# A SPATIAL AND TEMPORAL ANALYSIS OF TEXAS BAYS AND MARINE SPECIES

An Undergraduate Research Scholars Thesis

by

FIALA EMIKO-MAE BUMPERS-ISHII

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Masami Fujiwara

May 2019

Major: Ecological Restoration
Renewable Natural Resources

# TABLE OF CONTENTS

# ABSTRACT

A Spatial and Temporal Analysis of Texas Bays and Marine Species

Fiala Emiko-Mae Bumpers-Ishii
Department of Ecosystem Sciences and Management
Texas A&M University

Research Advisor: Dr. Masami Fujiwara
Department of Wildlife and Fisheries Sciences
Texas A&M University

Temporal and spatial analysis of marine species distributions within the Gulf of Mexico is important in recognizing trends as to how their population dynamics change. Recognizing these trends can help fisheries and bay managers take precautionary action to better manage species important to a system and prevent biodiversity loss. This paper explores (1) how the abundance of fish and invertebrate species across the 8 major Texas bays are changing over time and space; (2) the spatial variability between the Upper and Lower Laguna Madre bay systems depicted by Self-Organizing Map tools (SOM). Species observation data collected by TPWD for 7 bays in the Gulf of Mexico over 35 years (between 1982-2016) across three sampling methods (gillnet, bag seine, and bay trawl) for over 1200 species of fish and invertebrates were analyzed in the R Studio Programming environment. Linear regression and related analysis were performed on the fish and invertebrate species data to determine their changes across bays (space) and over time. SOMs were created to determine differences in environmental variables between the Upper and Lower Laguna Madre. The findings for this study will allow for updated species distribution trends to be recognized, and allow for the exploration of the use of SOM tools in marine species distribution analyses.

# ACKNOWLEDGEMENTS

I would like to thank my research advisor, Dr. Masami Fujiwara, for his guidance and support for the last 3 years throughout the course of this research.

Thanks also go to the TPWD Coastal Fisheries Division and Dr. Fernando Martinez-Andrade for their data collection and documentation.

I also wish to extend gratitude to NSF Division of Ocean Sciences for funding this project through award number 1656923 to MF.

I would like to thank Dr. Akihiro Tokai and the Tokai Lab at Osaka University in Japan, for providing use of the Viscovery SOMine software and allowing me to join as a member of their lab. Special thanks go to Yoshihiko Ika, for his guidance in understanding and utilizing the software and for acting as my mentor during my stay.

I also wish to extend my gratitude to my friend, Natsuki Ozawa, and the rest of the Ozawa family for hosting me during the duration of my research in Japan.

Finally, thanks go to my mother-in-law, husband, and the rest of the Simpson family for their support and love.

# CHAPTER I

# INTRODUCTION

## The Economic Significance of the Gulf of Mexico Fisheries Industry

### *National Economic Impact of Fisheries*

Fisheries are important to the American economy. The commercial and recreational fishing industries generated $207.6 billion in sales, added $96.7 billion to the gross domestic product, and supported 1.62 million full-and part-time jobs worth $62.4 billion in income across the U.S in 2015 (National Marine Fisheries Service, 2017). Of this, the commercial fisheries sector contributed $144.2 billion in sales, $60.6 billion in gross domestic product, and supported 1.18 million jobs worth $39.7 billion; while the recreational fisheries sector contributed $63.4 billion in sales, $36.1 billion in gross domestic product, and supported 439,242 jobs worth $22.7 billion in income (National Marine Fisheries Service, 2017).

### *The Gulf of Mexico Region*

The Gulf of Mexico (GOM) fisheries region is particularly impactful to the economy. The region is one of the 7 regions in the US which contribute to the above national fisheries statistics (National Marine Fisheries Service, 2017). Commercial fishing in the Gulf accounted for 21% of the US' commercial seafood landings between 1992-2003 (Adams, Hernandez, & Cato, 2004). For the recreational fisheries sector, the economic activity associated with the GOM is greater than that in any other federal Regional Fishery Management Council area in the U.S (Adams et al., 2004). The commercial and recreational fishing industries together generated $33.6 billion in sales, contributed $15.26 billion to the gross domestic product, and supported

253,103 million full-and part-time jobs worth $9.3 billion dollars in income for the GOM region in 2015 (National Marine Fisheries Service, 2017). Of this, the commercial fisheries sector generated $21.5 billion in sales, contributed $7.85 billion to the gross domestic product, and supported 146,004 jobs worth $4.73 billion dollars in income; and the recreational fisheries sector generated $12.1 billion in sales, contributed $7.41 billion to the gross domestic product, and supported 107,099 jobs worth $4.57 billion dollars in income (National Marine Fisheries Service, 2017). Sales in both sectors are significant to both the regional and national economy, and if those dollars are spent on additional goods or services, then that spending will generate additional economic activity for the region (Lallo, 2017).

