

JMP Academic Case Study: 062

Monitoring Fish Abundance in the Mesoamerican Reef

Exploratory Data Analysis

Produced by

Melanie McField, Healthy Reefs for Healthy People
mcfield@healthyreefs.org

Ross Metusalem, JMP Statistical Discovery
ross.metusalem@jmp.com



Monitoring Fish Abundance in the Mesoamerican Reef

Data summarization and exploration

Key Ideas

This case study demonstrates exploratory data analysis in the context of wildlife monitoring and nature conservation. It pursues an exploratory line of questioning using univariate summary statistics, crosstabulation, interactive data filtering, and exploratory graphing with histograms, bar graphs, box plots, line graphs, scatter plots, density ellipses, and heat maps.

Background

The Mesoamerican Reef (MAR) is the largest barrier reef in the western hemisphere, spanning nearly 700 miles along the coasts of Mexico, Guatemala, Belize, and Honduras. The MAR hosts a multitude of species of fish, coral, and other marine life, including numerous critically endangered species. Millions of people in the region rely on the reef for food and for their livelihoods; unfortunately, overfishing, pollution, and climate change have taken their toll on this important ecosystem and present a threat to the health of marine life and people alike.



(A stoplight parrotfish swimming among sea whips.)

The Healthy Reefs for Healthy People Initiative (HRI; www.healthyreefs.org) is a multi-institutional effort dedicated to conserving the MAR by promoting the use of reef health indicators by policy makers and other leaders; analyzing and reporting scientific data to improve reef management; and fostering communication and networking among conservation partners. As part of this effort, HRI produces the [Mesoamerican Reef Report Cards](#), identifying current trends in reef health and suggested actions for improving it. These data come from the open Atlantic and Gulf Rapid Reefs Assessment database (www.agrra.org), which contains data on the biomass of various types of fish across the MAR, among other measures of reef health. Fish biomass represents the total mass of fish within a given area, with higher biomass values indicating that the reef is able to support a greater amount of fish life.



(A researcher surveying the Mesoamerican Reef.)

The Task

- Identify changes in fish biomass over time across different fish types and MAR subregions.
- Use summary statistics and exploratory graphs to find interesting trends and then investigate those trends to develop a clearer understanding of how fish biomass is changing in the MAR.

The Data **fish-biomass.jmp**

The file contains data from surveys of numerous reef subregions conducted from 2006 to 2018. For each year and subregion in the data set, we have biomass values for 20 types of fish measured in grams per 100 square meters. The individual biomass values represent averages taken over each individual site surveyed for the given subregion in the given year.

Year	Year that the data were collected.
Subregion	Geographic subregion from which the data were collected.
N Sites	Number of individual sites surveyed within the subregion.
Total	Sum total biomass (in g/100m ²) across all 20 fish types. Values represent the mean biomass across all sites measured in that subregion.
[20 fish cols]	Biomass (in g/100m ²) of the specific fish type indicated by the column name. Values represent the mean biomass across all sites measured in that subregion.

(Data are from the AGRRA database: Marks, K.W. and J.C. Lang. 2018. "AGRRA Summary Products, version (2018-03)." Available online <<http://www.agrra.org/data-explorer/explore-summary-products/>>)

