A statistical probe into the longitudinal status of coral reef fish of Grand Cayman Island

Adam Bruce

Lawrence University

Abstract:

Fish assemblages play a critical role in maintaining the health of coral reef communities throughout the world via a mutualistic relationship with the reefs themselves. Therefore, maintaining an abundant and diverse community of fish is of critical importance to reef survival. In this report, data collected from 1998 to 2018 on Grand Cayman Island were analyzed to assess overall and species-specific variation in abundance. This assessment was performed using hypothesis testing, model building, and machine learning. Ultimately, the analysis revealed an overall decrease in species abundance over time. Additionally, the analysis provided evidence of increases in Lionfish predator abundance and mixed trends of abundance for species who fed predominantly on macroalgae.

Introduction

Island Background:

The most remote islands of the Caribbean are known collectively as the Cayman Islands. Together, Grand Cayman, Cayman Brac, and Little Cayman make up this three-island archipelago located between 19 to 20 degrees north and 79 to 82 degrees west. Of the islands, the largest, at about 35 km long and 14 km wide, is Grand Cayman. This island is also the most populous of the three and was the site of data collection for this analysis. Originally, only sparse visits to Grand Cayman were made from aboriginal Indians of the Caribbean (Fewkes, 1922). However, after its discovery by Columbus in 1503, there was a profound increase in visits from European sailing ships because of the availability of food and fresh water. By the 1800's, the island began receiving research interests for natural history collections. Eventually, the primary scientific focus on Grand Cayman became large scale research after the Oxford University Cayman Islands Biological Expedition in 1938 (Brunt and Davies, 1994).

Grand Cayman is located on the Cayman Ridge at the southern end of the North American Plate. This plate has several faults resulting in high tectonic activity, which has led to a wide diversity of marine habitats. Primarily, the landscapes of these habitats are composed of hard, carbonate limestone (Jones, 1994). Topographically, the reef structures surrounding the island are of shelf, slope, and fringing nature with strong tide-modulated currents allowing them to flourish. However, the island also features a large lagoon on its Northern side known as the North Sound, which features large mangroves with many reef-dwelling fish species (Roberts, 1994). Collectively, this rich marine environment, along with the diverse range of marine species, have made Grand Cayman a common site for marine biology related studies.

Data Collection, Structure, and Wrangling:

From 1998 to 2018, the Lawrence University Marine Program (LUMP) undertook a biyearly trip to Grand Cayman Island. The objective of this trip was to document abundance and diversity of reef-fish species, provide evidence towards the health of coral reefs surrounding the island, and to understand the ecological relationships between marine species and their environments. To accomplish this, students on the trips used Reef Environmental Education Foundation (REEF) roving diver specifications to collect data 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).

On each dive, data were collected using underwater paper and pencils, with readings lasting approximately 20 to 50 minutes depending on diver air consumption and non-decompression limits. After each trip, the data were used to quantify species richness and abundances according to REEF guidelines (Timpe, 2018). A key abundance measurement resulting from this quantification was density index, which represented the number of a species present at a site during each dive. Scientifically, the density index scale represented whole number integer categories from 0 to 4, where "0" represented no individuals spotted, "1" represented a single individual spotted, "2" represented two to ten individuals spotted, "3" represented eleven to one hundred individuals spotted, and "4" represented over one hundred individuals spotted for any given species. However, because some sites were surveyed by

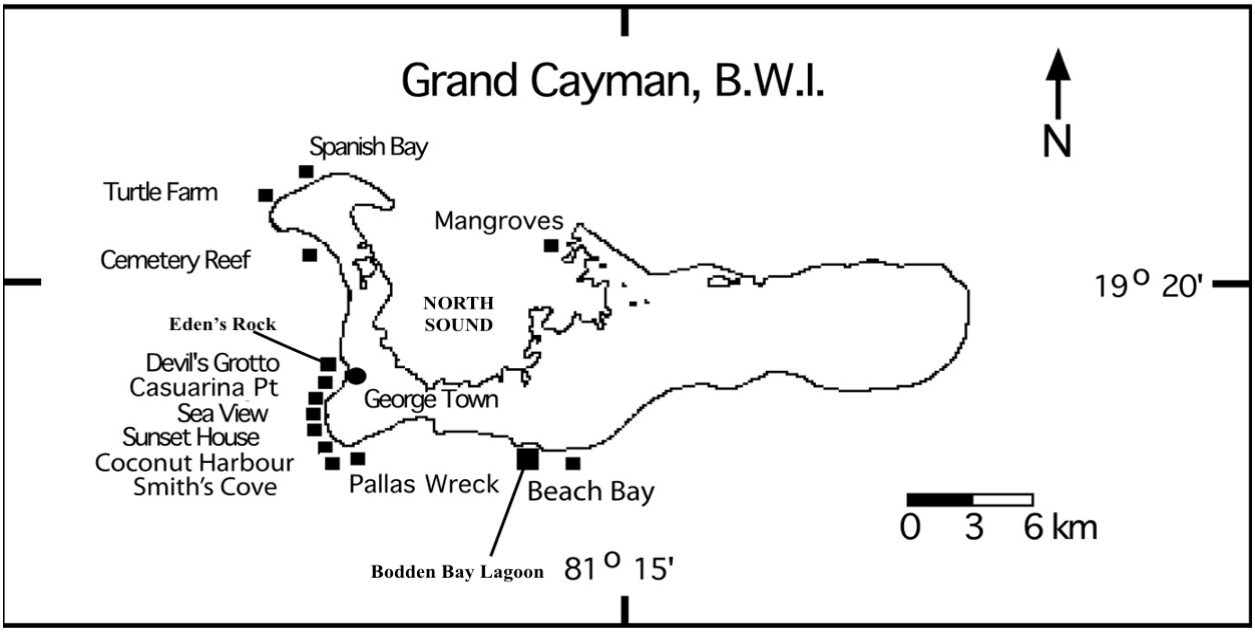
multiple divers each year, their density index measurements were averaged by species, which resulted in some non-integer values.

For this analysis, a .csv file containing ten years of data was obtained from the head of LUMP, Professor Bart De Stasio of Lawrence University. Collectively, the file contained the columns “SPECIES_NAME”, “DIET”, “SPECIES_ID”, “DENSITY_INDEX”, and “YEAR”. In full, this made the original data structure six columns consisting of 22,052 observations for 246 unique fish species. Ultimately, two additional variables were created and added to the data set. These variables were “SCIENTIFIC_FAMILY” and “MARINE_RESERVE”.

The first variable was created using information from the “SPECIES_NAME” column of the original dataset and the ‘case_when()’ function in R. In full, there were fifty-six families represented by the species in this dataset. Of those families, only those with six or more species within them were considered individual categories within the variable. All other families were grouped into the category “Other”, which resulted in twelve total subcategories in the variable. Altogether, this grouping allowed for easier graphical visualizations because it reduced the species variable from 256 species to 12 families containing the species.

Additionally, the variable “MARINE_RESERVE” was added to the dataset. This variable was created using data from the Cayman Islands Government Department of Environment marine parks boundary map (Appendix A). In total, four of the elven locations surveyed were found outside of marine reserves on the island. These included Beach Bay, Spanish Bay, Turtle Farm, and Smith’s Cove, and collectively they were represented by a “No” subcategory within the variable. With the feature engineering complete, this study aimed to understand how species abundance has varied on Grand Cayman Island. Therefore, exploratory data analysis, hypothesis testing, statistical modeling, and machine learning techniques were used to probe the dataset.

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 twelve other sites were surveyed, with Mangroves being the only dive site located within the North Sound lagoon.



Exploratory Data Analysis

Since their introduction in 1970 by statistician John Tukey, the techniques of exploratory data analysis have become a principal component to any data analysis. Collectively, these techniques allow for the detection of patterns, relationships, and basic mathematical summaries, like those relating to central tendencies or variance, for variables of interest. Ultimately, these techniques also allow for assumptions of statistical models and hypothesis tests to be checked, which is critical to their proper use. Because the biological questions surrounding this study focused primarily on changes in abundance, the variable "DENSITY_INDEX" was considered the response variable, Y, and was the focal point of this exploratory data analysis.

Response Patterning:

Behind almost all parametric statistical techniques, there is an assumption about the distribution of the response variable. For instance, the most studied distribution is a bell-shaped curve known as a Gaussian Distribution. Behind the scenes, these models, algorithms, and tests distributional assumptions to build an expectation of the response as it relates to the predictors. To visualize the distribution for "DENSITY_INDEX", a density plot was built in R using the `ggplot()` function. Ultimately, the resulting distribution was a collection of Gaussians centered around the values zero, one, two, and three (Appendix B). This was abnormal as far as common distribution go but was representative of the way the data were collected. Primarily, a single diver surveyed each site each year, which resulted in most datapoints, 20,136 of 22,056, being whole number REEF integers. Alternatively, datapoints that were averaged because of several dives, 1916 observations, created the Gaussian-like variation away from the centers of these truncations.

Interestingly, the largest of these truncations was centered around zero. In the dataset, zero was representative of a true-zero, because it occurred when no individuals of a particular species were observed during the dive. In total, 16,844 of the 22,056 or 76.4% of observations in the dataset were true zeros, which makes the response variable zero-inflated. This finding would prove influential later to model building. Additionally, it raised the question about the shape of the response without the true zero observations. Therefore, another density plot was created based solely on the distribution for non-zero observations. The resulting plot showed a higher frequency of observations around the value two, with a slightly more Gaussian shape (Appendix B). However, there was also a major truncation around the value one, and a minor hump around three. With a firm understanding of the response distribution, the next task was to identify how the other variables in the data related to it.

Variable Relationships:

It can often be tempting to make assumptions about expectations for biological data based on prior knowledge. For instance, it might be assumed that, given extensive research on coral reef decay due to global warming, the "DENSITY_INDEX" values will decrease over time in the Grand Cayman dataset. However, the biological world is extremely variable and often unpredictable, which makes exploring graphical relationships between variables essential. Therefore, violin plots were built to explore the response to categorical variable relationships, and lattice plots were built to explore the response to numeric variable relationships. However,

these plots were built for both the zero-inflated data and the non-zero data to investigate the impact of the zero values.

Essentially, violin plots are a hybrid mixture of box and whisker plots and kernel density plots (Hintz and Nelson, 1998). These plots combine local densities of observations and basic summary statistics, which makes them extremely useful for interpreting categorical relationships. For this analysis, the plots were generated with means for their "DENSITY_INDEX" values represented by shapes. Ultimately, both complete and non-zero fits on the categorical variables "LOCATION" and "MARINE_RESERVE", showed almost no variation in densities and means for the sub-categories of these variables. However, the violin plots for "DIET" and "SCIENTIFIC_FAMILY" showed some notable variations in sub-categories (Appendix C & D).

In the full data violin plot for "DIET", it was observed that 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. Collectively, the densities of observations were very similar across diets aside from coral, which had the smallest bulge at zero of the groups. However, in the plot for non-zero values, there was variation in these findings. Ultimately, the primary bulges were centered around two, but zooplankton feeders had a primary bulge around three. Of the means, zooplankton and algae feeders were now the highest, with sponge and fish feeders at the lowest.

As for "SCIENTIFIC_FAMILY", the full data violin plot showed a much more elongated observation density structure for the families Holocentridae and Pomacentridae. These two families also had significantly higher means than all other families. In the non-zero data, these two families remained the highest means, but they were joined by Sparidae and Gobiidae as the only families with mean density index values greater than two. Interestingly, both Sparidae and Gobiidae went from two of the lowest averages in the full data plot, to some of the highest averages in the non-zero plot. This indicates that species of these family are either very common or extremely rare. 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.