*Texas*

Texas is one of the regions within the GOM that is reliant upon the economic activities generated by the fisheries industry. In Texas, the commercial and recreational fishing industries together generated $2.95 billion in sales, contributed $1.71 billion to the gross domestic product, and supported 30,197 full-and part-time jobs worth $1.09 billion dollars in income during the 2015 year (National Marine Fisheries Service, 2017). Of this, the commercial fisheries sector generated $1.02 billion in sales, contributed $509.8 million to the gross domestic product, and supported 14,829 jobs worth $361.4 million dollars in income; and the recreational fisheries sector generated $1.93 billion in sales, contributed $1.20 billion to the gross domestic product, and supported 15,368 jobs worth $726 million dollars in income (National Marine Fisheries Service, 2017).

Management of the GOM is important, nationally and locally. Nationally, the GOM provides invaluable natural resources and economic activity to the US (Adams et al., 2004).

Locally, the GOM provides an important source of jobs and earnings for its coastal communities, and the non-resident tourists it attracts represent an important source of new revenue for the local economies (Adams et al., 2004). Resource managers across the nation are becoming increasingly aware of their need to manage the Gulf in a sustainable manner to ensure actors in the marine-related industries and its surrounding communities continue to have full access to the natural resources provided by the Gulf of Mexico (Adams et al., 2004). In Texas, the Texas Parks and Wildlife Department (TPWD) plays that important role of managing the GOM area (Adams et al., 2004). TPWD regulates the fishing season, allowed fish catch and size, helps maintain the population of important commercial and recreational marine species, and protects marine habitats along Texas coast. If trends in the change of marine species distribution can be determined, TPWD will have a better idea as to which economically critical species may become depleted in the future and can begin taking precautionary action.

**Purpose of the Research**

Temporal and spatial analysis of species distributions are important in recognizing trends in how their population dynamics change. Recognizing these trends can help TPWD take precautionary action to better manage species important to a system and prevent biodiversity loss. While previous studies have been done in Texas bays analyzing spatial and temporal differences between bays and species, analysis has never been done with 35 years of data collected across a multitude of bays all together as in our case. This study will allow for more comprehensive trends in changes of marine species distribution to be recognized than previous studies.

# CHAPTER II

# BACKGROUND AND DATA COLLECTION

## Background

Data were collected by Texas Parks and Wildlife Department's Coastal Fisheries Division as part of their long-term Marine Resource Monitoring Program. The program aims to collect long-term data to assess changes in their relative abundance and size, their spatial and temporal distributions, species composition of the community, and environmental parameters known to influence their distribution and abundance (Martinez-Andrade, 2015). All data collection methodologies described below are outlined in their "Marine Resource Monitoring Operations Manual" (Martinez-Andrade, 2015).

## Site Description

Sampling has been conducted in seven major and three minor bays along the coast of Texas since 1982. The bays were numbered from 1-10, based on geographic location in order from North to South. Data were collected for all ten bays, but our studies will only analyze data collected in the seven major bays (numbered 2-8) listed in Table 1. Those seven bays (in order from North to South) are Galveston Bay, (West) Matagorda Bay, San Antonio Bay, Aransas Bay, Corpus Christi Bay, Upper Laguna Madre, and Lower Laguna Madre.

Table 1. Number of Samples Collected per Bay.

| Bay Number | Bay Name | Samples per Season* | | | Samples per Year | | |
|---|---|---|---|---|---|---|---|
| | | Bag Seine | Bay Trawl | Gill Net | Bag Seine | Bay Trawl | Gill Net |
| 2 | Galveston Bay | 20 | 20 | 45 | 240 | 240 | 90 |
| 3 | West Matagorda Bay | 20 | 20 | 45 | 240 | 240 | 90 |
| 4 | San Antonio Bay | 20 | 20 | 45 | 240 | 240 | 90 |
| 5 | Aransas Bay | 20 | 20 | 45 | 240 | 240 | 90 |
| 6 | Corpus Christi Bay | 20 | 20 | 45 | 240 | 240 | 90 |
| 7 | Upper Laguna Madre | 20 | 10 | 45 | 240 | 120 | 90 |
| 8 | Lower Laguna Madre | 20 | 10 | 45 | 240 | 120 | 90 |

*Samples are taken monthly for bag seine and bay trawl, and twice a year for gill net.