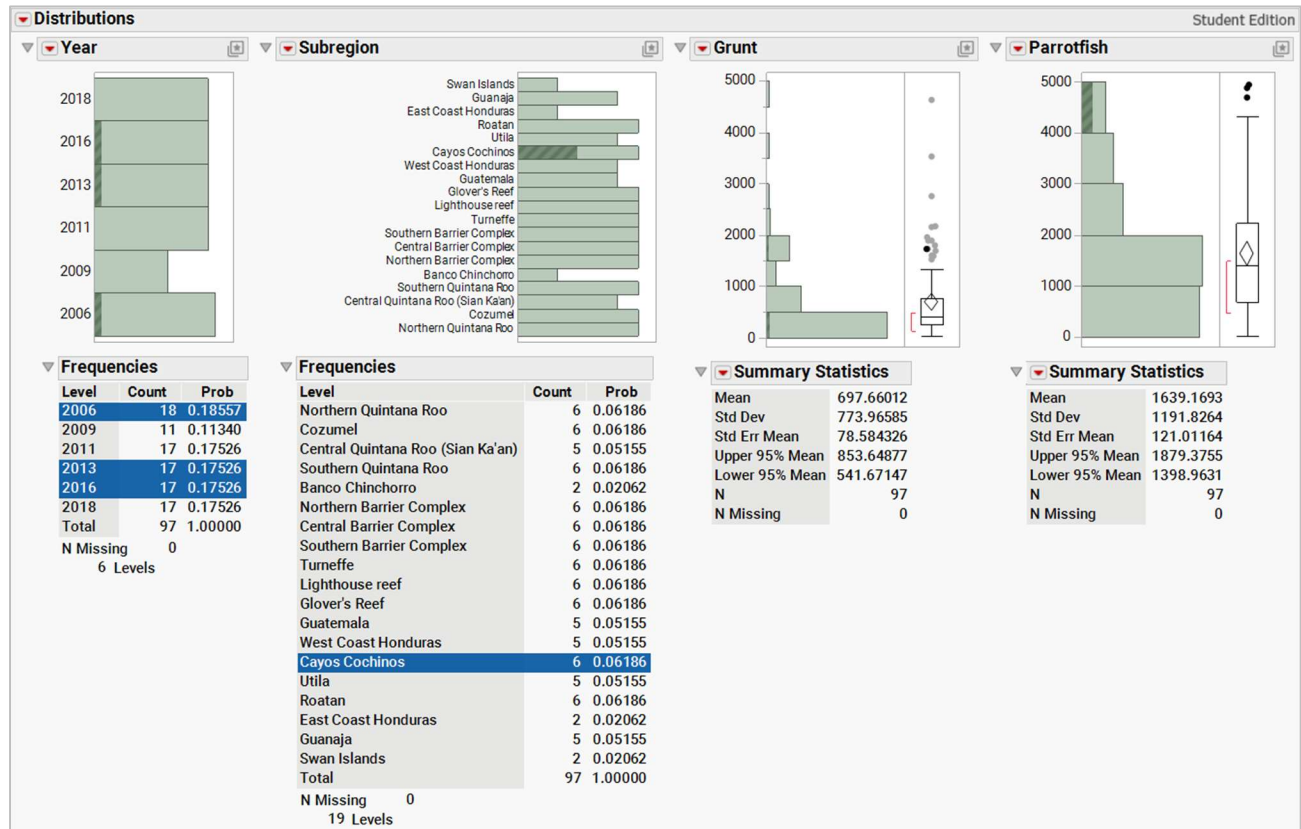
Analysis

Data summarization

Our goal is to identify important changes in fish biomass over time across the different reef subregions and fish types. Before we do that, though, we need to understand what data we have: How many different subregions are in the data set? Do we have data for each subregion in each year? Are there any notably high or low biomass values that may be worth investigating?

We begin by summarizing and graphing each variable in the data set using the Distribution platform. Exhibit 1 displays a subset of the variables. (In practice, we would look at all variables; we look at only four here to easily display the output on this page).

Exhibit 1 Summary statistics and graphs for Year, Subregion, Grunt, and Parrotfish. Outlier values for Parrotfish are selected. The report is truncated to fit on the page.



(Go to Analyze > Distribution, place all columns into the Y, Columns role, and click OK. Click and drag inside the Parrotfish outlier box plot to select the outlier values. All graphs are interactive, so click inside any graph to see where the selected data points lie in the other graphs.)

The report shows that we have six years of data across 19 subregions. The Count column in the Frequencies tables for Year and Subregion provides the number of rows in the data table for each level of these categorical variables, and the bar graphs represent these values visually. We can see variability in both Year and Subregion, meaning that not every subregion was observed every year. Looking at Year, we see that 11 subregions were surveyed in 2009, while 17 or 18 subregions were surveyed in all other years. Looking at Subregion, we see that all but three of the subregions were observed in at least five years. We'll further assess Year and Subregion in a moment, but first we'll use interactive graphing in this report to explore areas of high fish biomass.

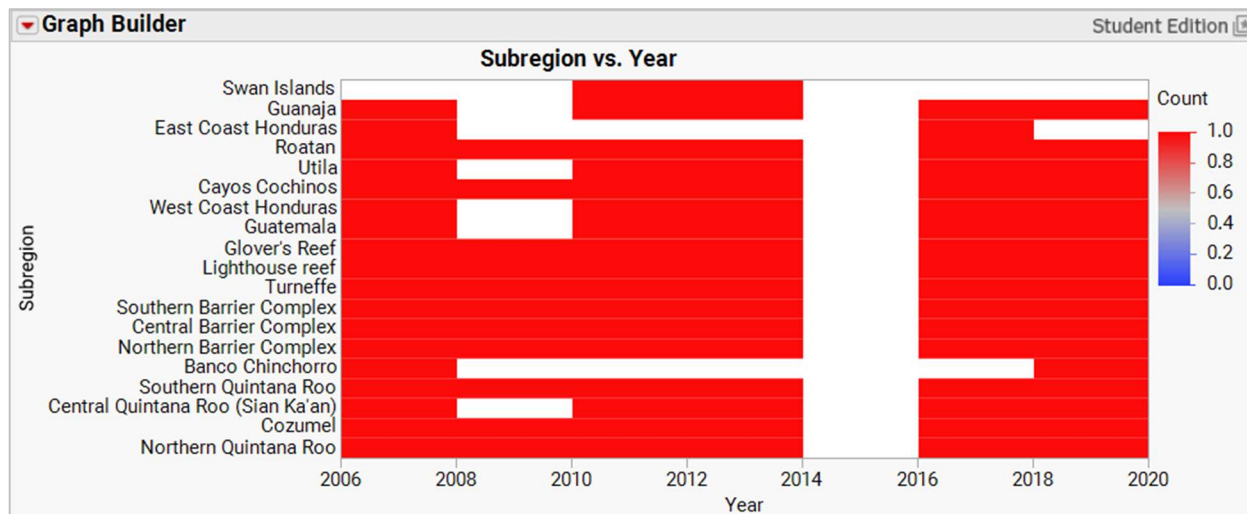
The histograms for Grunt and Parrotfish exhibit positive skew, with each showing several years of relatively high biomass. (These are the points that lie beyond the top whisker of the box plot.) Selecting the three outliers for Parrotfish as shown in Exhibit 1, we see that each of these high biomass values was observed in the Cayos Cochinos subregion off the coast of Honduras. The Honduran government protected Parrotfish in 2010, so perhaps this is a result of that effort. (Parrotfish have also been protected by other Central American countries: Belize in 2009, Guatemala in 2015, and Mexico in 2018.)