Perhaps the best visualization for numeric relationships of two or three variables is observed in a scatter plot. These plots include a point for every individual observation in the cartesian plane between the variables of interest. Additionally, they have the capability to work in conjuncture with categorical variables, which makes them even more powerful. In the Grand Cayman data, the primary numeric variable, outside of the "DENSITY_INDEX" response, is the variable "YEAR". When investigating its relationship to the response, a specialized scatter plot, known as a lattice plot, was used. This was done for both the zero-inclusive and non-zero data. Collectively, a lattice plot was made for the two numeric variables in conjuncture with the categorical variables "LOCATION", "MARINE_RESERVE", "DIET", and "SCIENTIFIC_FAMILY", in addition to a basic one-to-one plot of "DENSITY_INDEX" and "YEAR"

In the initial one-to-one lattice plot, a very intriguing trend was observed. When including the zero values, it appeared as though the abundance of species was generally increasing over time. However, when the zeros were removed, the trend was reversed, and abundance decreased over time (Appendix E). Interestingly, when produced in conjuncture with the categorical variables, the same trend dominated (Appendix F). In every zero-inclusive

lattice, the initial trend across the subcategories was an increase in abundance. While a majority of the non-zero lattices reversed, a select few remained increasing or were stable over time. Of the six plots that remained increasing or were stable, four of them were in the "SCIENTIFIC_FAMILY" subcategory. This included the only three plots that remained increasing in Carrangidae, Gobiidae, and Haemulidae families. The other plots which remained constant were Pomacentridae, Coral, and Sunset House of the "SCIENTIFIC_FAMILY", "DIET", and "LOCATION" plots respectively. This may be indicative of certain families and diets thriving with the decay of other families.

Mathematical Summaries:

The last aspect of the exploratory data analysis aided in understanding the tendencies of the data. Tabular summaries allow for generalizations to be made without having to dive into specific observations. In the Cayman Islands dataset, summary tables were generated for the four primary categories of interest and for the overall data based on zero-inclusive and non-zero datasets (Appendix G & H). Outside of the means, which were evaluated previously, the other statistic of focus was the standard deviation. Interestingly, almost all standard deviations decreased from the zero-inclusive to non-zero summary across all tables. In the summaries relating to the categorical variables, this indicates the zero-inclusive data has more variability within families than between them, and the non-zero data has more variability between families than within them. Overall, though, this indicates the collective data has more dispersion than the non-zero data.

Methods

Hypothesis Testing:

To investigate assumptions of parametric statistics and compare 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. In parametric statistics, the three primary assumptions are independence of samples, normality of the response distribution, and homogeneity of variance. In this study, data were collected on different individual fish across many years and sites. Because of this structure, the assumption of normality was not formally tested. Notedly, it could have been justifiable to argue that samples from the same location or year violated the assumption of independence. 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".

Essentially, even though the distribution of the response was non-Gaussian shaped, violating standard parametric assumptions, the Central Limit Theorem allowed it to be tested as though it was a Gaussian distribution (Kwak and Kim, 2017). This meant the assumption of normality was valid for parametric testing. Finally, homogeneity of variance was assessed using Levene's Test (Levene, 1960). Ultimately, this test assessed if the variance of *k* samples were equal. In R, this was accomplished using the `leveneTest()` function found within the `car` library. After each Levene Test, the comparison of means was made using either the standard ANOVA and T-Test or Welch's ANOVA and T-Test for data of unequal variance.

Interpretive Modeling:

In this study, modeling techniques were used for interpretation of abundance trends over time. Modeling approaches served to relate the response variable, "DENSITY_INDEX", to the explanatory variables, such as "SCIENTIFIC_FAMILY" or "YEAR". Initially, ordinary least squares linear models of the form $Y_i = B_0 + B_1X_i \ \ldots \ B_pX_{ip} + \epsilon_i$ were considered for the zero-inclusive Grand Cayman data because of their relative simplicity for interpretation. However, when these models were fit, the resulting residual plots indicated major modeling violations. Therefore, other modeling techniques had to be considered. At first, a Zero-Inflated Poisson modeling approach was evaluated, but the response was not fully discrete, which meant this model was also invalid. Further investigation into other models were also unsuccessful because the response distribution of "DENSITY_INDEX" did not match the assumed distribution of these models, such as those using Tweedie, exponential, beta, and gamma distributions. Fortunately, the techniques surrounding modeling with two parts were eventually identified as a plausible approach to modeling the Grand Cayman Data.

The two-part modeling approach models outcomes 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 all other non-zero observations fell under $Y_i > 0$. The first piece in this approach was thus used to model the response observations satisfying or dissatisfying the condition $Y_i = 0$, which was thought of as a binary categorical outcome of "yes" or "no". Almost immediately, binomial logistic regression was recognized as being ideal for this piece because the technique uses binary outcomes and is relatively simple to interpret.

The binomial logistic regression model was fit using only "SCIENTIFIC_FAMILY" as an explanatory variable because the twelve families encompassed all individual species. Doing this allowed the model to be more accurate, as including all individual species would have increased the total number of coefficients in the model leading to overfitting. When this occurs, the estimates associated with each of the coefficients become inflated, which would have yielded inaccurate interpretations in the logistic regression model. Therefore, the overall model was built using the equation $E(Y_i) = P(Y_i \neq 0) = \pi_i = \frac{e^{B0+B1X_{i1}+\cdots B_pX_{ip}}}{1+e^{B0+B1X_{i1}+\cdots B_pX_{ip}}}$, where π_i represented the probability that the "DESNITY_INDEX" value for the *ith* family was non-zero. Overall, this logistic model accomplished the modeling goal of the first piece of the two-part model.

In the second piece of the two-part modeling approach, response values of zero were not used because the focus was on modeling "DENSITY_INDEX" values of $Y_i > 0$. This meant the non-zero dataset was used. Multi-level longitudinal techniques were identified as a useful modeling approach for this data because the data was observed biyearly. This type of modeling is built on the foundations of linear mixed-effects models, which use both fixed and random effects to investigate difference or relationships between variables. Fixed effects are considered for variables which are of focal importance to the study or model, while random effects are for variables which are not focally important but play a role in accounting for variability and correlation in the data (Roback and Legler, 2021). Longitudinal versions of these models are built on multiple levels and consider the inclusion of random slope terms. However,

most importantly, their interpretations mirror ordinary least squares regression models, which made them ideal for the second piece of the two-part model.

Level one variables in this model were considered observations that changed over time. Therefore, the variables: “YEAR”, “LOCATION”, and “MARINE_RESERVE” were considered plausible measures at this level. Meanwhile, level two variables were considered observations which remained constant over time, which included “SCIENTIFIC_FAMILY” and “DIET” as considerations for this level. Ultimately, this model structure allowed for comparisons of variability within and between species. Finally, the model could implement random slope terms, which allowed for rates of change for an explanatory variable to vary from species to species. An example of a plausible random slope term that was investigated was “YEAR1998”, which accounted for the fact that species “DENSITY_INDEX” values may have changed at different rates over time (Figure 2).

Overall, a general structure of model building was used for this longitudinal approach. First, an Unconditional Means Model was fit followed by an Unconditional Growth Model from which more complex models were built. The Unconditional Means Model allowed for calculation of an intraclass correlation coefficient (ICC) for measuring variability within and between species with the equation $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma^2}$. In this equation, σ^2represented the total variance within datapoints for a single species over time, while σ_u^2 represented the total variance between datapoints of all species over time. Alternatively, the Unconditional Growth Model allowed for an investigation into the longitudinal structure of the data by including only “YEAR1998” as a covariate. Collectively, these models served as foundations for finding the optimal model fit for interpretation using Maximum Likelihood-Ratio Tests.

By adding more explanatory variables to an initial model a better fit can often be found. In the multi-level longitudinal models, the Unconditional Means Model was the most basic model, followed by the Unconditional Growth Model. From these models, additional explanatory variables were added, and thus additional coefficients were produced. Ultimately, the addition of these variables was either impactful or not, which was indicative of whether they would produce meaningful interpretations. To evaluate whether an addition to the model was useful, comparisons were made using the `add1()` function in R. This function tested Maximum Likelihood-Ratios between the nested and additive models based on a Chi-Square distribution (Etz, 2018). Additionally, the function produced outputs based on Akaike’s Information Criterion (AIC) which was also considered for evaluation (Kuha, 2004).

The significance level in the Maximum Likelihood-Ratio test was evaluated at α =0.05, and plausible interactions between “YEAR1998” and “DIET” as well as “YEAR1998” and “SCIENTIFIC_FAMILY” were investigated alongside simple variable additions. These specific interactions were investigated because of graphical evidence that these interactions may have existed (Figure 3). After finding the optimal additive fit, evaluations were performed using AIC only for the addition of random slopes. In some cases, the AIC and Maximum Likelihood tests were indicative of better models, but all interpretability was lost due to loss of significance in the coefficients of the model. Therefore, the model that produced the lowest AIC value, had significant additions, and was most interpretable was selected for final interpretations. Ultimately, this optimal model was used to achieve the interpretation goals of the second piece of the two-part model.

Machine Learning:

The final approach used for understanding abundance over time was an unsupervised machine learning technique known as K-Means Clustering. This technique involves grouping numeric observations together that are most similar, and functions based on a distance measurement between observations (Steinley, 2010). For this analysis, the distance measure used was Euclidean Distance. Initially, the Grand Cayman dataset was broken down by years using the `filter()` and `group_by()` functions in R. Groupings for each year were based on the variable "SPECIES_NAME", which meant the number of observations, known as the sample size or simply n, within a grouping was indicative of the number of sites the species was observed at during the year. Thus, this allowed species to be grouped not only by their density index numeric measurements, but also by a numeric representation of locations they were observed at named "N_SITES". To ensure the groupings were indeed most similar, the resulting datasets for each year were scaled so that the grouping variables were of the same measure.

With the yearly datasets prepped, it was imperative to determine how many grouping clusters, our K, would be used. To determine this, the function `fviz_nbclust()` was used. Ultimately, this function produced a plot for choosing a K value which optimized the drop in total sum of squares, which was optimized at an elbow-like bend in the plot. For each of the ten years, it was decided to use a K of four when clustering primarily because four was a reasonable choice observed in each `fviz_nbclust()` plot, and this made for easier comparisons between years (Appendix I). With K decided, the actual clustering was performed on each year using the `kmeans()` function in R. This function allowed for a specification of random starting points for each clustering, which was always set at twenty-five.

Once complete, visualizations of the four optimal clusters were produced using the `fviz_cluster()` function (Appendix J). Additionally, the resulting centers for "DENSITY_INDEX" and "N_SITES" were calculated along with the sample size of each cluster. However, the centers had to be transformed back to their unscaled values for reliable interpretation (Appendix K). Ultimately, the clusterings in which an individual species fell were also assigned, using the "SPECIES_NAME" variable. Assignment was carried out because of the interest in comparing results to previous studies on similar species in the Caribbean, such as those on herbivore biomass and Lionfish densities (Pattengill-Semmens and Semmens, 2003; Benkwitt, 2002; Williams and Polunin, 2001). Specifically, the species used for comparisons were Lionfish, Rainbow Parrotfish, and Ocean Surgeonfish. The Rainbow Parrotfish was randomly chosen to represent Parrotfish, while the other two species were specifically mentioned in previous work. The data on the clusterings and centers of "DENSITY_INDEX" values in 1998, 2008, 2014, and 2018 were used to draw conclusions about the movement of these species over time.