## Sampling Design

A stratified cluster sampling design was used to determine the sampling location at each bay. Each of the bays act as a non-overlapping stratum from which a fixed number of samples are drawn every season (Table 1). A different point within the bays were sampled for each cluster sample, without selecting the same cluster in the same month. Every species captured were recorded. The species mostly included algae, plant, invertebrate, and fish species. Each species was referred to by a species code number that is listed in the "Marine Resource Monitoring Operations Manual". For each sample, the bay number, sampling location within the bay, year, month, salinity, dissolved oxygen, temperature, depth, and turbidity were also recorded. Sample numbers varied by sampling gear type (Table 1).

## Sampling Gear Description

Data were collected with 3 different gear types: gill net, bag seine, and bay trawl. Gill nets mainly targeted subadult and adult fish; bag seines mainly targeted juvenile fish and invertebrates; and bay trawls mainly targeted juvenile fish, subadult fish, and invertebrates. Data were collected 20 times per month, every month using bag seine and bay trawl (10 with the bay trawl for the Upper Laguna Madre and Lower Laguna Madre). Data were collected 45 times per season using gill nets, with each season consisting of the months of April-June and September-November.

### *Gill Net*

Gill nets are set perpendicular to the shoreline at or near sunset and retrieved as soon as possible after sunrise the next day. They consist of four connected panels with stretched

monofilament mesh sizes 76 mm, 102 mm, 127 mm and 152 mm. A gill net is set so that the smallest mesh (76 mm) is closest to the shore.

*Bag Seine*

Bag seines are pulled bayward and perpendicular to the shoreline for deployment, then the sampling pull begins parallel to the shoreline. The sampling period should be between 0.5 hours sunrise and 0.5 hours after sunset. 19 mm stretched nylon #5 multifilament mesh is used for the wings, and 13 mm stretched nylon #5 multifilament mesh is used for the bag.

*Bay Trawl*

Trawls are towed for 10 minutes at speeds of 3 mph in a circular fashion. The sampling period is between 0.5 hours sunrise and 0.5 hours after sunset. A 38 mm stretched nylon multifilament mesh is used throughout.

# CHAPTER III

# LINEAR REGRESSION ANALYSIS OF FISH SPATIAL AND TEMPORAL DISTRIBUTIONS

## Introduction

In Texas, there are 8 major and 3 minor bay and estuary systems distributed along the Gulf Coast (Redwine, 1997). We hope to determine how the abundance of fish and invertebrate species across 7 of the major bays are changing over space and time. By determining the trends in species distribution over the last 35 years across the 10 bays, we can assist TPWD's mission to manage and maintain fisheries resources by informing them as to which critical fish populations may decrease in the future so that they can intervene before it is too late and help secure Texas' economy.

## Methods

Table 2. Number of Species observed 40-70% of the time (Number of species used as the subsetted data).

| | Bag Seine | Bay Trawl | Gill Net |
|---|---|---|---|
| Fish | 11 | 12 | 12 |
| Invertebrate | 5 | 7 | 1 |

Table 3. Number of samples used in linear regression analysis

| | Bag Seine | Bay Trawl | Gill Net |
|---|---|---|---|
| Fish | 64044 | 65650 | 34427 |
| Invertebrate | 27434 | 42495 | 5045 |

Data from all 35 years, each of the 7 bays, and all three gear types were analyzed using the R programming environment to allow for a comprehensive study encompassing the full range of different marine species and their life history stages captured in the samples. Only four of the data columns were used for this part of the analysis: the bay number, sampling location within the bay, year, and month. Data collected for each gear type, as well as fish and invertebrate data were treated separately for our analysis. For each gear type, the data was first subsetted to only include fish species with an observance rate between 40-70% of samples to limit our analysis to species that have enough observations for our regression analyses to have a high confidence level (Table 2). Then, the data was subsetted once more to only keep one record of observation for each month of sampling (this applies to the gill net samples as well). In doing this, a monthly presence-absence record for all fish and invertebrate species with a 40-70% sample observance rate for each gear type was created. The new data set was then used to conduct a linear regression analysis in R to determine if space and/or time affects species prevalence. The total sample size now remaining to be used in this analysis is included in Table 3.

Two associations were tested with the linear regression analysis: Number of times a species was observed (sampled) vs. Year and Number of times a species was observed (sampled) vs. Bay number (north to south). The first test hoped to determine if there was a temporal pattern of change in species distribution by measuring the number of times a species was observed each year across all 7 bays. The x-axis ranged from 1982-2016 and there was at maximum 84 observances per species per sampling gear recorded on the y-axis (1 observation per month x 12 months x 7 bays). The second test hoped to determine if there was a spatial difference in species distribution. Since the bays are numbered in order based on their geographic locations from north to south, observation rates among them could be analyzed to determine the latitudinal gradient of

species, possible differences in preferred environmental conditions, and potential invasion of warmer northern waters by southern species. The x-axis ranged from 2-8 (since the first of the seven bays we kept in our sample was numbered 2) and there was at maximum 420 observances per species per sampling gear recorded for the y-axis (1 observation per month x 12 months x 35 years).