Because we have found variability in the number of subregions measured each year, we next make a graph of Year by Subregion to see, for each subregion, which years we have data for. The graph in Exhibit 2, a heat map, has one cell for each possible combination of Year and Subregion, with the cell colored in red if we have data for the given subregion in the given year and white if not. We see, for example, that for the Swan Islands subregion we have data only in 2011 and 2013, giving us a restricted

time window to assess biomass changes over time at that location. The other subregions show a longer time between the first and most recent measurements, so we should be able to assess biomass changes over longer time periods at those subregions.

Having learned more about the data we have available, as well as having uncovered a hint of high Parrotfish biomass in Cayos Cochinos, we'll move on to exploring how fish biomass changed over time and subregion.

Exhibit 2 Heat map of Year and Subregion



(Go to Graph > Graph Builder. Ensure the modeling type for Year is set to Ordinal. Drag Year to the X axis and Subregion to the Y axis. Select the heat map element (📊) in the ribbon of graph elements and click Done.)

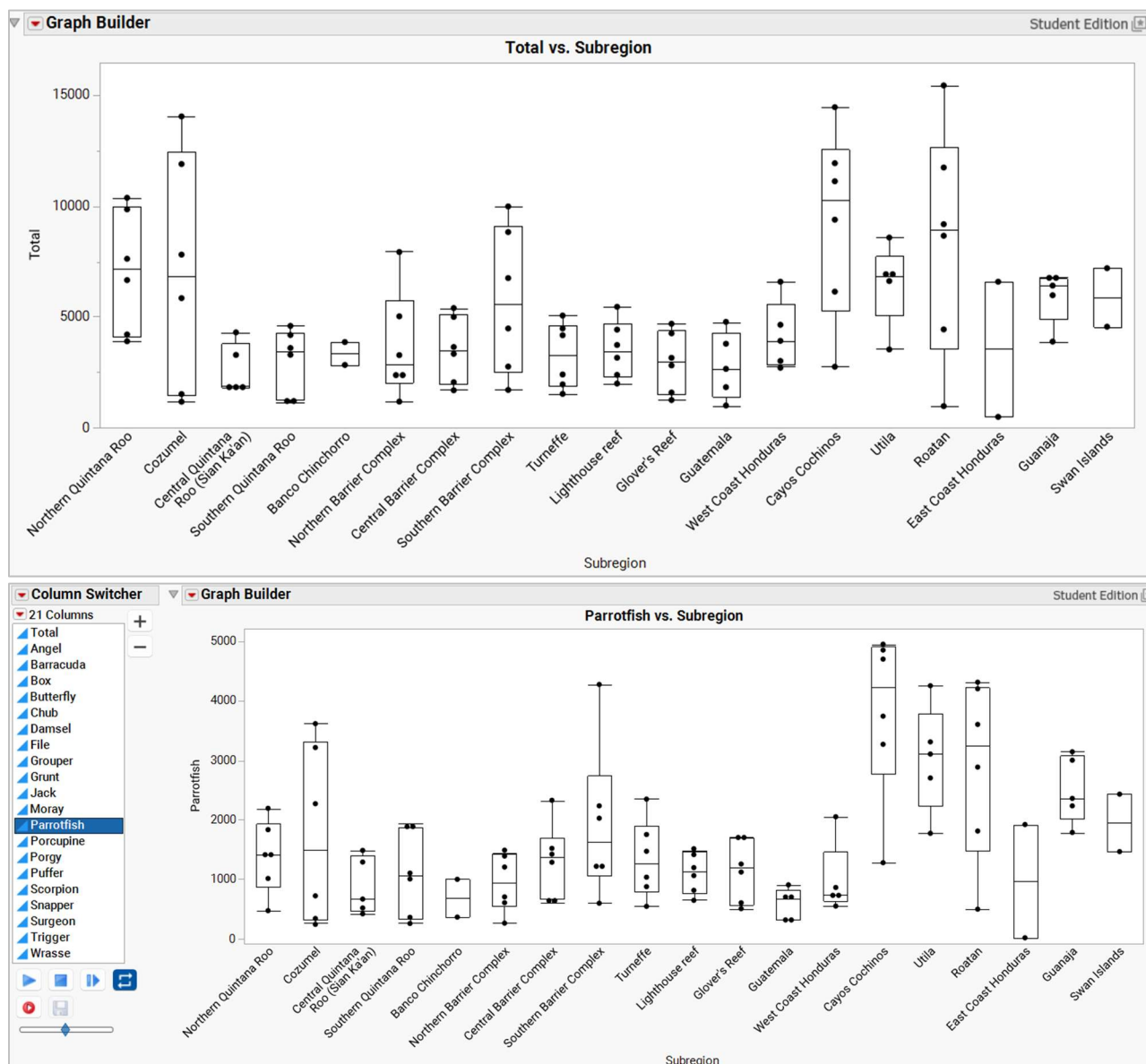
Exploring fish biomass across subregions

