Course Project

Analysis of Benthic Marine Life in the Gulf of Mexico around the Deep Water Horizon Oil Spill

DSP 539: Big Data Analysis

Instructor: Rachel Schwartz

Ian Poe

100677280

Date of Submission

April 30, 2025

Contents

List of Figures

1 Introduction

The Gulf of Mexico is home to a large variety of marine life. The access to warm waters and an abundance of structure provides marine life with area to thrive, breed, and grow. With the large marine life that is present in the Gulf, it is important to understand how actions we take may impact the biodiversity of these regions. One of the best examples of human impact on the marine life in the Gulf is through the many oil rigs located within this region. These rigs use drilling to extract the oil beneath the seafloor. These rigs can have both positive and negative effects on the ecosystem. The added structure creates a type of artificial reef where an increase of marine life can be found. However, if these oil drills are not correctly maintained or operated, an oil spill can be caused which can be detrimental to marine life taking years for the environment to recover. One of the most famous examples of an oil spill occurred on April 20, 2010 at the Deep Water Horizon drill site. Oil continued to spill into the Gulf until they were able to place a cap on the drill line. "Before it was capped three months later, approximately 134 million gallons of oil had spilled into the Gulf" (DARRP, 2025). After the spill, many organizations took an interest in how these events could have affected the marine life. Using public fisheries data from the region, we can determine what kind of effect the oil spill had on fish in the benthic zone. As the oil spills into the Gulf, it raises to the surface, directly impacting marine birds and other marine life. However, this does not tell us what occurs at the bottom of the Gulf. This project will focus on the effect of the oil spill on benthic fish in order to determine if a significant change occurred as a result of the oil spill.

2 Data Description

The National Oceanic and Atmospheric Administration, Department of Commerce (NOAA) accumulated a variety of data in order to see effects of the oil spill on marine life. For this project, two main types of data will be used. The first of which is collected by Natural Resources Damage Assessment (NRDA) in order to determine a baseline for the benthic marine life of the Gulf. This data consists of 282 variables and 13,248 observation. This information provides data collected from bottom trawling surveys from 1987 to 2010. The dateset includes variables for geographical coordinates, unique trip information, time of servery, type of equipment, time spent trawling, and the total amount of live specimens collected. In addition, every species collected contains a variable for both the amount found and their weight. The second type of data is long-line fishery data. This data is recorded in multiple datasets. Each dataset provides slightly different information. For this project, the data for station and catch will be used. This provides geographical coordinates, basic site information, time of collection, Type of Species, Number and Weight. Using these datasets, an analysis

will be performed in order to determine if and how the oil spill has affected the benthic marine life in the region.

3 Analysis

3.1 Question 1: Using Bottom Trawling Data, is Marine Life Affected by the Oil Spill?

The data collected by the NRDA is an accumulation of Bottom Trawl Surveys from 1987 to 2010. This dataset contains a small proportion of data which was collected after the oil spill. In order to determine a significant difference in marine life, the data variables related to the amount of live specimens sample: overall, finfish only, crustation only, and for other marine line is used. This data then must be scaled relative to trawling time. To solve this scaling issue, the variable related to time trawling is used to scale the variable associated with amount of live specimens sampled. Once these results were summarized, the data was grouped by the date taking the mean of all observations which occurred on the same day. This allows for a single point to be found for any given day and these results can be graphed to see how the the different groups change over time. This figure can be seen in the R-markdown code provided with this document. A relatively flat trend line is found; however, there are clear seasonal changes occurring within the data along with larger population cycles occurring over many years. In order to compare the data before and after the oil spill, a bar graph is created. This bar graph compares data from 2009 and 2010 in order to minimize population trends. To control for seasonal trends, the data was converted to day of the year where the data from 2010 is overlapped on the data from 2009. Figure 1 shows the four bar graphs comparing the mean total live for each group over the days of the year. From the bar graph it is difficult to determine a significant difference as data is missing from parts of the year and the difference that is seen could be caused by normal population changes. In order to determine significance, two linear models were created. The first linear model was a full model which included a variety of variables in order to determine significance. This model used total catch as a response with the longitude, latitude, minutes fished, depth, time of day, mesh size, tows, vessel speed, month, year, and is_2010 as variables. By adding a binary dummy variable for the 2010 data we can see if this data is significantly different that the other years. From this linear model, a low r-squared (0.1321) was found which means the model does not fit the data very well. However, eight of the explanatory variables were found to be significant. The variable is_2010 was found to be significant with (p-value = 7.00e-10) with a coefficient of 2.953e1. This regression indicates that after the spill there is an increase in benthic marine life. To ensure the variable is_2010 was not being influenced by