Figure 2: Species' "DENSITY_INDEX" values changed at different rates over time. Scatter plots with loess fits, shown as red lines, 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 "YEAR1998".

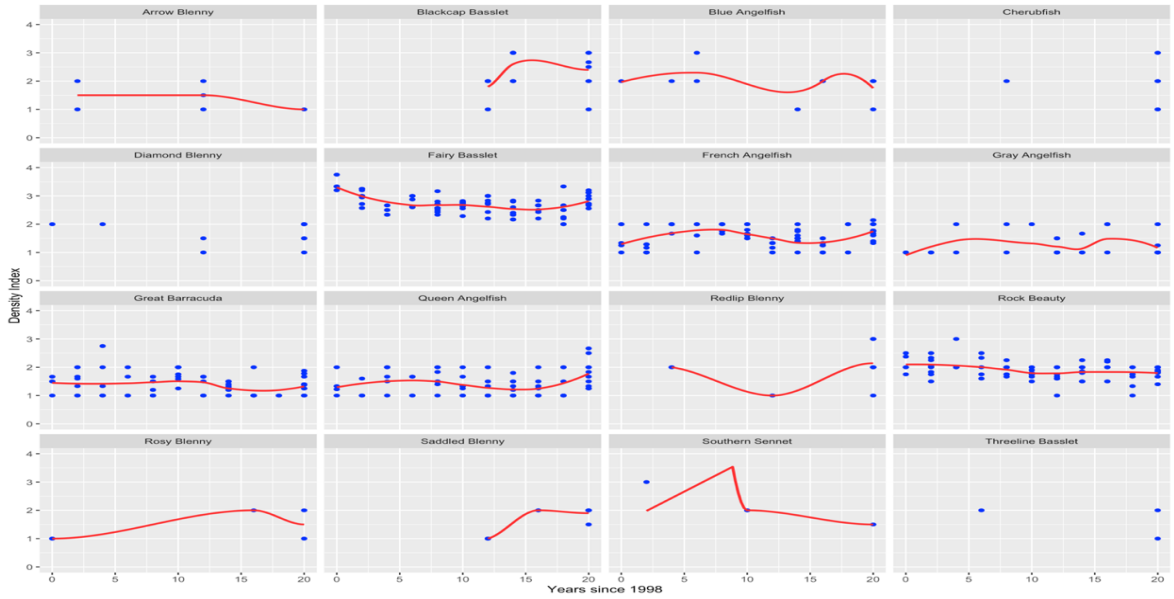
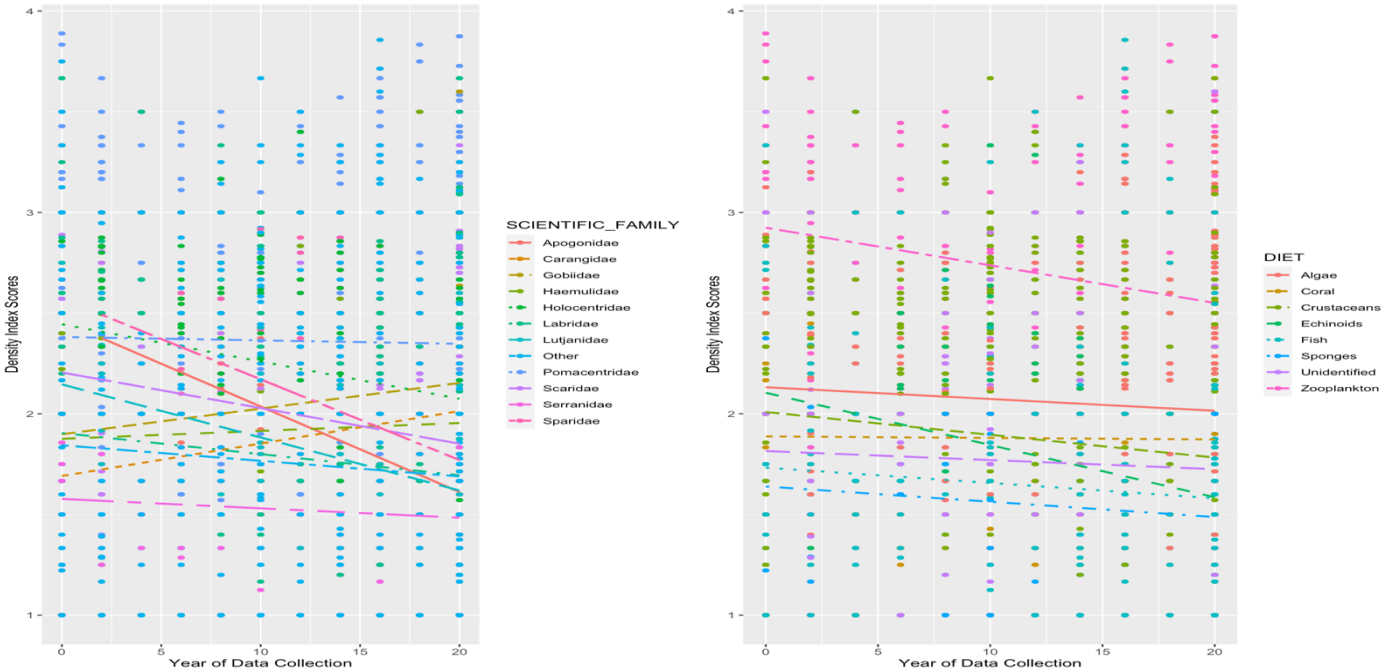


Figure 3: Plausible interactions between diet and time as well as family and time. Scatter plots with linear fits by "SCIENTIFIC_FAMILY" (left) and "DIET" (right) are pictured. The rates of change between families are clearly not constant, and neither are the rates of change between diets. This indicates the rate of change in "DENSITY_INDEX" scores by "YEAR1998", a measure of time, depend on the "SCIENTIFIC_FAMILY" or "DIET" justifying interaction.



Results

Hypothesis testing was caried out using both the zero-inclusive and non-zero data. The 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 provided evidence that the variances between the sub-categories of these variables were unequal (Appendix L). In the Levene 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" in this data (Appendix L). The significant p-value groups were concluded to have evidence supporting unequal variance, while the insignificant p-value groups were concluded to have evidence supporting equal variance. Ultimately, these findings were used to determine the correct testing procedure for making comparisons of means.

Because the zero-inclusive categorical variables were all concluded to have unequal variance, testing for comparison between means of their sub-categories had to be performed with this considered. Therefore, Welch's ANOVA was used for "DIET", "SCIENTIFIC_FAMILY", and "LOCATION" since they had more than two subcategories. Alternatively, Welch's T-Test was used for "MARINE_RESERVE" because it had only one sub-category. Overall, the resulting p-values from this test were all significant at an alpha of 0.05, which provided evidence that the true mean differences of the sub-categories were not zero (Appendix M).

In the non-zero data, testing structures for "LOCATION" and "MARINE_RESERVE" had to be built on the assumption of equal variance. Therefore, the standard ANOVA and equal variance t-tests were used for these variables respectively. As for "DIET and "SCIENTIFIC_FAMILY", the tests were conducted using Welch's ANOVA. Overall, significant p-values were obtained in the tests for "DIET and "SCIENTIFIC_FAMILY" subcategories, but not for the subcategories of "LOCATION" and "MARINE_RESERVE" (Appendix M). This provided evidence that the true mean differences of their sub-categories were not zero. Additionally, for the second two variables, it provided evidence that the true mean differences of their subcategories could plausibly be zero.

Upon conclusion of hypothesis testing, model building for interpretation was carried out with a two-part approach. Ultimately, a logistic regression model was fit to the zero inclusive data and indicator terms were used to estimate increases or decreases between families. The baseline estimate for this model, β_0, was measured for the family Apogonidae. Overall, there were eleven additional model coefficients in the model, with eight achieving significant p-values at an alpha of 0.05 (Figure 4). Additionally, the model was used to calculate the probability that an observation of a certain family was non-zero, which if satisfied allowed for interpretation with the longitudinal multi-level linear mixed effect models (Figure 5).

In the non-zero data, the initial longitudinal multi-level linear mixed effect model was an Unconditional Means Model which did not include predictors of "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 N). The intraclass correlation coefficient (ICC) was also found using this model by applying the equation:

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

The resulting ICC from the model was 0.423. This indicated that 42.3 percent of the total variation in "DENSITY_INDEX" values was attributed to differences between species as opposed to differences within the same species. Thus, there was more total variability in "DENSITY_INDEX" values within individual species (57.7%) than between them. Also derived from this model were standard deviations associated with the random effects table. These included $\sigma = 0.5151$ and $\sigma_u = 0.4413$. This σ was the standard deviation in "DENSITY_INDEX" values for the same species between different years, while σ_u was the standard deviation in "DENSITY_INDEX" values between species averaged across twenty years (1998 to 2018). After fitting this model, a more complex Unconditional Growth Model was built.

The Unconditional Growth Model introduced time, indicated by the "YEAR1998" variable, as a level one covariate but did not include any predictors at level two (Appendix N). Additionally, the model included a random slope term for "YEAR1998". Overall, this model was of the form:

$$Y_{ij} = B_0 + B_1 YEAR1998_{ij} + u_i + v_i YEAR1998_{ij} + \epsilon_{ij}, \text{ with}$$
$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & 1 \\ \rho_{uv}\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix}\right) \text{ and } \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Again, Y_{ij} defined the "DENSITY_INDEX" value for species i in year j. This model provided fixed effects estimates of 1.84 for B_0 and -0.008 for B_1, with both estimates obtaining significant p-values. B_0 represented the average "DENSITY_INDEX" value across all species in 1998 while B_1 indicated the average yearly 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.

This σ represented the standard deviation in "DENSITY_INDEX" values for an individual species, σ_u represented the standard deviation in "DENSITY_INDEX" values between all species 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}, which was -0.52. This ρ_{uv} measured correlation between species 1998 "DENSITY_INDEX" values and their rates of change in this value between 1998 and 2018. Since both B_1 and ρ_{uv} were less than zero, species with higher "DENSITY_INDEX" values in 1998 had larger decreases in this value 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.

From the Unconditional Growth Model, more complex models were built using estimates of Maximum Likelihood Ratios, AIC, and interpretability as deciding factors for the optimal model. Overall, variable additions which yielded lower AIC values and higher Maximum Likelihood Ratio test values were considered to have improved the model. First, the categorical variables "DIET", "SCIENTIFIC_FAMILY", "LOCATION", and "MARINE_RESERVE" were tested as additive effects. Overall, the baseline AIC of 8413.0 fell with the addition of "DIET" and

"SCIENTIFIC_FAMILY", but not for "LOCATION" or "MARINE_RSERVE". Additionally, the Maximum Likelihood Ratio test values, denoted under "LRT" in the outputs, were higher in "DIET" and "SCIENTIFIC_FAMILY", but not in "LOCATION" or "MARINE_RESERVE" (Appendix O). This led to both "LOCATION" and "MARINE_RESERVE" being excluded from the model, and both "DIET" and "SCIENTIFIC_FAMILY" being added to the model.