We continue by exploring fish biomass across the different subregions, looking for any interesting patterns: for example, a subregion with notably high or variable biomass or a certain fish type showing high biomass variability within a specific subregion, which would suggest fluctuation over time.

We create the box plots in Exhibit 3, which show total and individual fish biomass by subregion. In the top panel, we see that Cayos Cochinos has the highest median biomass at 10,248 g/100m². (We found this value by hovering our pointer over the box plot for Cayos Cochinos, as shown in Exhibit 3.) We also see that Cayos Cochinos, Cozumel, and Roatan are highly variable, with low values among the lowest in the data set and high values among the highest in the data set. Substantial change over time must have occurred in these subregions. (We'll investigate this in the next section.)

We recall from Exhibit 1 that Cayos Cochinos showed several high Parrotfish biomass values, so we next plot Parrotfish biomass in place of total biomass. At the bottom of Exhibit 3, we use the Column Switcher to change which biomass measure is plotted on the Y axis. We select Parrotfish and again see high variability at Cayos Cochinos, Cozumel, Roatan, suggesting that changes in Parrotfish biomass may be a primary driver of the total biomass variability we see in these subregions. In practice, we might use the Column Switcher to explore other fish types, but for now we'll move forward with a focus on trends over time in total and Parrotfish biomass in Cayos Cochinos, Cozumel, and Roatan.