any strong associations with other explanatory variables, a second linear model was fitted using total catch as the response variable and only is_2010 as the explanatory. This linear model also concluded significance of the year 2010 with a positive coefficient. The regression can be seen in the r-markdown file attached.

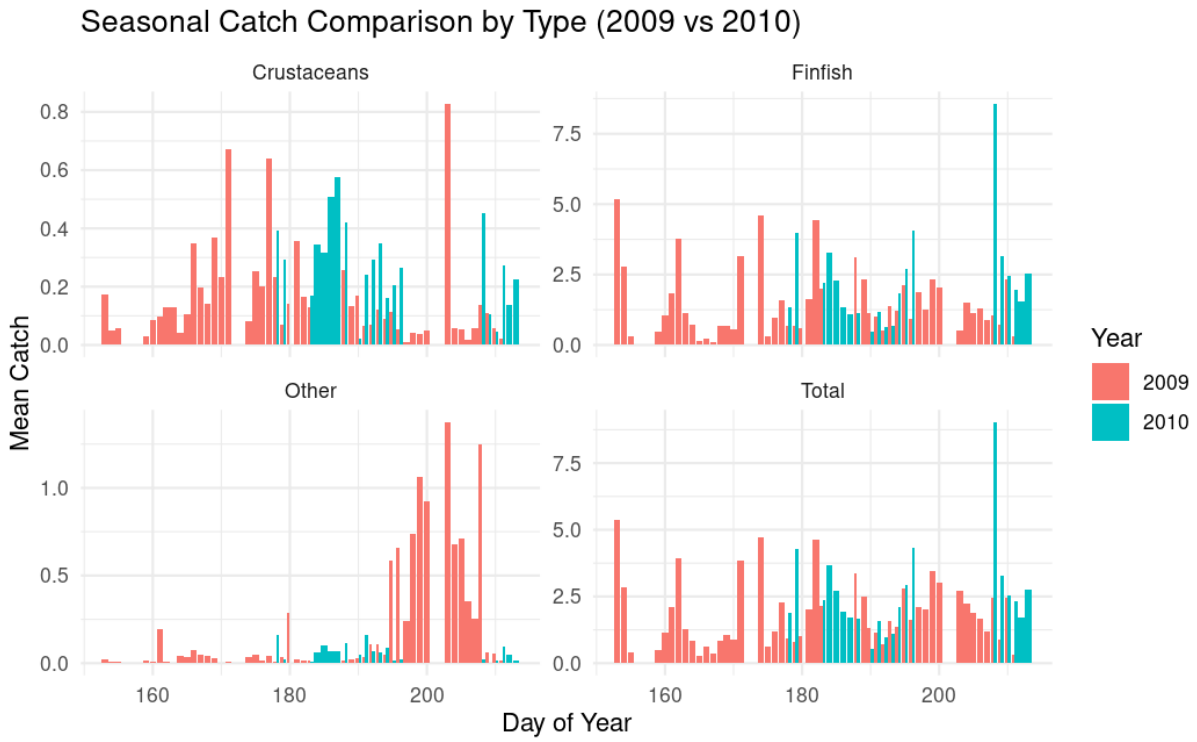


Figure 1: The mean of Live Specimens for each group in the Gulf show no meaningful change while observing between 2009 and 2010. Any differences in the data can be caused by lack of information or normal population changes between the years. The use of bar graphs, and plotting by day of the year minimizes seasonal and population trends which could occur. A regression model was fitted to the data which found a positive association with a p-value of 7.00e-10 and a r-squared value of 0.1321. Concluding the data from 2010 is significantly higher than the rest of the data.