Next, the interaction effect of "YEAR1998" and "DIET" as well as "YEAR1998" and "SCIENTIFIC_FAMILY" were tested. In this case, AIC was improved with the addition of an interaction between "YEAR1998" and "SCIENTIFIC_FAMILY", but not with the addition of the "YEAR1998" and "DIET" interaction. Additionally, the Maximum Likelihood Ratio test for "YEAR1998" and "SCIENTIFIC_FAMILY" yielded a large increase, while the test for "YEAR1998" and "DIET" did not (Appendix O). While the addition of the "YEAR1998" and "SCIENTIFIC_FAMILY" was considered an improvement based on these tests, it was ultimately excluded in the final model because of the loss of significance in many coefficients. For instance, only three diets and three interaction terms had significance in this model, half of those only being significant when tested at an alpha of 0.10 (Appendix P). Regardless of the fact the interaction was excluded, this model brought to light the likelihood that rates of change for families are not constant, which would be beneficial if the data were being used for prediction. Finally, the addition of a random slope term for "YEAR1998" was tested using AIC, which resulted in a significant drop in AIC (Appendix O). This random slope term was easily interpretable, so it was added to the final model.

After testing, the final model included additive effects for "YEAR1998", "DIET", and "SCIENTIFIC_FAMILY" along with a random slope effect for "YEAR1998". Before making interpretations from this model, diagnostics were performed to check the assumptions of normality, independence, and linearity. Overall, the normality assumption appeared reasonable with little deviation observed in the QQ plot. Additionally, the residual versus fitted plot showed only slight violations of linearity due to some patterns in the residuals. Finally, the assumption of independence was not violated in the structure of the data (Appendix Q). Therefore, interpretations from the model were valid.

Overall, the final fitted model produced twenty fixed-effect coefficient estimates of which ten were significant at an alpha of 0.10 (Figure 6). The intercept, B_0, was 1.916 and estimated the expected "DENSITY_INDEX" value in 1998 for the family Apogonidae with a diet of algae. The estimated betas for diets of coral and crustaceans in addition to families of Carangidae, Haemulidae, Labridae, Lutjanidae, Other, Scaridae, Serranidae, and Sparidae were all insignificant in this model. Therefore, their estimated "DENSITY_INDEX" values in 1998 were also 1.916. All other coefficients had significant p-values and were thus interpretated.

The coefficient for "YEAR1998", B_1, was -0.0165, which indicated the average yearly decrease in "DENSITY_INDEX" was -0.0165. 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. The coefficient for a diet 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. The coefficient for 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 fish with 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. 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 zooplankton compared to fish feeding on algae. In addition to these baseline 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. As for the random effects table for 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 interpretation for these estimates were the same as in 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 declines in this value over time than was observed in the Unconditional Growth Model.

In the analysis of centers for K-Means Clustering, it was observed that the range of maximum "DENSITY_INDEX" centers from 1998 to 2008 was between 2.18 and 3.25, while the range of minimum centers was between 0.039 and 0.11. Alternatively, 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 Lionfish 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 R). 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.

Figure 4: R output for the Logistic Regression Model fit to the zero-inclusive dataset. Model equation, coefficient estimates, and corresponding p-values are depicted.

```
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 5: Resulting probabilities that an observation from a particular family was non-zero as determined by the logistic regression model are pictured. Corresponding p-values from the model are also provided.

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%)

Figure 6: R output for the final model from the non-zero data fit using longitudinal linear mixed-effects techniques. 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.