Exhibit 3 Box plots of Total biomass and Parrotfish biomass across Subregion

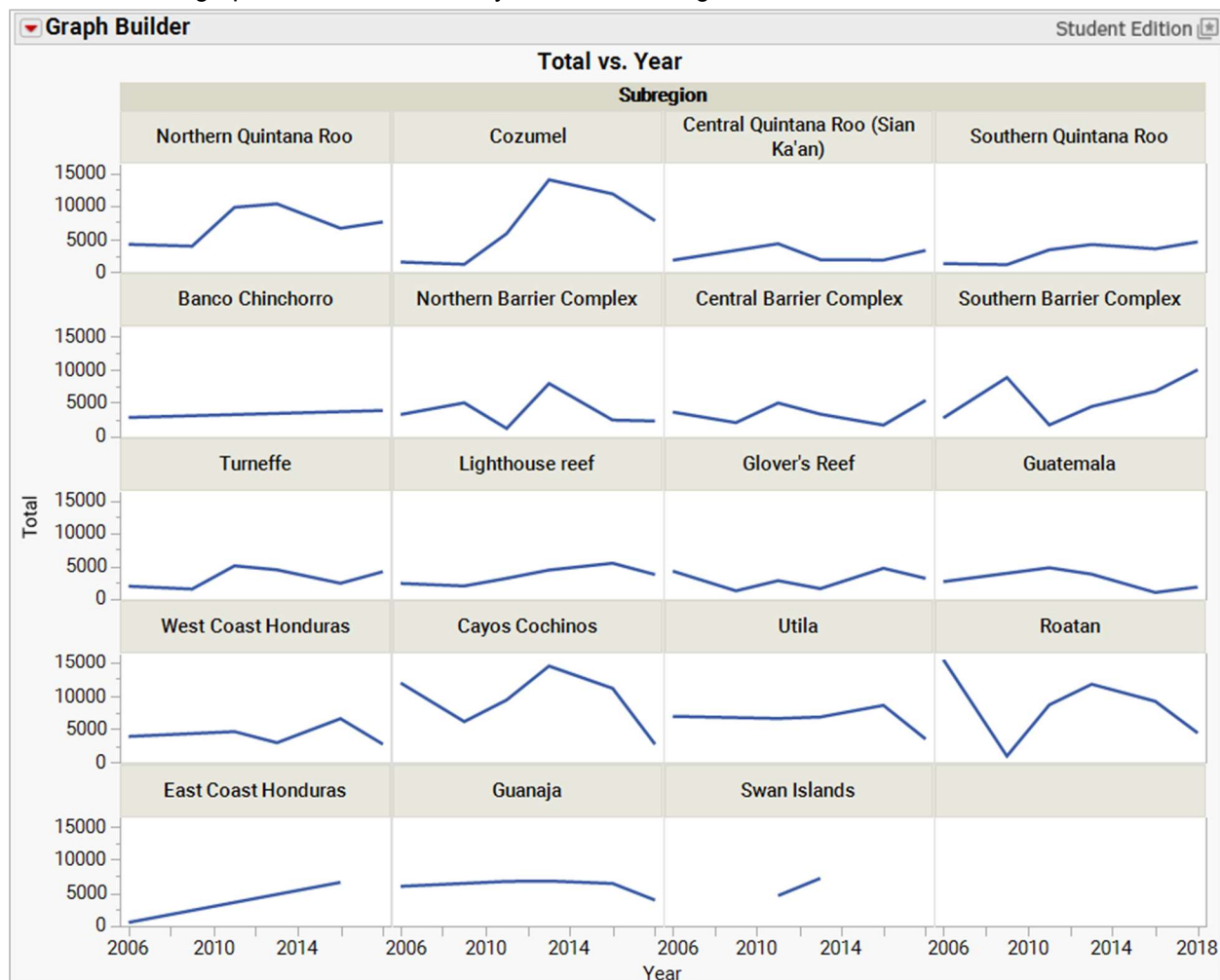


(Go to Graph > Graph Builder and drag Subregion to the X axis and Total to the Y axis. Drag the box element into the graph to add the boxes over the points and click Done. To invoke the Column Switcher, go to the red triangle > Redo > Column Switcher, select Total as the column to switch and then all fish biomass columns as the replacement columns. Hover the arrow over a point to bring up the hover label as in the top panel.)

Exploring biomass change over time

Exhibit 4 shows line graphs of total fish biomass across time for each subregion. Note that we have changed the modeling type of Year to Continuous so that the space between each tick mark on the X axis represents an equal unit of time, which would not be the case if Year were Ordinal.

Exhibit 4 Line graphs of Total biomass by Year and Subregion



(Change the modeling type of Year to Continuous by right clicking on the green bar icon next to Year in the column list and selecting Continuous. Go to Graph > Graph Builder and drag Year to the X axis and Total to the Y axis. Next, select the line element. Drag Subregion to the Wrap zone and click Done.)

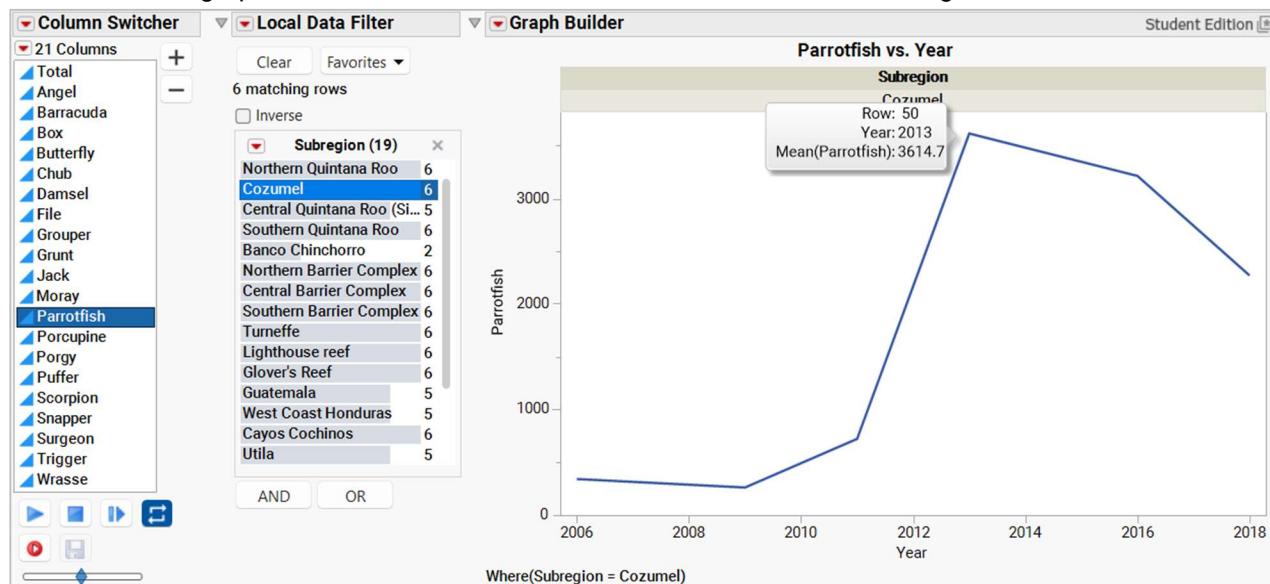
Cayos Cochinos, Cozumel, and Roatan all show high variability as we expect from the box plots. We look first at Cayos Cochinos: Is its biomass rising or falling? We see a sudden drop in biomass in 2009 before a sharp rise through 2013 and subsequent decline. We've learned that while Cayos Cochinos showed the highest median Total biomass, it recently has been in decline. We see a similar trend over time for Roatan.