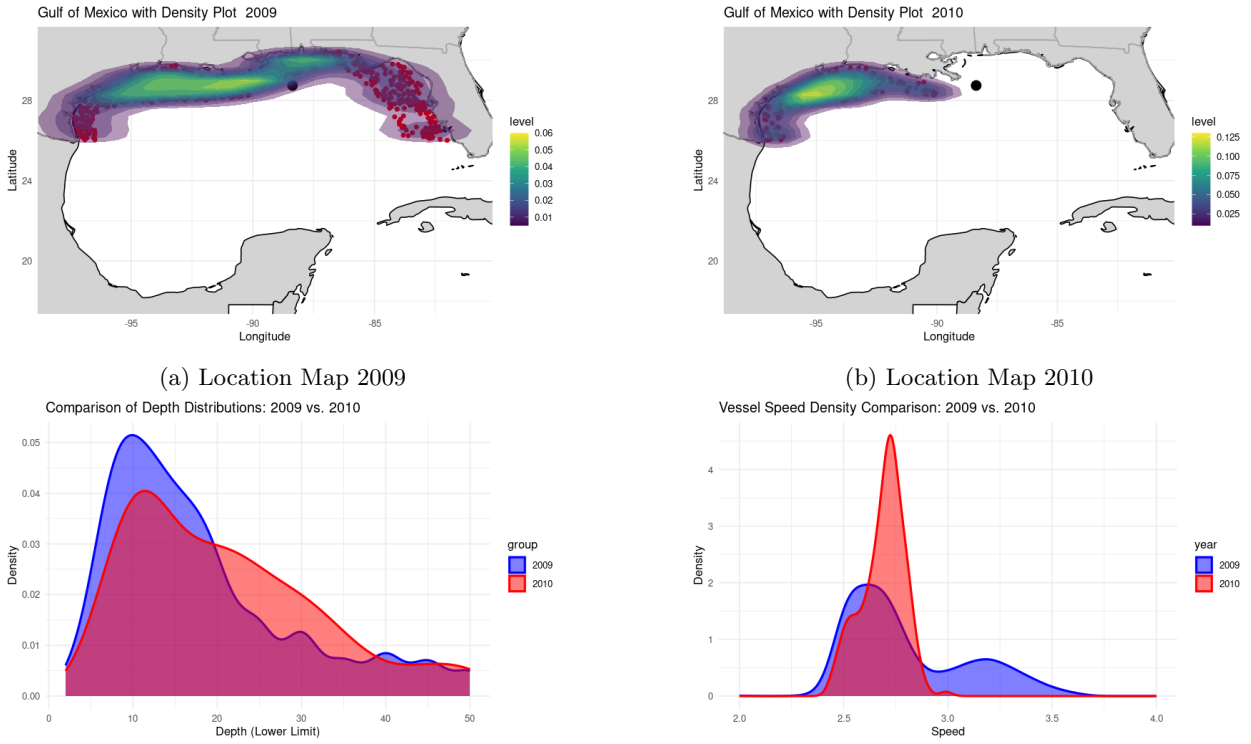
3.2 Question 2: Are we Able to Compare the Data from 2010 to the Data of Other Years?

