1.48424325	8	1.6265220	0.06375767
0.07641189	8	-0.3713298	0.06375767

2002 Centers

DI_CENTERS <dbl>	SITES_CENTERS <dbl>	MEAN_YEARLY_DI <dbl>	N_Sites <dbl>
2.08224209	6.000000	2.2038192	-2.3407942
0.08018755	6.994819	-0.4705805	0.3360832
3.25000003	4.000000	3.7637425	-7.7224331
1.22321432	7.000000	1.0563062	0.3500253

2004 Centers

DI_CENTERS <dbl>	SITES_CENTERS <dbl>	MEAN_YEARLY_DI <dbl>	N_Sites <dbl>
1.00245042	7	1.0491517	0.06375767
0.03916642	7	-0.4151812	0.06375767
1.17934056	6	1.3180506	-15.62062947
2.17511388	7	2.8317720	0.06375767

Appendix N
Summary of Levene Test Results

Summaries of Levene Test p-values with resulting hypothesis test interpretations are depicted for both the zero-inclusive and non-zero containing Grand Cayman Island Data.

Summary of Levene Test Results

Test Variable	Resulting P-Value	Interpretation
Diet	2.2e-16	Unequal Variance
Family	2.2e-16	Unequal Variance
Location	2.151e-11	Unequal Variance
Reserve	0.0008255	Unequal Variance

Non-Zero Data Summary of Levene Test Results

Test Variable	Resulting P-Value	Interpretation
Diet	8.094e-12	Unequal Variance
Family	2.2e-16	Unequal Variance
Location	0.6516	Equal Variance
Reserve	0.0636	Equal Variance

Appendix O
Summary of ANOVA and T-Test Results

Summaries of ANOVA's and T-Test's with p-values and resulting hypothesis test interpretations are depicted for both the zero-inclusive and non-zero containing Grand Cayman Island Data. Welch's versions were used for variables with unequal variance determined by the Levene Test.

Summary of ANOVA and T-Test Results

Test Variable	Test Type	Resulting P-Value	Interpretation
Diet	Welch ANOVA	2.2e-16	True Mean Difference Not 0
Family	Welch ANOVA	2.2e-16	True Mean Difference Not 0
Location	Welch ANOVA	1.144e-12	True Mean Difference Not 0
Reserve	Welch T-Test	0.0007742	True Mean Difference Not 0

Non-Zero Data Summary of ANOVA and T-Test Results

Test Variable	Test Type	Resulting P-Value	Interpretation
Diet	Welch ANOVA	2.2e-16	True Mean Difference Not 0
Family	Welch ANOVA	2.2e-16	True Mean Difference Not 0
Location	Standard ANOVA	0.317	True Mean Difference = 0
Reserve	Equal Variance T-Test	0.5413	True Mean Difference = 0

Appendix P

Model Outputs of Unconditional Mean and Unconditional Growth Models

Output for the Unconditional Means Model (Top) shows random effects estimates. Variability between versus within was assessed from this model by calculating the ICC. Unconditional Growth Model (Bottom) output shows random effects outputs with the addition of "YEAR1998" as a fixed effect. The significant coefficient estimate for this model was used to assess the overall change in "DENSITY_INDEX" over time for all species.

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: DENSITY_INDEX ~ 1 + (1 | SPECIES_NAME)
   Data: CAYMAN_NO_ZEROS

REML criterion at convergence: 8436.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6607 -0.6235 -0.0347  0.6186  3.8503

Random effects:
 Groups       Name        Variance Std.Dev.
 SPECIES_NAME (Intercept) 0.1947   0.4413
 Residual                 0.2653   0.5151
Number of obs: 5208, groups:  SPECIES_NAME, 231

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   1.73686    0.03138 227.75030   55.35   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: DENSITY_INDEX ~ YEAR1998 + (YEAR1998 | SPECIES_NAME)
   Data: CAYMAN_NO_ZEROS

REML criterion at convergence: 8358.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.3176 -0.6167 -0.0359  0.6183  4.0456

Random effects:
 Groups       Name        Variance  Std.Dev. Corr
 SPECIES_NAME (Intercept) 0.2584244 0.50835
              YEAR1998    0.0003597 0.01897  -0.52
 Residual                 0.2539652 0.50395
Number of obs: 5208, groups:  SPECIES_NAME, 231

Fixed effects:
             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)   1.838342   0.041660 176.497368  44.127  < 2e-16 ***
YEAR1998     -0.007947   0.002003 114.764712  -3.968 0.000127 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr)
YEAR1998 -0.669
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.0291389 (tol = 0.002, component 1)
```

Appendix Q
Maximum Likelihood and AIC Model Determinants

Results for determining the best fit and most interpretable model using Maximum Likelihood Ratio Tests (Chi-Square) and AIC criterion are shown. First image was for consideration of additive terms, second image was for consideration of interaction terms, and third image was for consideration of adding a random slope term for year to the model.

1.