The pattern in Cozumel looks different. This subregion began very low in total biomass before a steep increase from 2009 to 2013, ultimately ending 2018 higher than it was in 2006, despite a decline from its 2013 peak. We decide to investigate this trend further. Could Parrotfish biomass be playing a critical role?

In Exhibit 5, we use the Local Data Filter to drill down into the Cozumel subregion, and we use the Column Switcher as in Exhibit 3 to cycle through biomass for individual fish types. Looking at Parrotfish, we see a trend that looks much like that for total biomass: a sharp increase from 2009 to 2013, followed by a moderate decline. The biomass values for Parrotfish here are a relatively large proportion of total biomass, too; for example, they peak in 2013 at over 3,600 g/100m², which is approximately 25% of the total biomass observed in Cozumel that year. We conclude that the increase in total biomass at Cozumel was driven in part by the trend we've uncovered for Parrotfish.

Cycling through the other fish types, we see that Grunt, Snapper, and Surgeon show similar trends to that of Parrotfish, albeit with slightly lower magnitudes. These fishes also seem to have played a role in the increase in total biomass in Cozumel, but perhaps not as importantly as that of the Parrotfish.

Exhibit 5 Line graph of Parrotfish biomass across time for the Cozumel subregion

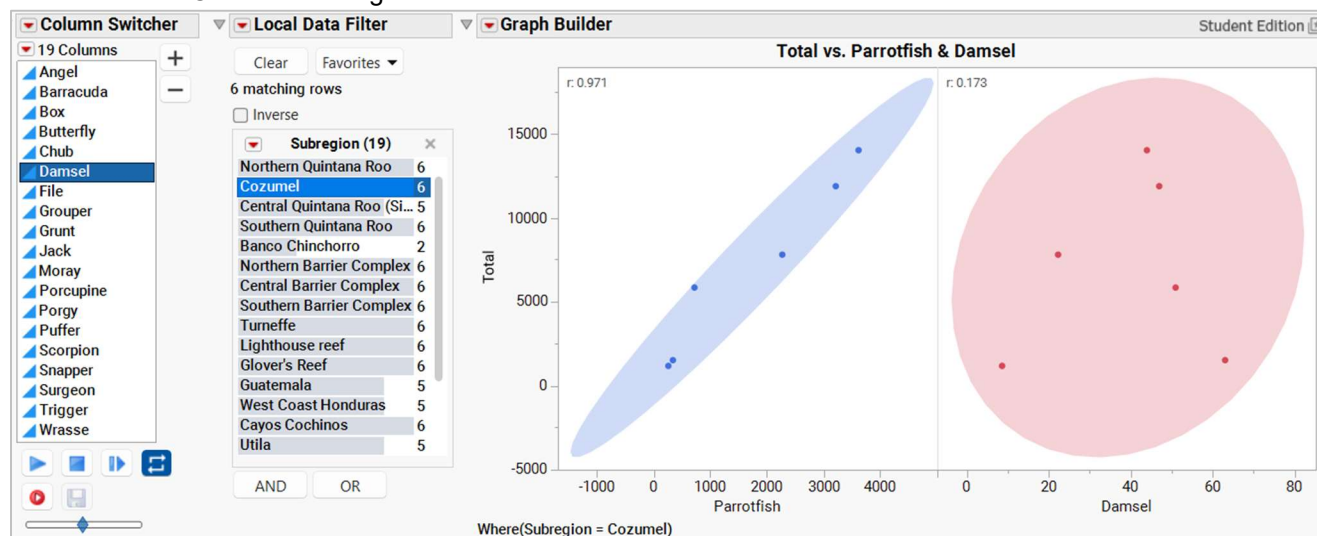


(Starting with the graph from Exhibit 4, go the red triangle > Local Data Filter. Select Subregion from the list of columns, click the plus sign, and then select Cozumel from the list. Invoke the Column Switcher as you did in Exhibit 3 and then select Parrotfish from the list of fish types. Hover over the line at 2013 to reveal the depicted hover label.)

We now believe Parrotfish may be a primary driver of Cozumel's trend in total biomass over time. If this is true, we'd expect to see high correlation between Total and Parrotfish biomass, suggesting that they track together tightly. Exhibit 6 depicts these correlations. The correlation coefficient (r) in the top left of each pane represents the strength of the relationship, with values closer to 1 indicating a strong positive relationship and -1 indicating a strong negative relationship. The shaded density ellipses provide a visual depiction of the correlation, with more elongated ellipses indicating stronger correlation. We see a correlation of Total to Parrotfish of 0.971, indicating that as Parrotfish biomass increases or decreases, we tend to see a clear corresponding increase or decrease in total fish biomass.