The results found in the previous section indicate an increase in benthic marine life after the oil spill. To determine if this analysis is a valid assessment of the population the sample, we must show it comes from the same population as the rest of the data. To check if this is true, three functions were created in order to compare the location of samples, the density of depth distributions, and the density of speed distributions. In an ideal world, if all of these figure look alike, then a general statement can be made that it is a sample of the true population. The first function location-per-year takes two inputs. First is the original data and the second input is the year of interest. By filtering by specific year and isolating the variable corresponding

to the geographic coordinates, spatial analysis library's in r can be used to create a map and density plot to find areas of importance. Through the use of the "sp" and the "maps" library in r, the Map of the Gulf of Mexico was plotted. A point coded into the map plots the location where the DWH Oil Rig was located. Then the sample location was marked in red and a density plot is overlayed. To determine if 2010 was representative of the true population, it was compared to the prior year as seen in Figure 2a and 2b. It can be seen that the data from 2010 only shows a small portion of what is typically sampled. Next, a function to compare the density distribution of depth where created. This function depth-per-yer-density is a function which takes three inputs: data, year1 and year2. This function creates 2 different datasets from the original filtered by each respective year. The depth variable is a string of 4 values: the first 2 correspond to the lower limit in fathoms as the upper 2 are the upper limit in fathoms. This function plots the density of the lower limit, so the first 2 values are used from the depth variable. A data frame is then created which includes the data from both years and a single figure with both density plots is created. This figure can be seen in 2c, plotting 2009 to 2010 we can see the 2009 data was sampled more from a 5-20 fathom depth while 2010 had a larger proportion from 20 to 40 fathoms. The final comparison is the speed of sampling. A function speed-per-year-density is created which also takes the same three inputs was above. The same method was used (Speed Variable Did Not Need to be Modified) in order to obtain the figure with the two density plots. Comparing 2009 to 2010 (figure 2d), the data from 2009 was more spread out with a peak of about 2.6 while 2010 has very little variation with a large spike around 2.7. Due to the differences in these figures, the conclusion is the data from 2010 does not represent the true population of benthic marine life sampled in previous years. As a result, the results obtained in Question 1 should be reevaluated with more data or a different data set.

3.3 Question 3: Using the Bottom Trawling Dataset as a Baseline; Can We Identify Species of Interest and Apply it to the Long-line Fishery Dataset?

Species of interest are defined as the species with the greatest growth and greatest decline between years. A function was created to identify the Species of Interest named Species-Greatest-Change which takes three inputs: the original data, a reference year (not 2009) and a value of species for increasing and decreasing. In the previous question, we determined data from 2010 should be removed from this dataset as it is not representative. The data is then filtered which takes the data from the reference year all the way to 2009 the new end of the data. By grouping the data by year, the mean of each species was found by identifying columns that start with "N_". A new data frame is created which shows the change from year to year of each species and for loop is used to find the species with the maximum difference from the reference year to



(a) Location Map 2009

(b) Location Map 2010

(c) Depth Density Distribution

(d) Speed Density Distribution

Figure 2: These Figures show that the data compared from 2010 is not a reasonable representation of the data that was collected from the rest of the dataset. The map comparison shows a clear difference as the data from 2010 where all sampled from Northwest Region of the Gulf as compared to the entire norther section of the Gulf normally sampled. In addition, the depth density plot shows difference in both the depth of the sampled regions. 2009 had an increased sampling region in shallow water 5 to 20 fathoms while 2010 had a increase in sampling from 20 to 35 fathoms. Additionally, the speed on the ship while sampling in 2009 was more spread out with a peak of about 2.6 knots while in 2010 there is very little variation with a large spike around 2.7 and very little spread.

the end of the dataset. Two list are created: one for increasing and one for decreasing which outputs the top n species for each side. Finally, the lists are modified to only include the scientific name. These lists can be seen in the r-markdown attached. Using the list of species obtained from the species of interest function, we want to apply these fish to the long-line fishery dataset. The data is collected by leaving a line of hooks at the bottom of the ocean and seeing what species are caught on it. The data is split into two datasets, one for site information and one for catch information. The first step is to join these two datasets together by station key. Next, using the Taxon variable in the Long-Line Data the scientific names of the fish are found. By filtering Taxon by this list above a dataset with only the ten species should be obtained. Due to the change of sampling methods, the key species found by trawling have minimal observations from the long line. Only three of the ten listed species are present and none of them have enough observations to make

a meaningful claim about the change of its population over time. As a result, we can conclude that even though we are able to identify the key species, they must be evaluated by trawling and we can not use the long-line data to evaluate how these species.