```
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
```

Discussion

This study aimed to understand how species abundance has varied on Grand Cayman Island. During hypothesis testing on the non-zero data, the variables "LOCATION" and "MARINE_RESERVE" did not yield significant p-values for differences in means of their subcategories. This finding indicated that fish on Grand Cayman Island did not differ in their "DENSITY_INDEX" values from site to site or when they were observed in protected areas. This supported previous research on Marine Protected Areas (MPA) compared to Non-Marine Protected Areas on Grand Cayman Island where researchers observed that the density of species compared between sites in an MPA and sites outside of an MPA were not significantly different (McCoy et al., 2009). They concluded from this finding that the insignificant difference was due to the efficiency in design of the MPA's on Grand Cayman. Essentially, these sites allowed species to grow to larger sizes and produced more homogeneity, which was independent of abundance. However, there were also significant p-values for differences of subcategory means observed during hypothesis testing. This occurred in the variables "DIET" and "SCIENTIFIC_FAMILY". This finding indicated that fish on Grand Cayman Island differed in their abundance based on what they primarily ate and what family they were from, which was ultimately supported by modeling.

When modeling the zero-inclusive data with logistic regression, it was discovered that Holocentridae and Pomacentridae had relatively high probabilities of observations being non-zero compared to the other families over time. Overall, this indicated that species of these two families were more abundant in the zero-inclusive data compared to the other families. This is supported by the slopes in the lattice plots of "DENSITY_INDEX" over time for the zero-inclusive data (Appendix F). Conversely, this model indicated that the families Apogonidae, Carangidae, Gobiidae, and Sparidae had relatively low probabilities of observations being non-zero compared to other families over time. This indicated that species of these families were less frequently observed over time compared to the other families.

Perhaps the most meaningful conclusions from this study come from the longitudinal linear mixed effects model estimates. Collectively the Unconditional Means Model and Unconditional Growth Model conferred both information on variability in the data and the general trend in abundance over time. While it was important to recognize the nature of this variability, the findings of decreasing abundance and those associated with ρ_{uv} were most relevant to this study. Outside of these models, the final interpretive model, which primarily portrayed changes in abundance over time based on families and diets, was used to assess differences in abundance.

Overall, this model found that "DENSITY_INDEX" values have decreased on Grand Cayman Island over time. Support for this finding was obtained by looking at the maximum cluster centers of the K-Means models from 1998 to 2008 and from 2010 to 2018. Ultimately, it was observed that between 1998 and 2008, the maximum cluster centers were much higher than those from 2010 to 2018. Collectively, these findings supported previous research which had used 3727 abundance readings across numerous Caribbean reefs to identify a decline in both reef structures and abundance over a thirty-year period (Alvarez-Filip et al., 2015). Additionally, thanks to the negative correlation estimate of ρ_{uv} in the model, it was discovered that species that had higher "DENSITY_INDEX" values in 1998 had faster decreasing