```
## Single term additions
##
## Model:
## DENSITY_INDEX ~ YEAR1998 + (1 | SPECIES_NAME)
##                     Df    AIC    LRT  Pr(>Chi)
## <none>                 8413.0
## DIET                 7 8382.2 44.889 1.437e-07 ***
## SCIENTIFIC_FAMILY 11 8376.3 58.711 1.608e-08 ***
## LOCATION           10 8418.8 14.245    0.1621
## MARINE_RESERVE      1 8414.7  0.353    0.5525
```

2.

```
## Single term additions
##
## Model:
## DENSITY_INDEX ~ YEAR1998 + DIET + SCIENTIFIC_FAMILY + (1 | SPECIES_NAME)
##                            Df    AIC    LRT Pr(>Chi)
## <none>                        8369.7
## YEAR1998:DIET               7 8376.9  6.809    0.449
## YEAR1998:SCIENTIFIC_FAMILY 11 8345.4 46.336 2.82e-06 ***
```

3.

```
##          df      AIC
## M3_LMER 22 8426.150
## M6_LMER 24 8365.182
```

Appendix R
Model Output of Potential Interaction Model

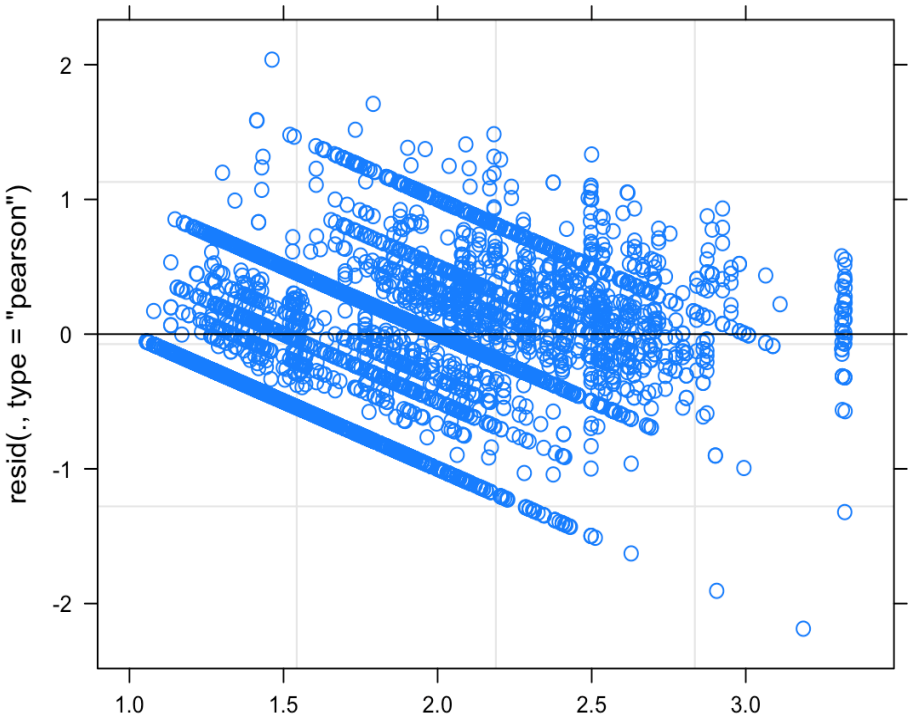
Output of the model testing for an interaction being “YEAR1998” and “SCIENTIFIC_FAMILY” is shown. Note all model coefficients are insignificant, which made drawing conclusions from this model difficult. Ultimately, the interaction was not included for interpretation, but should be used when modeling for prediction

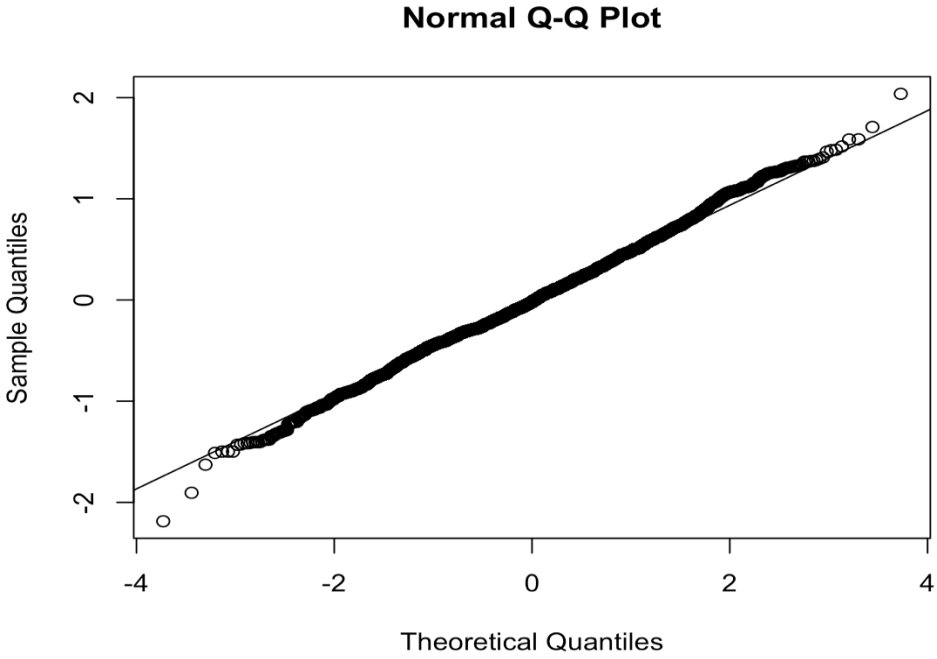
```
REML criterion at convergence: 8378.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2977 -0.6159 -0.0406  0.6274  4.0305