3.4 Question 4: Using Long-line Data From after the Oil Spill Can Any Conclusions be made From That the spill affected Benthic Marine Biology

In Question 3, we where unable to perform a analysis due to lack of observations. To ensure that species being observed have enough information the top four observed species in the dataset was found. This was done by combining the station and the catch datasets again. Then, using the Taxon variable a count was preformed and the names where arranged in descending order. A list of the top 4 names were then created and used to filter the original data by finding the observations where the names in the list are seen in the Taxon variable. The start time variable was then modified to take the 6 and 7 statement which refer to the year. Using a "if else" statement if year is less than 25 a 20 is placed in front and if year is greater than or equal to 25 a 19 is placed in front. This creates a year variable which can be plotted. The dataset is then grouped by Taxon and Year and the mean number of species per catch is found for each year. These plots can be seen in figure 3. By plotting the mean number caught per sample for every year a time plot can be created that allows a visual representation of how species change over the years; especially after 2010. It can be seen that after 2010 there is a decrease in all the mean number of observations; however, it is difficult to conclude if this is from the oil spill or variation in the data. In order to determine the significance of the oil spill, a linear regression was created using mean number as the response and year, Taxon, and after_spill (a binary variable for before or after the oil spill). This regression found the variable after_spill had a coefficient of -0.96557 with a p-value of 0.0386. This shows that "after the spill" is a significant variable for the four top observed species and since the coefficient is negative, it is associated with a decrease in the number of fish caught per sample. This regression was a r-squared of 0.8028 which means that the regression line fits the data well. As a result, we should use our long line data to conclude that at least for the top four observed species their population decreased as the mean number of species caught decreased after the oil spill.