"DENSITY_INDEX" values over time. This meant that species with higher initial abundance saw much greater decreases in abundance than those with lower initial abundances.

Specifically, species of the families Holocentridae, Gobiidae, and Pomacentridae as well as those feeding on algae, coral, crustaceans, and zooplankton would have experienced this trend, as they had the highest baseline abundance measures in the final model. In addition, this model provided evidence that the decreasing abundance trends over time did not vary significantly between the families Apogonidae, Carangidae, Haemulidae, Lutjanidae, Scaridae, Serranidae, Sparidae, and those grouped under "Other". Also, they did not vary between species with primary diets of algae, coral, and crustaceans. Collectively, all other families and diets showed variation in their decreasing abundance trends over time in accordance with their coefficient estimates from the final model.

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 indicated that abundance of the Lionfish has likely increased over time. A previous study has indicated 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. Previous research had revealed increasing macroalgae abundance on Grand Cayman Island over time (Williams and Polunin, 2001). While these researchers hypothesized that increases in herbivore abundance, the species feeding on these macroalgae, would lead to a decrease in the macroalgae cover due to grazing, other researchers observed that this hypothesis did not always hold true (Pattengill-Semmens and Semmens, 2003). For instance, these researchers observed that increasing abundance of Parrotfish algal feeders led to an increase in algal cover, while increasing abundance of Surgeonfish algal feeders led to a decrease in algal cover. Using the K-Means Clusters for the Ocean Surgeonfish and Rainbow Parrotfish, this finding was assessed in the Grand Cayman Data. Overall, the Rainbow Parrotfish was grouped with clusters of lower "DENSITY_INDEX" centers over time, while the Ocean Surgeonfish was grouped with clusters of higher "DENSITY_INDEX" centers over time. This contradicted the findings by Pattengill-Semmens and Semmens (2003), as the increasing macroalgae cover over time on Grand Cayman should have led to Rainbow Parrotfish moving to clusters with higher centers and Ocean Surgeonfish moving to clusters with lower centers over time.

Collectively, the findings in this report indicate that species abundance on Grand Cayman Island has declined between 1998 and 2018. Understanding the broad and specific differences in declines produced through modeling and clustering here could prove beneficial in identifying target species, families, or diets for recovery initiatives. However, what kinds of initiatives would prove most impactful on these targets will have to be the focus of future research.