Random effects:
 Groups       Name        Variance  Std.Dev. Corr
 SPECIES_NAME (Intercept) 0.2100044 0.45826
              YEAR1998    0.0003372 0.01836  -0.61
 Residual                 0.2539256 0.50391
Number of obs: 5208, groups:  SPECIES_NAME, 231

Fixed effects:
                                        Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                            2.226e+00  3.759e-01  3.464e+02   5.923 7.61e-09
YEAR1998                              -3.213e-02  2.211e-02  4.197e+02  -1.453   0.1469
DIETCoral                             -5.515e-04  2.009e-01  1.551e+02  -0.003   0.9978
DIETCrustaceans                       -1.449e-01  1.353e-01  1.747e+02  -1.071   0.2855
DIETEchinoids                         -4.815e-01  2.448e-01  2.125e+02  -1.967   0.0505
DIETFish                              -2.970e-01  1.440e-01  1.775e+02  -2.062   0.0407
DIETSponges                           -3.517e-01  1.452e-01  1.683e+02  -2.423   0.0165
DIETUnidentified                      -2.865e-01  1.352e-01  1.883e+02  -2.119   0.0354
DIETZooplankton                        2.477e-01  1.528e-01  1.809e+02   1.621   0.1068
SCIENTIFIC_FAMILYCarangidae           -3.914e-01  4.363e-01  2.760e+02  -0.897   0.3704
SCIENTIFIC_FAMILYGobiidae             -2.085e-01  4.099e-01  2.883e+02  -0.509   0.6114
SCIENTIFIC_FAMILYHaemulidae           -4.311e-01  4.061e-01  2.688e+02  -1.061   0.2895
SCIENTIFIC_FAMILYHolocentridae         2.289e-01  4.200e-01  2.385e+02   0.545   0.5863
SCIENTIFIC_FAMILYLabridae             -2.269e-01  3.889e-01  2.810e+02  -0.583   0.5601
SCIENTIFIC_FAMILYLutjanidae            1.703e-01  4.102e-01  2.662e+02   0.415   0.6783
SCIENTIFIC_FAMILYOther                -2.930e-01  3.642e-01  3.062e+02  -0.805   0.4217
SCIENTIFIC_FAMILYPomacentridae         1.712e-01  3.917e-01  2.917e+02   0.437   0.6625
SCIENTIFIC_FAMILYScaridae             -1.090e-01  4.047e-01  2.850e+02  -0.269   0.7880
SCIENTIFIC_FAMILYSerranidae           -3.651e-01  3.807e-01  2.950e+02  -0.959   0.3383
SCIENTIFIC_FAMILYSparidae             -2.616e-01  4.835e-01  2.553e+02  -0.541   0.5889
YEAR1998:SCIENTIFIC_FAMILYCarangidae   4.656e-02  2.499e-02  3.206e+02   1.863   0.0634
YEAR1998:SCIENTIFIC_FAMILYGobiidae     4.453e-02  2.479e-02  3.941e+02   1.796   0.0732
YEAR1998:SCIENTIFIC_FAMILYHaemulidae   3.996e-02  2.385e-02  3.313e+02   1.676   0.0948
YEAR1998:SCIENTIFIC_FAMILYHolocentridae 2.017e-02  2.406e-02  2.843e+02   0.838   0.4025
YEAR1998:SCIENTIFIC_FAMILYLabridae     2.306e-02  2.329e-02  3.463e+02   0.990   0.3228
YEAR1998:SCIENTIFIC_FAMILYLutjanidae   1.517e-03  2.375e-02  3.097e+02   0.064   0.9491
YEAR1998:SCIENTIFIC_FAMILYOther        2.533e-02  2.232e-02  4.046e+02   1.135   0.2571