We next check to see if any fish exhibits a stronger correlation with total biomass than does Parrotfish. We add a second fish to the X axis and use the Column Switcher to cycle through the other fish and compare their correlation with that of the Parrotfish. Several fishes show correlation values above 0.8, but none is as highly correlated with Total as Parrotfish. Some fishes also show very low correlation with Total: Damsel (depicted in Exhibit 6), Porgy, Puffer, and Scorpion. These fishes generally have low biomass values, so it is unsurprising that they are not highly correlated with total fish biomass; their influences on total biomass are minor compared to those of the other fishes.

Exhibit 6 Scatter plot, density ellipse, and correlation of Total vs Parrotfish and Total vs Damsel biomass in the Cozumel subregion



(Go to Graph > Graph Builder and drag Total to the Y axis and Parrotfish to the X axis. Select the Ellipse element (). In the Ellipse element options to the left, select Correlation. Drag any other fish to the rightmost X axis drop zone to create a second scatter plot to the right of the first one. Use the Local Data Filter as in Exhibit 5 to drill down into only the data from the Cozumel subregion. Use the Column Switcher to cycle through different fishes in the rightmost pane.)

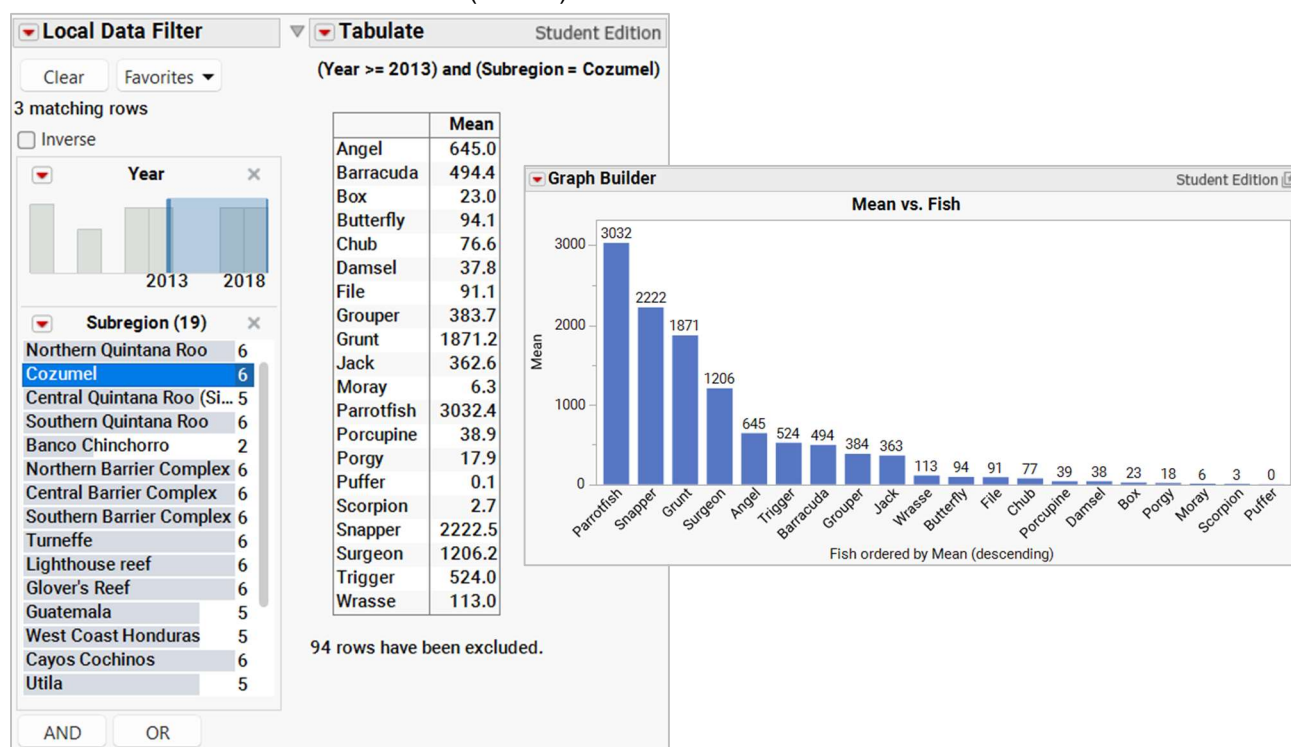
Identifying the most abundant fish in the Cozumel subregion

We have found that the Cozumel subregion has increased in total biomass between 2006 and 2018. This appears due in large part to an increase in Parrotfish biomass. We'll finish this line of exploration by determining which fish types have been the most abundant overall in Cozumel in the time period following the biomass increase starting in 2009; after all, the fish that showed the biggest increase in biomass is not necessarily the most abundant overall.