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 detailed notes from STAT 455 on modeling longitudinal data. Finally, I would like to thank Professor Bart De Stasio of Lawrence University for providing the data used in this report and for advisory role throughout the project.

Project Repository

The complete Rmarkdown containing the code used in this report can be accessed at my Github repository at this url: 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.

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.

Brunt, M. A., and Davies, J. E. (1994). Scientific studies in the Cayman Islands. In *The Cayman Islands: Natural History and Biogeography*, pp. 1-12. Berlin, Germany: Springer Science+Business Media.

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

Fewkes, J. W. (1922). Prehistorical island culture. *Annual Report Bureau of American Ethnology.* 34, 49-281.

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

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.

McCoy, C. M. R., Dromard, C. R., and Turner, J. R. (2009, 11 2-6). An evaluation of Grand Cayman MPA performance: A comparative study of coral reef fish communities. *Proceedings of the 62nd Gulf and Caribbean Fisheries Institute,* Cumana, Venezuela.

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.

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.

Roberts, H. H. (1994). Reefs and lagoons of Grand Cayman. In *The Cayman Islands: Natural History and Biogeography,* pp. 75-104. Berlin, Germany: Springer Science+Business Media.

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

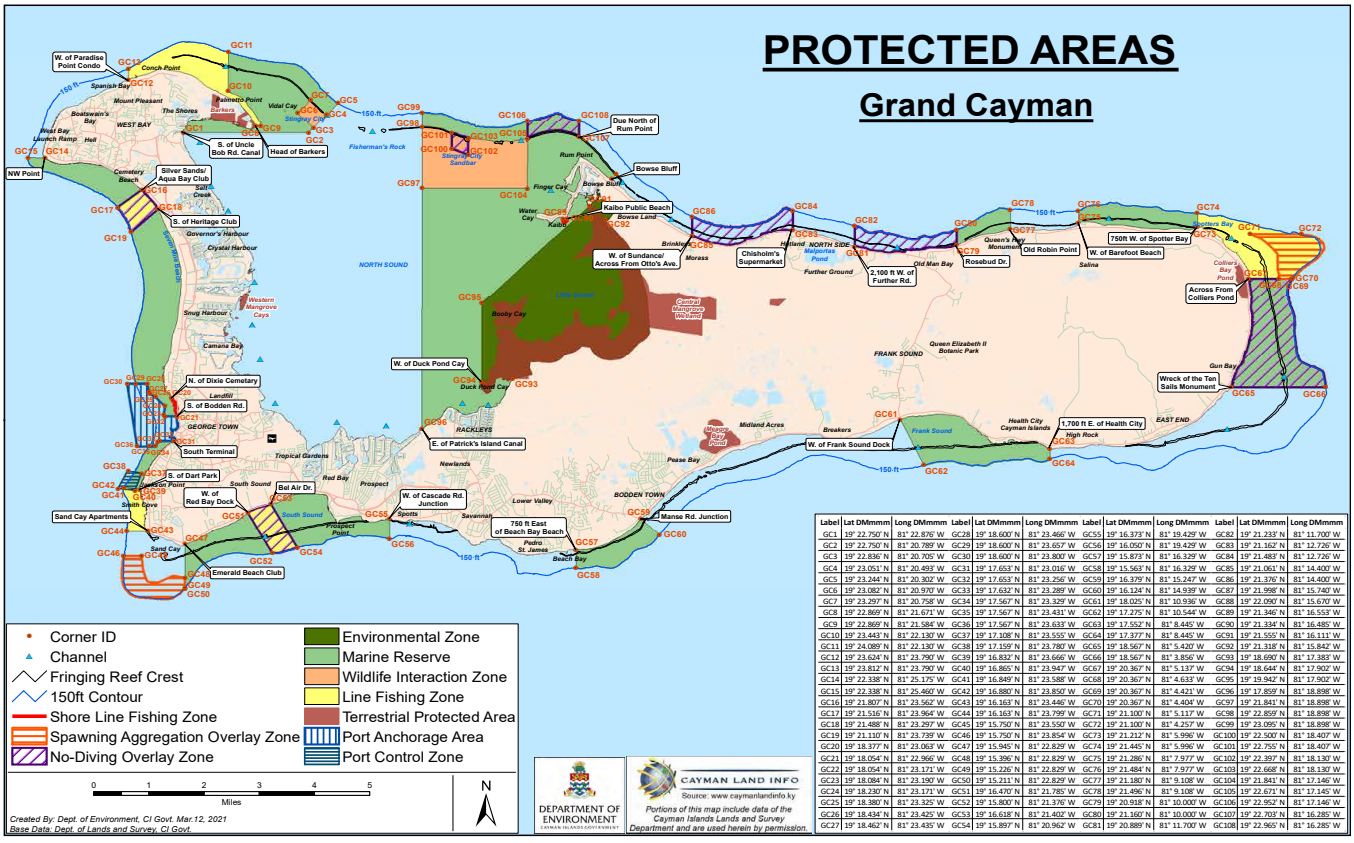
Timpe, A. (2018). A survey of Grand Cayman reefs. *Unpublished Manuscript.*

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, 358-356.

Appendix A

Cayman Island Government Department of Environment Marine Park Map

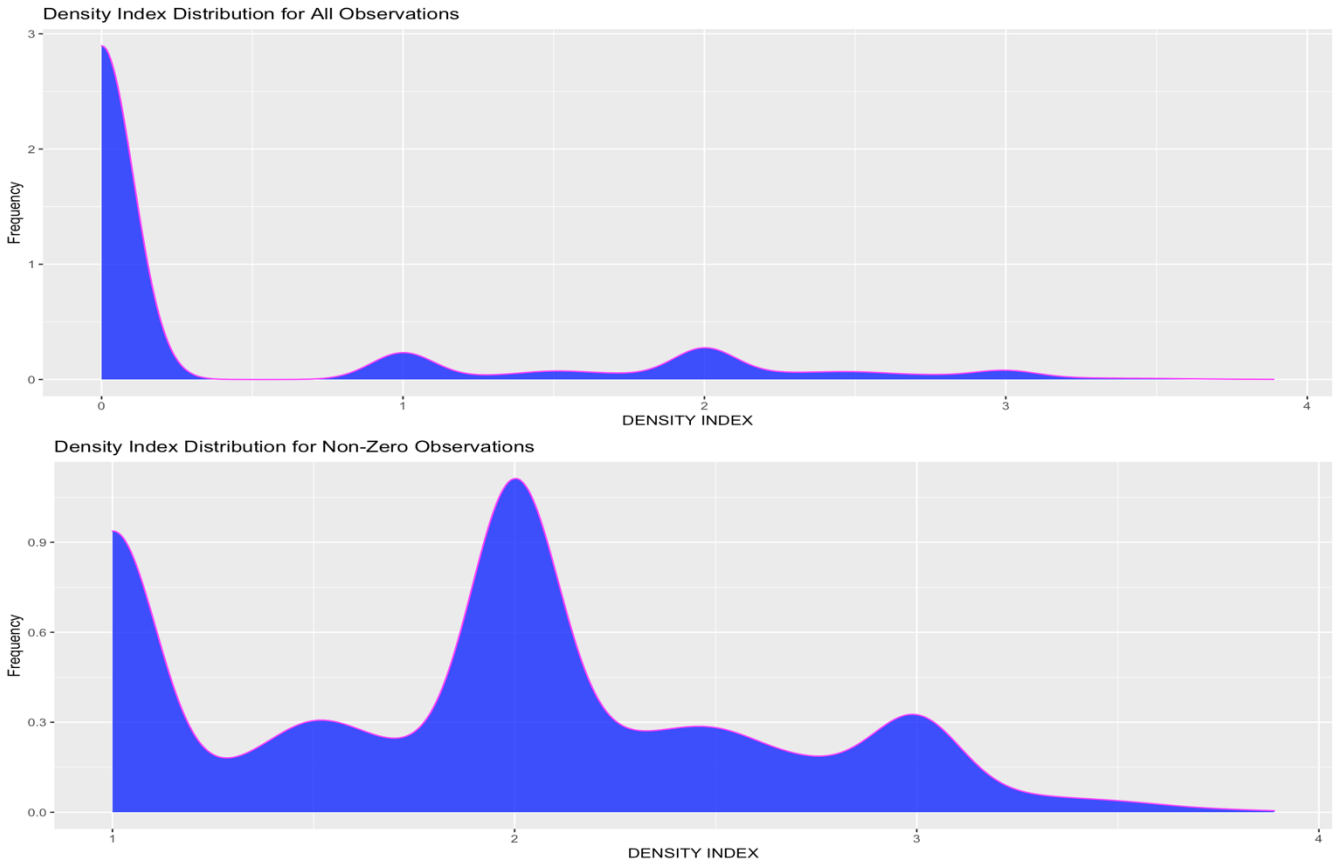
The following map was produced in 2021 and was used to analyze whether a dive site was contained within a marine reserve area or not.



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

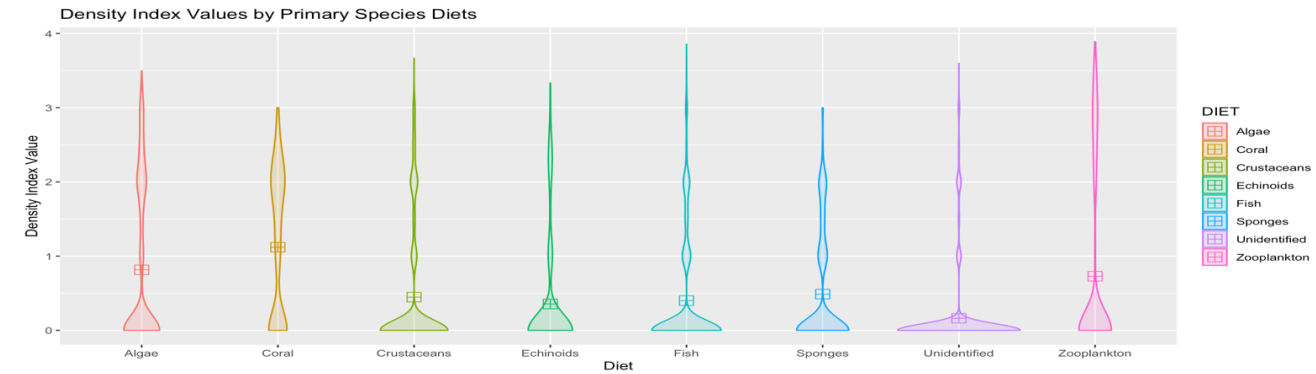
Appendix B
Density Plots for the Response Variable "DENSITY_INDEX"

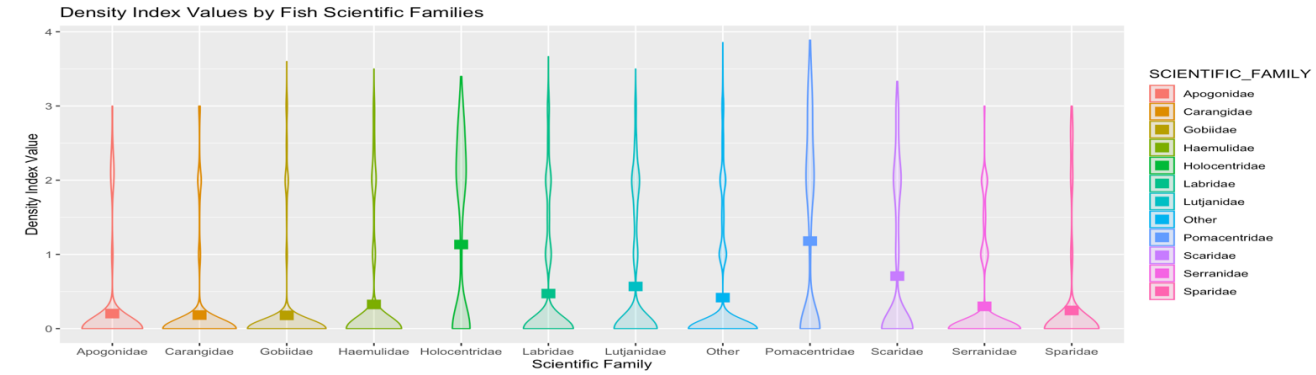
The complete and non-zero density distributions for the abundance measurement response variable density index. Heavy zero-inflation is observed in the complete distribution with minor Gaussian-like truncations centered around whole number integer responses. A more Gaussian-like shape is observed in the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

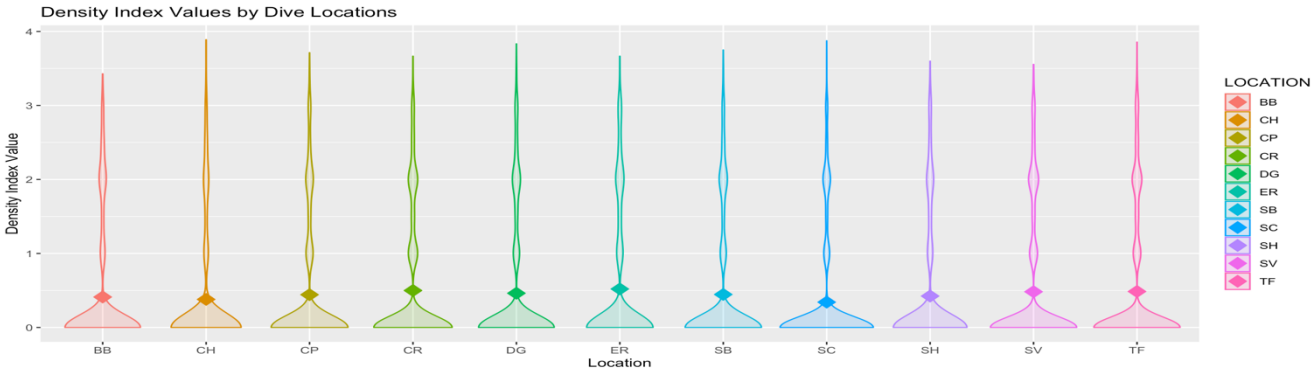


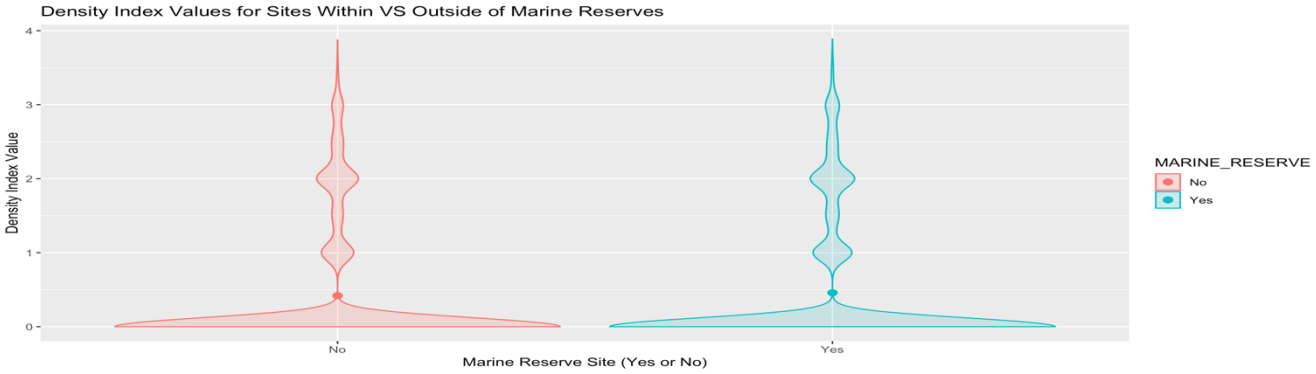
Appendix C
Zero-Inclusive Violin Plots of "DIET", "SCIENTIFIC_FAMILY", "LOCATION", and "MARINE_RESERVE"

Variation in observation density and sub-category means is observed by violin plots. Means are represented by shapes with colors denoting sub-categories.





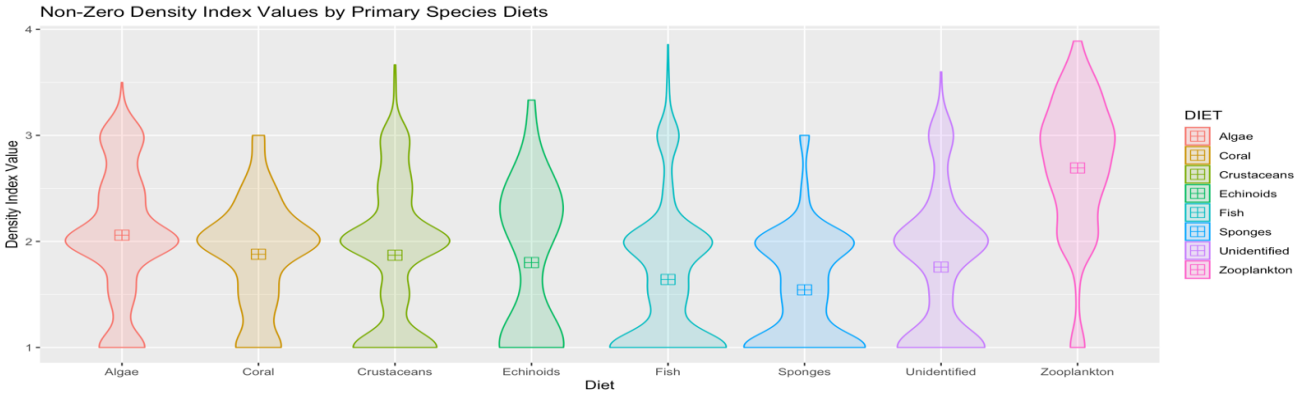


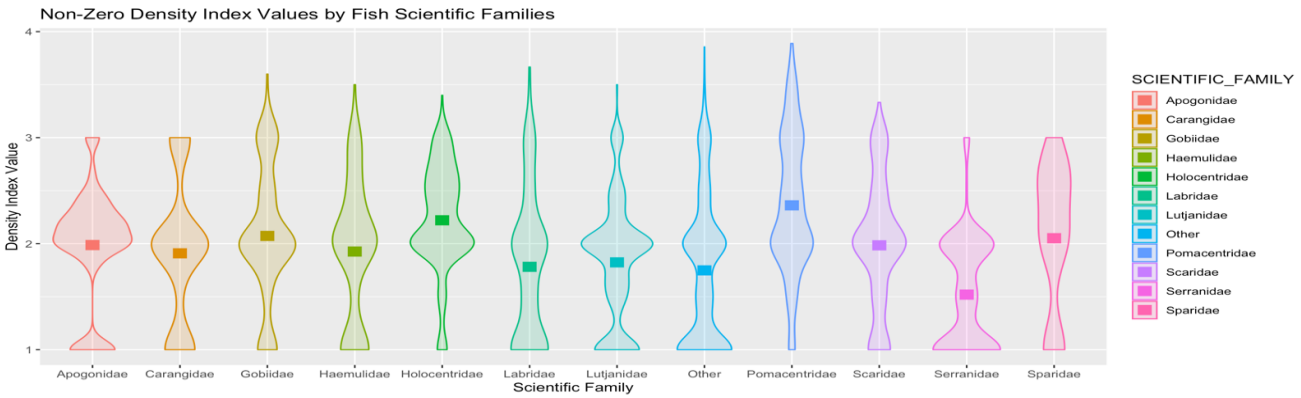


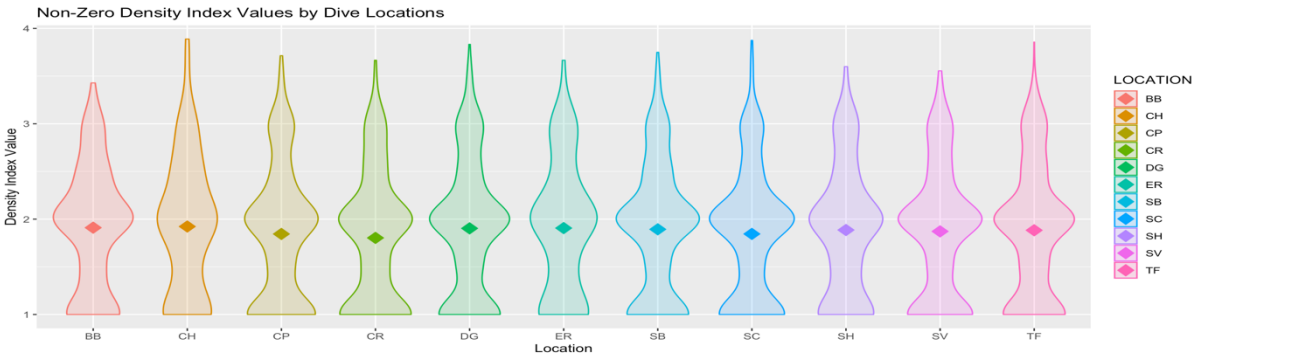
Appendix D

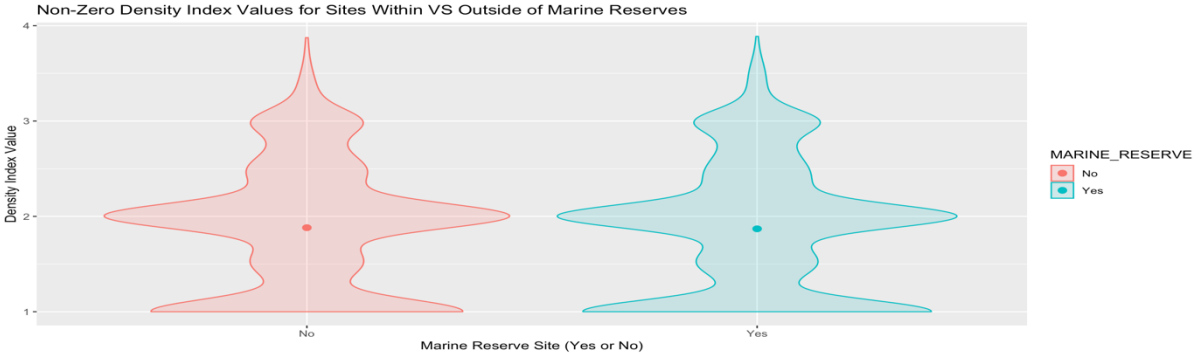
Non-Zero Violin Plots of "DIET", "SCIENTIFIC_FAMILY", "LOCATION", and "MARINE_RESERVE"

Variation in observation density and sub-category means is observed by violin plots. Means are represented by shapes with colors denoting sub-categories





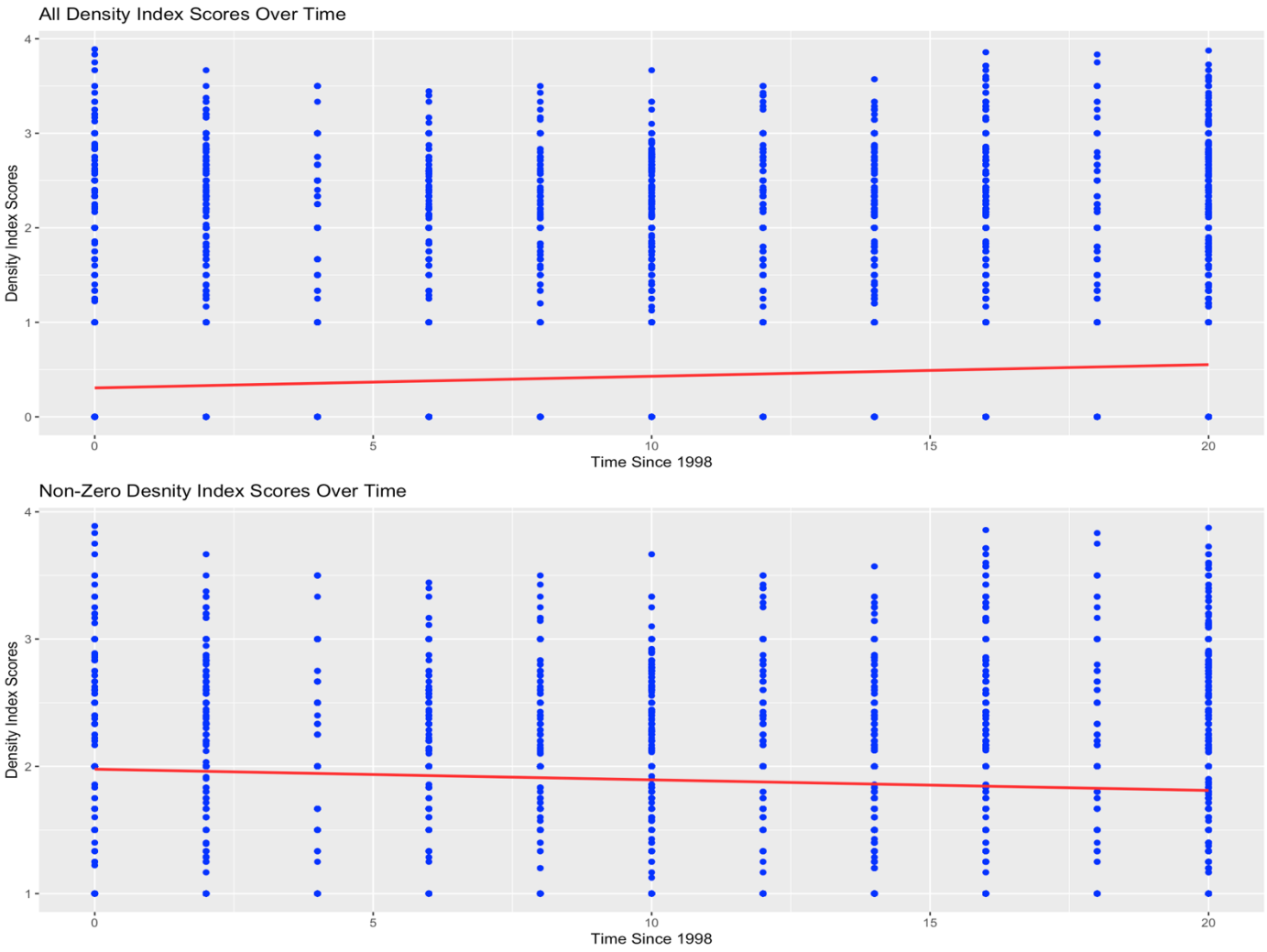




Appendix E