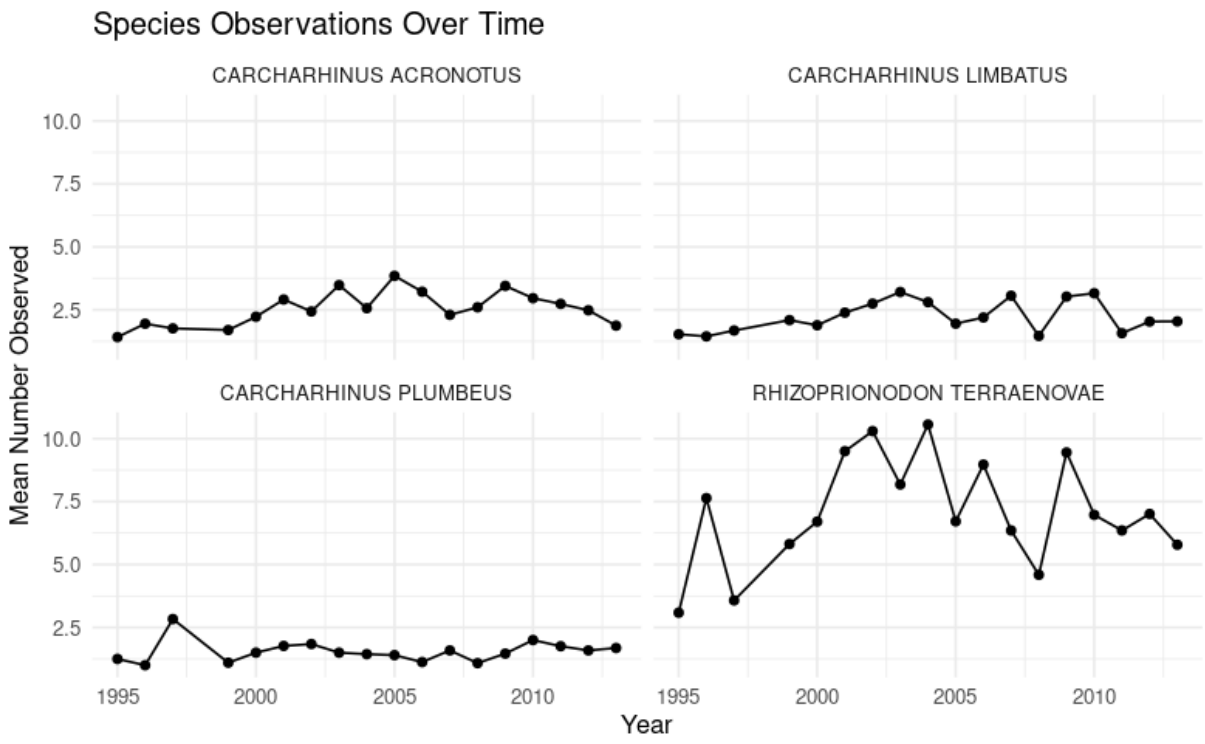


Figure 3: Mean Number of Observation of the top 4 Observed Species in the Long-Line dataset. By plotting the mean number caught per sample for every year a time plot can be created that allows a visual representation of how species change over the years; especially after 2010. It can be seen that after 2010 there is a decrease in all the mean number of observations. However, it is difficult to conclude if this is from the oil spill or variation in the data. A linear regression was fit with a binary value for after the 2010 in order to determine if there is a significant difference for the top 4 species. A coefficient of -0.96557 was found for after_spill (p-value of 0.0386). This regression model fit the data well with a r-squared of 0.8028.

4 Discussion

4.1 Limitations of the Study

Even though there is along of data provided within these two datasets, there are still limitations to it. As seen in Questions One and Two due to missing data and discrepancy in collection the data collected after 2010 could not be applied to a analysis. When the trawling data was applied, the result obtained was actually the opposite of the conclusion made from correctly distributed data. If the trawling dataset continued, the data would have been more representative of the true population and once enough data is added this analysis could be preformed. The second limitation was having two datasets using different collection methods. The key species identified by trawling were unable to be analysis through the long-line dataset as these species never or rarely where caught on a long-line. This means we must adjust the comparisons to the method being used to obtain key species.

4.2 Conclusions Made from Analysis

Based on the long line data set, through the selection of the top observed long line species, an analysis was preformed to find a decline in species caught after there oil spill. Since this analysis provided a significant p-value with a negative coefficient, we know there is a negative effect on the population. In this analysis has a r-squared of about .80 this model fits the data very well showing it does a good job modeling the true data. The uses of the trawling dataset found that an analysis of trawling fishery data from before and after the oil spill is impossible because the data is incomplete and not representative of the entire population. The list of key species were able to be created; however, due to different catch method these key species are unable to be compared by a different sampling method.

5 References

DARRP. (2025, March 18). Deepwater horizon: Oil spills: Damage assessment,remediation, and restoration program. Deepwater Horizon | Oil Spills | Damage Assessment, Remediation, and Restoration Program. https://darrp.noaa.gov/oil-spills/deepwater-horizon

NOAA. (2024, November 1). National Oceanic and Atmospheric Administration, Department of Commerce - Deepwater Horizon - Baseline Dataset (NCEI accession 0150631). Data Catalog. https://catalog.data.gov/dataset/deepwater-horizon-baseline-dataset-ncei-accession-0150631

NOAA. (2024a, October 3). Catch. NOAA Fisheries. https://www.fisheries.noaa.gov/inport/item/31654

NOAA. (2024b, October 3). DWH baseline SEAMAP Groundfish. NOAA Fisheries. https://www.fisheries.noaa.gov/inport/item/32334

NOAA. (2024c, October 3). Station. NOAA Fisheries. https://www.fisheries.noaa.gov/inport/item/31659