In Exhibit 7, we use a crosstabulation to calculate mean biomass for each fish type from 2013 through 2018 in the Cozumel subregion. The most abundant fishes in this time window are the Grunt, Parrotfish, Snapper, and Surgeon, which aligns with our findings in Exhibit 5. However, these biomass values may be difficult to understand by using a numeric table alone, so in Exhibit 7 we have also created an ordered bar chart of the crosstabulation values to illustrate more clearly just how much more abundant these four types of fish are than the others. Clearly, these four were far more abundant than others in the Cozumel subregion from 2013 to 2018. However, we also see a second group of moderately abundant fish: Angel, Trigger, Barracuda, Grouper, and Jack. We didn't notice these in our crosstabulation table but were able to notice them by using a graph.

We draw several conclusions from this line of data exploration. First, the Cayos Cochinos, Roatan, and Cozumel subregions showed high variability in fish biomass over time, with Cozumel showing the more promising trend of growth. Changes in Parrotfish biomass seem to be a primary driver of this trend. Following the large overall biomass increase in Cozumel starting in 2009, we find that the most abundant fish are the Parrotfish, Snapper, Grunt, and Surgeon.

Exhibit 7 Crosstabulation of mean biomass in the Cozumel subregion 2013-2018 (top) and an ordered bar chart of the crosstabulation values (bottom).



(To create the crosstabulation table, go to Analyze > Tabulate and drag Mean from the list of statistics to the Drop Zone for Columns, drag all of the individual fish biomass columns to the Drop Zone for Rows, and click Done. Then use the Local Data Filter to drill down to 2013-2018 and Cozumel. To create the bar chart, inside the Tabulate window go to the red triangle > Make into Data Table; in the resulting data table, change the column headers to Fish and Mean Biomass. Go to Graph > Graph Builder and drag Fish to the X axis and Mean Biomass to the Y axis, and then right click on the X axis and select Order By > Mean Biomass, Descending. Finally, select the bar element (), show labels and click Done.)

Summary

Statistical insights

We explored the fish abundance data using a combination of graphs and summary statistics to identify potentially important patterns in the biomass of various fishes over time and across different subregions of the Mesoamerican Reef. Our exploratory analysis was fluid, starting with relatively high-level questions and data visualizations to answer them, and then using what we found to form further questions. The exploratory process ultimately led us to focus on the Cozumel subregion and several types of fish, most notably the Parrotfish. We arrived at a clearer understanding of how total fish biomass has increased in the Cozumel subregion and which specific fish have contributed most greatly to this increase.

Implications

Exploratory data analysis is an open-ended inquiry process. We search for meaningful patterns in the data and use them to form and answer subsequent questions. The fluidity of the process requires us to use whichever data analysis and visualization tools we need at the time. In this case study, we used many different graphs and analyses: bar charts, histograms, heat maps, box plots, line graphs, density ellipses, means, medians, and correlations. We also saw how exploratory data analysis involves drilling down into data subsets and considering many different variables when looking for patterns.

JMP features and hints

This case study primarily used Graph Builder, a general-purpose interactive graphing tool. It also used the Distribution platform for initially summarizing the data and the Tabulate platform for performing crosstabulations. We also made extensive use of the Column Switcher and Local Data Filter, two tools found throughout JMP that are useful for efficiently updating graphs and analyses in real time by cycling through different variables and drilling down into data subsets.

Exercises

Use the **fish-biomass.jmp** data set to answer the following questions. Note that there may be multiple ways (graphs, crosstabulations, data summarizations) to arrive at your answer for each question.

1. We found that Parrotfish were the most abundant in the Cozumel subregion from 2013-2018. Still, it's possible that other subregions were more abundant in Parrotfish than was the Cozumel subregion. Which subregion(s), if any, show a higher mean Parrotfish abundance than the Cozumel subregion in the 2013-2018 time period?
2. As Parrotfish biomass fluctuated in the Cozumel subregion, other fish likely fluctuated along with it. Across all years, which fish shows the strongest positive correlation with Parrotfish in the Cozumel subregion? Are there any fish(es) that show a *negative* correlation with Parrotfish in the Cozumel subregion? If so, which one(s)?
3. We found evidence that Snapper was also among the drivers of the total biomass increase at Cozumel. Looking across all subregions and years, are there any high outliers of Snapper biomass values? If so, in which subregions were these values observed?
4. Exhibits 3 and 4 showed that the Roatan subregion had high total biomass variability. Between Parrotfish and Snapper, which seems like the more important driver of the total biomass trend over time at Roatan? Explain your reasoning and provide a graph to support it.
5. We've focused on the more abundant fish, but it is important to attend to the less abundant fish, as well. If you sum each fish's biomass across all subregions, which was the least abundant overall in 2018? (Hint: Analyze > Tabulate could be useful here, but it's not the only way to get the answer.)