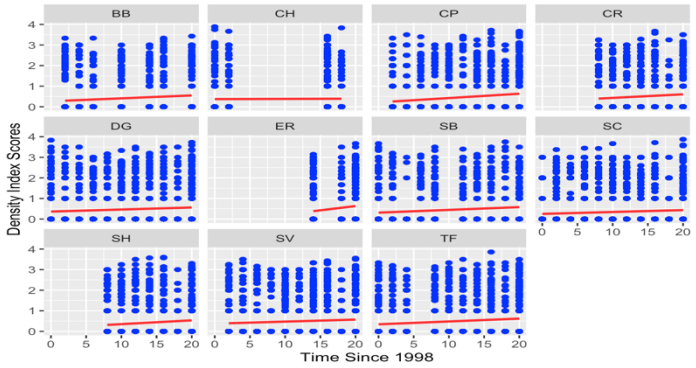
Zero-Inclusive and Non-Zero Lattice Plots of "DENSITY_INDEX" by "YEAR"

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.

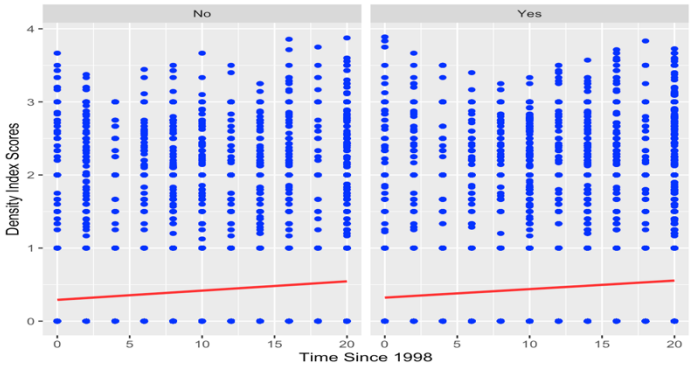


Appendix F
Zero-Inclusive and Non-Zero Lattice Plots in Conjuncture with Categorical Variables

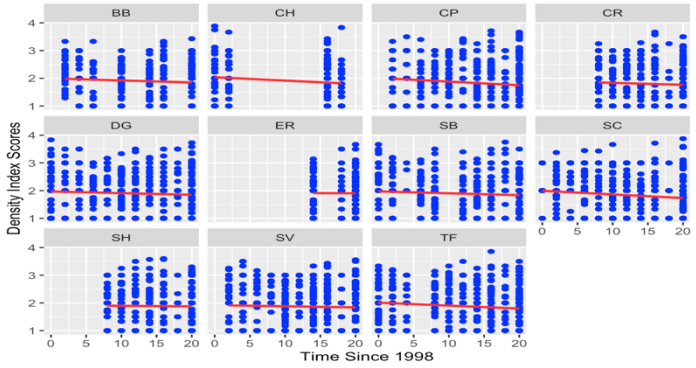
All Density Index Scores by Location



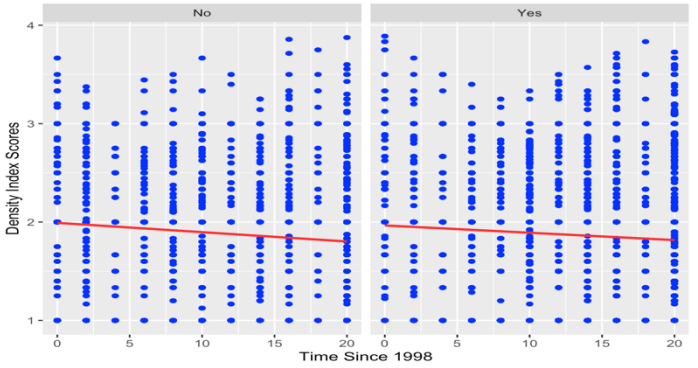
Non-Zero Density Index Scores for Marine Reserve Vs. Non-Reserve



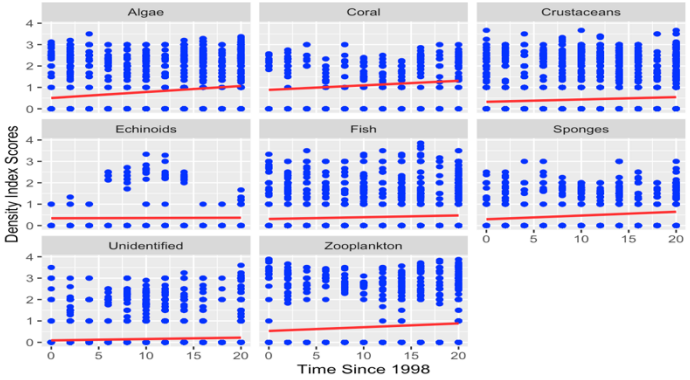
Non-Zero Density Index Scores by Location



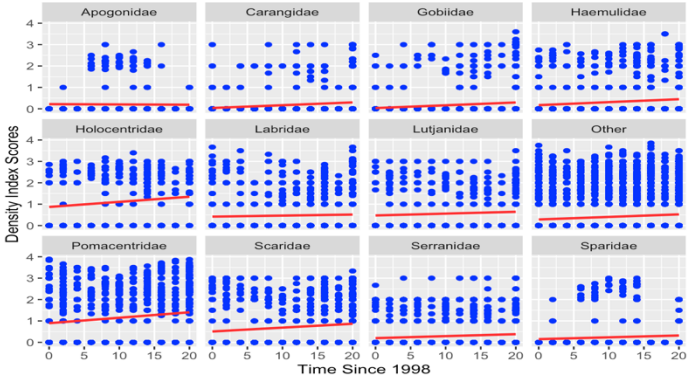
All Density Index Scores for Marine Reserve Vs. Non-Reserve Sites



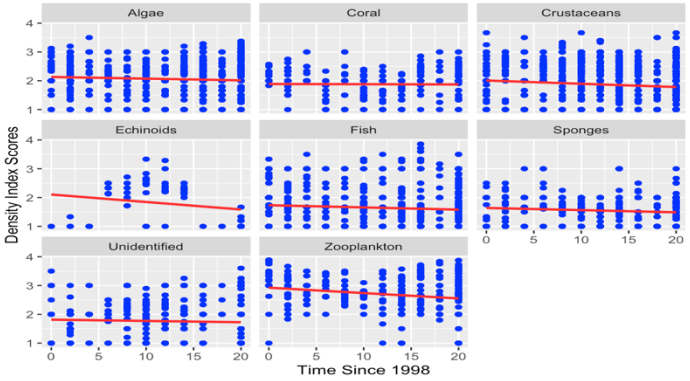
All Density Index Scores by Diet



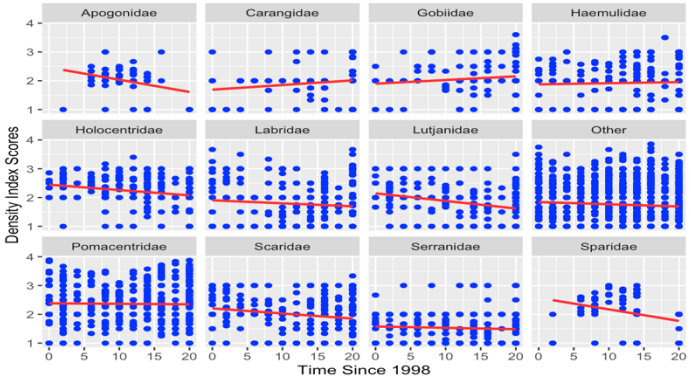
All Density Index Scores by Scientific Family



Non-Zero Density Index Scores by Diet



Non-Zero Density Index Scores by Scientific Family



Appendix G
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 H
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

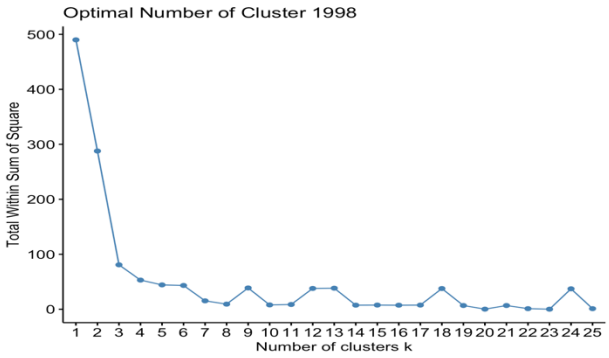
3. Non-Zero "DIET"

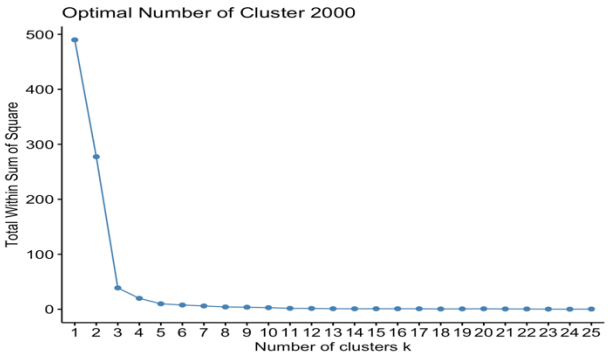
Non-Zero Summary Statistics for Species Primary Diets on Grand Cayman Island

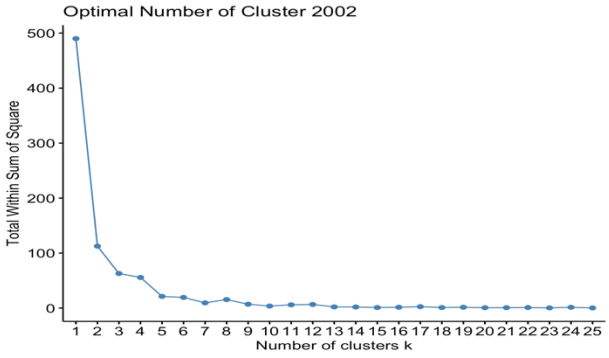
Diet	Mean Density Index	Standard Deviation	Standard Error	Sample Size
Algae	2.058689	0.6160146	0.0192036	1029
Coral	1.878836	0.5118014	0.0313217	267
Crustaceans	1.869881	0.6551331	0.0168986	1503
Echinoids	1.800451	0.7087813	0.0841169	71
Fish	1.640814	0.6391348	0.0197429	1048
Sponges	1.543108	0.5171518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

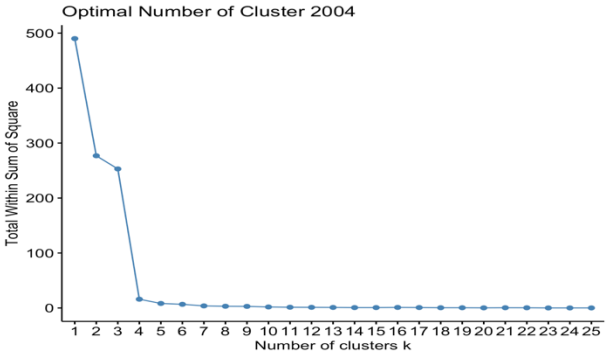
Appendix I
Optimal K Choice Plots for K-Means Clustering

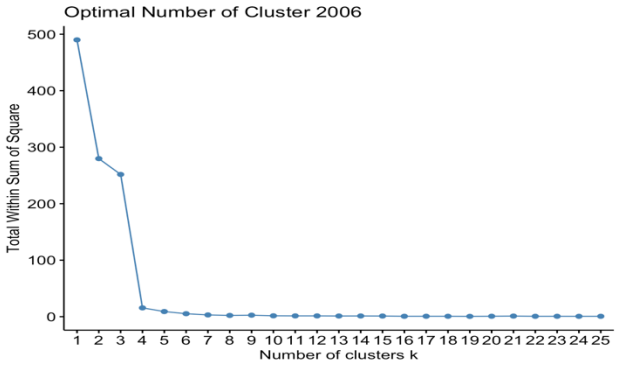
Optimization of the number of clusters, K, as determined by drop in total sum of squares is visualized in `fviz_nbclust()` plots. Optimization occurs at the elbow-like bends in the plots, which made K=4 a reasonable choice for clustering across all years.

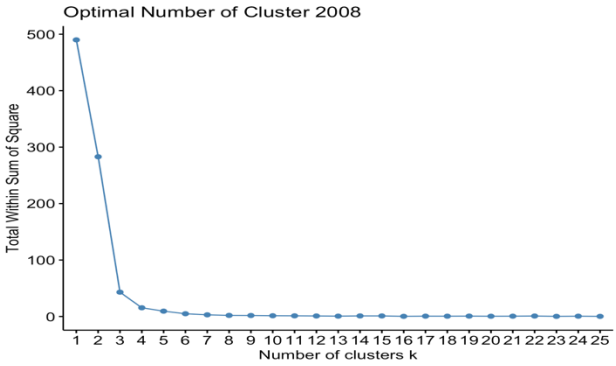


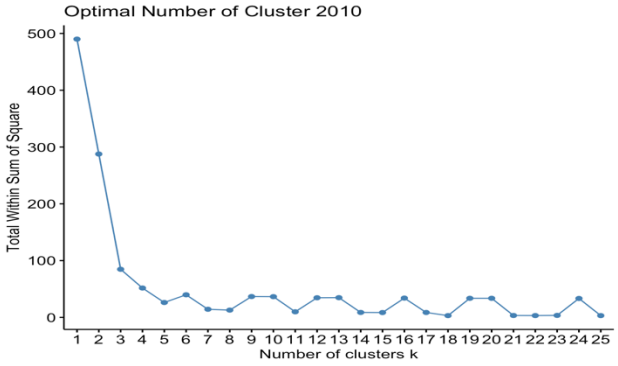


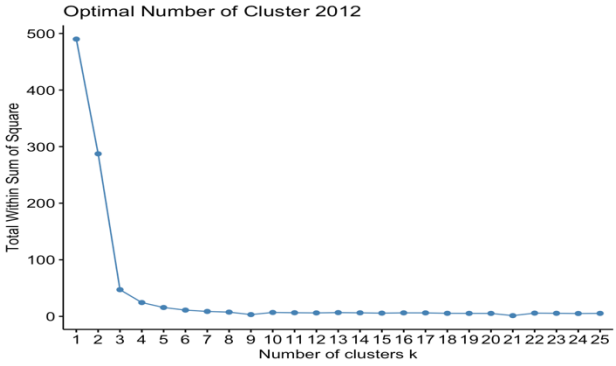






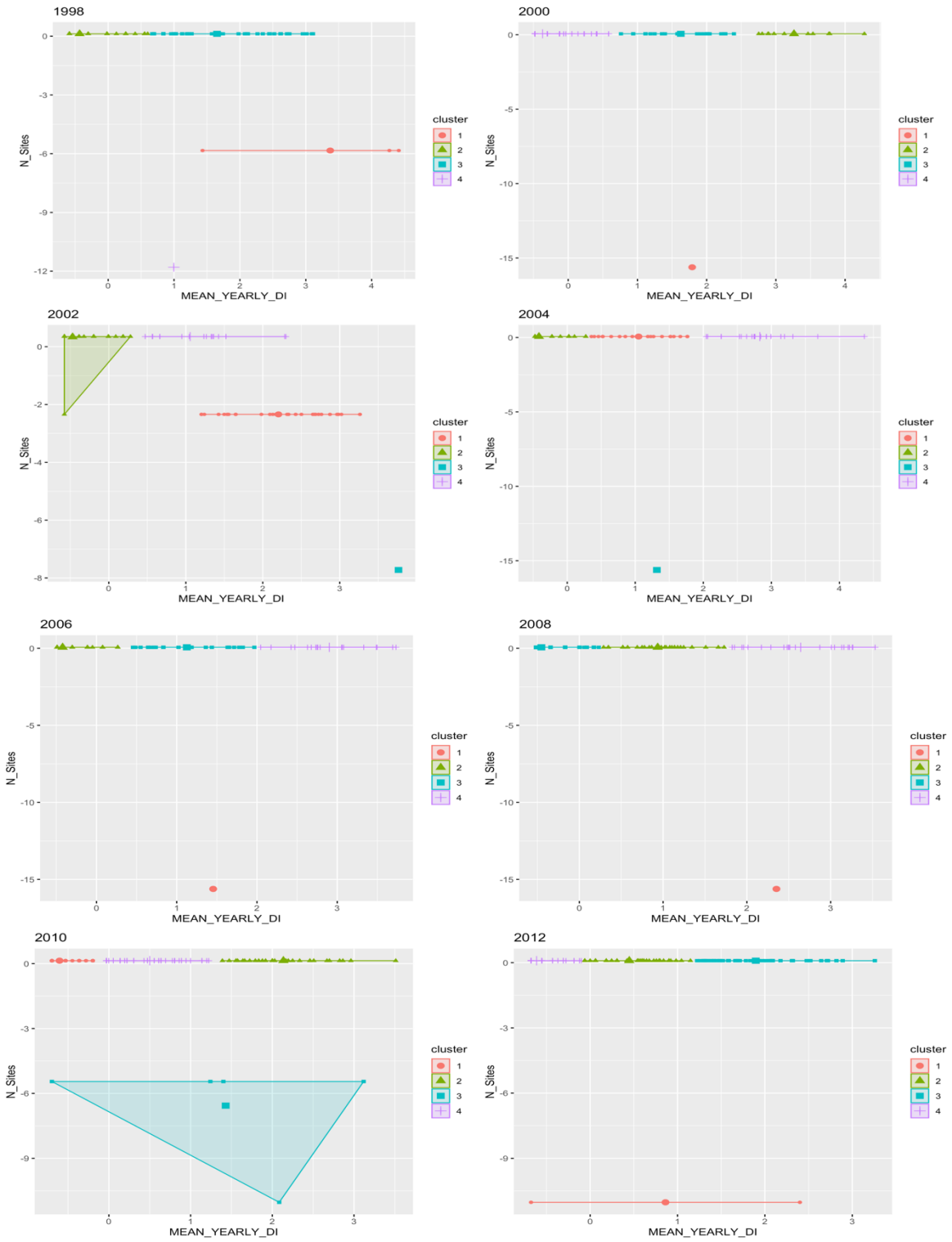






Appendix J
K-Means Clusters Visualization

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 K
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

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 L
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 M
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 N

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 O

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 P
Model Output of Potential Interaction Model

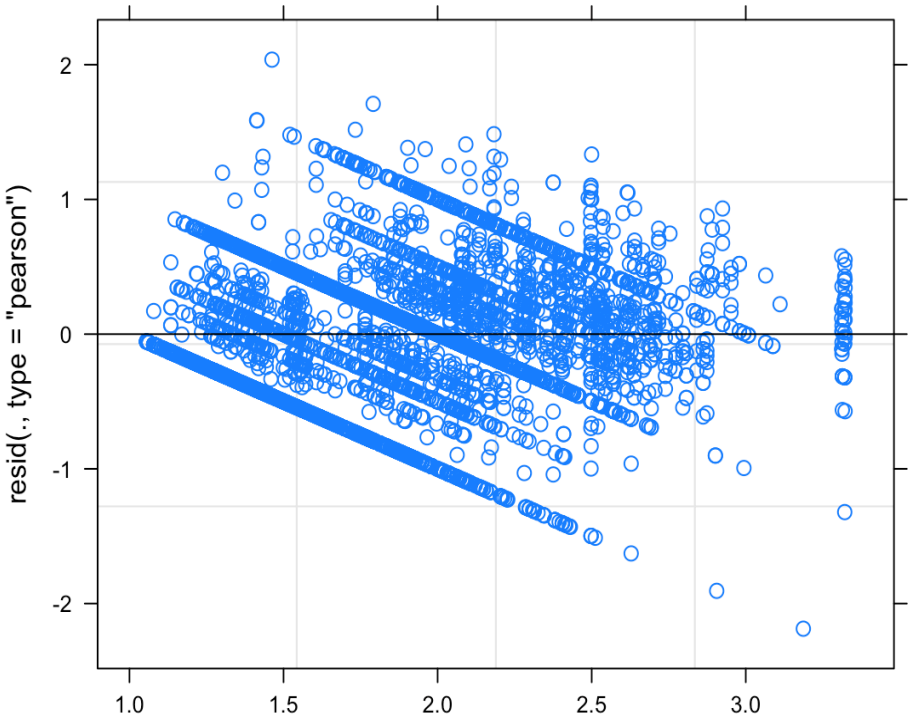
Output of the model testing for an interaction being “YEAR1998” and “SCIENTIFIC_FAMILY” is shown. Note the insignificance of model coefficients made drawing broad conclusions from this model difficult. Ultimately, the interaction was not included for interpretation, but should be used when modeling for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

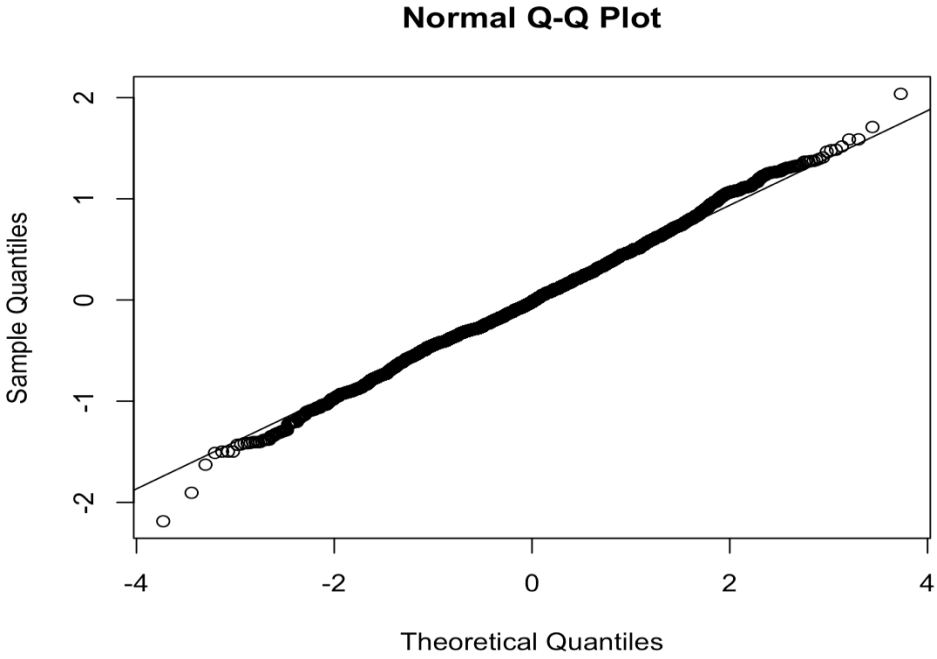
```
REML criterion at convergence: 8378.6

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 Q

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 R
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