YEAR1998:SCIENTIFIC_FAMILYPomacentridae 2.543e-02  2.296e-02  3.390e+02   1.107   0.2689
YEAR1998:SCIENTIFIC_FAMILYScaridae     1.285e-02  2.316e-02  3.301e+02   0.555   0.5794
YEAR1998:SCIENTIFIC_FAMILYSerranidae   2.222e-02  2.285e-02  3.768e+02   0.973   0.3314
YEAR1998:SCIENTIFIC_FAMILYSparidae     2.245e-02  2.744e-02  3.015e+02   0.818   0.4140
```

Appendix S

Final Interpretation Model Diagnostic Plots

Residuals versus fitted and normal qq-plot diagnostic graphs are pictured based on the final interpretation model. Residuals versus fitted displays slight violation of linearity, while qq-plot does not indicate any violations of normality in the model. Straight lines in residual versus fitted plot result from structure of the data. Most observations are discrete integers giving this result.





Appendix T
Cluster Centers Over Time and For Specific Species

Specifications for each of the K = 4 "DENSITY_INDEX" centers of the K-Means Clustering models are depicted for every year of data collection. Additionally, centers and cluster specifications are depicted for the species Lionfish, Rainbow Parrotfish, and Ocean Surgeonfish.

K-Means Density Index Cluster Centers From 1998 to 2008 for Grand Cayman Data

	1998 Centers	2000 Centers	2002 Centers	2004 Centers	2006 Centers	2008 Centers
1	2.7750663	2.6365477	3.2500000	2.1751139	2.2421548	2.2899310
3	1.5708361	1.6000001	2.0822421	1.1793406	1.2857142	2.0802720
4	1.1111112	1.4842433	1.2232143	1.0024504	1.0686190	1.0520447
2	0.1087986	0.0764119	0.0801875	0.0391664	0.0456053	0.0446581

K-Means Density Index Cluster Centers From 2010 to 2018 for Grand Cayman Data

	2010 Centers	2012 Centers	2014 Centers	2016 Centers	2018 Centers
2	2.1001345	2.0912047	2.2011562	2.0423332	2.1440958
3	1.5784578	1.2509260	0.8624584	1.4839505	1.7745536
4	0.8914290	0.9143393	0.0654832	0.8553333	0.9258102
1	0.0759519	0.0518004	0.0000000	0.0259390	0.1493352

Density Index K-Means Cluster Movements of the Lionfish

Species	Year	Cluster Classification	Cluster Center
Lionfish	1998	2	0.1087986
Lionfish	2008	3	0.0446581
Lionfish	2014	4	0.0862458
Lionfish	2018	3	0.9258102

Density Index K-Means Cluster Movements of the Ocean Surgeonfish

Species	Year	Cluster Classification	Cluster Center
Ocean Surgeonfish	1998	3	1.5708361
Ocean Surgeonfish	2008	3	0.0446581
Ocean Surgeonfish	2014	3	2.2011560
Ocean Surgeonfish	2018	1	2.1440958

Density Index K-Means Cluster Movements of the Rainbow Parrotfish

Species	Year	Cluster Classification	Cluster Center
Rainbow Parrotfish	1998	3	1.5708361
Rainbow Parrotfish	2008	2	1.0520447
Rainbow Parrotfish	2014	3	2.2011560
Rainbow Parrotfish	2018	3	0.9258102