6 Appendix

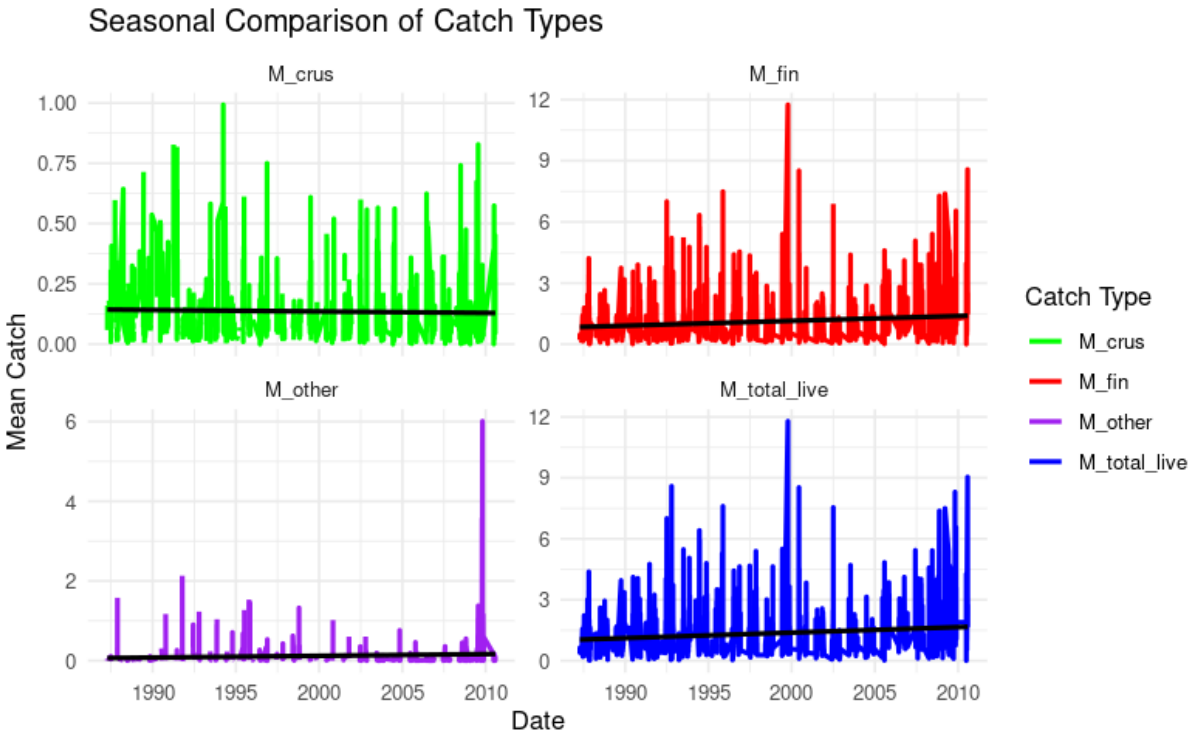


Figure 4: Overall Trends for the Dredging Data aver all Years of Data

```
> print(species_increased)
[1] "MICROPOGONIAS UNDULATUS" "TRICHIURUS LEPTURUS"     "CALINECTES ALL"
[4] "CALLINECTES SIMILIS"     "CYNOSCION NOTHUS"
> print(species_decreased)
[1] "CHLOROSCOMBRUS CHRYSURUS" "TRACHYPENEUS SIMILIS"     "CYNOSCION"
[4] "ANCHOA HEPSETUS"          "SERRANUS ATROBRANCHUS"
```

Figure 5: List of Species Which increased and Decreased form the Year 2000 compared to 2009

```
> print(species_counts)
                      TAXON  n
1        CYNOSCION NOTHUS  2
2 MICROPOGONIAS UNDULATUS  4
3     TRICHIURUS LEPTURUS 12
```

Figure 6: Count of Key Species Identified From Dredging applied to the Long-line Dataset