All packages, codes, and related references used in this process is included in Appendix A-E.

## Linear Regression Results

*Number of Times Fish Species were Observed (sampled) vs. Year*

For the bag seine (Figure 1), all species except southern flounder showed a significant trend (p-values $< 0.05$). For the bay trawl (Figure 2), all species except Atlantic stingray, fringed flounder, pigfish, and spotted seatrout showed a significant trend. For gill net (Figure 3), all species except alligator gar and Atlantic stingray showed significant trends. Overall, 24 of the 31 fish species showed significant trends.

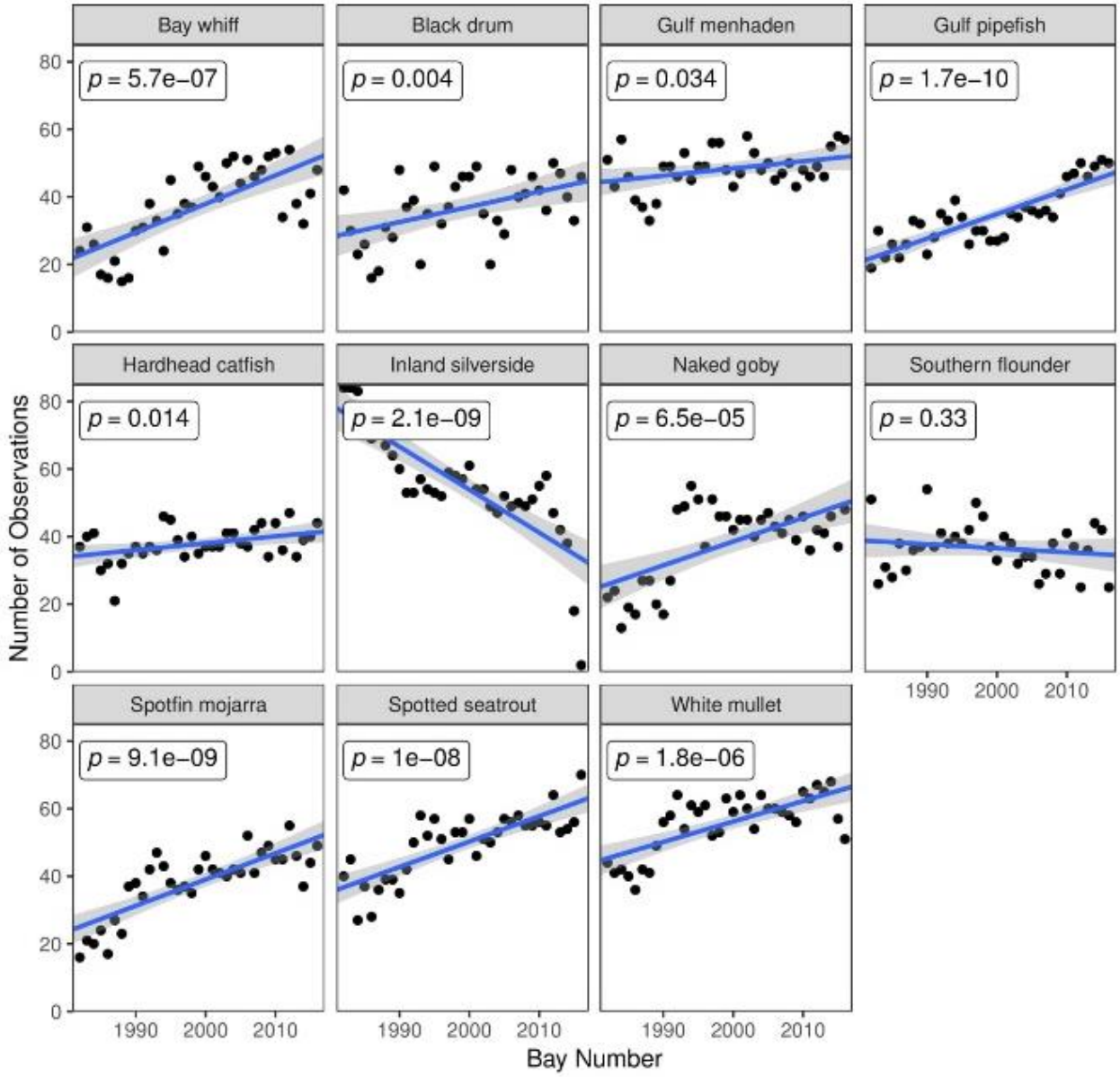


Figure 1. Linear regression of the number of observations (presences) of fish species plotted against time (in Years) for bag seine data.

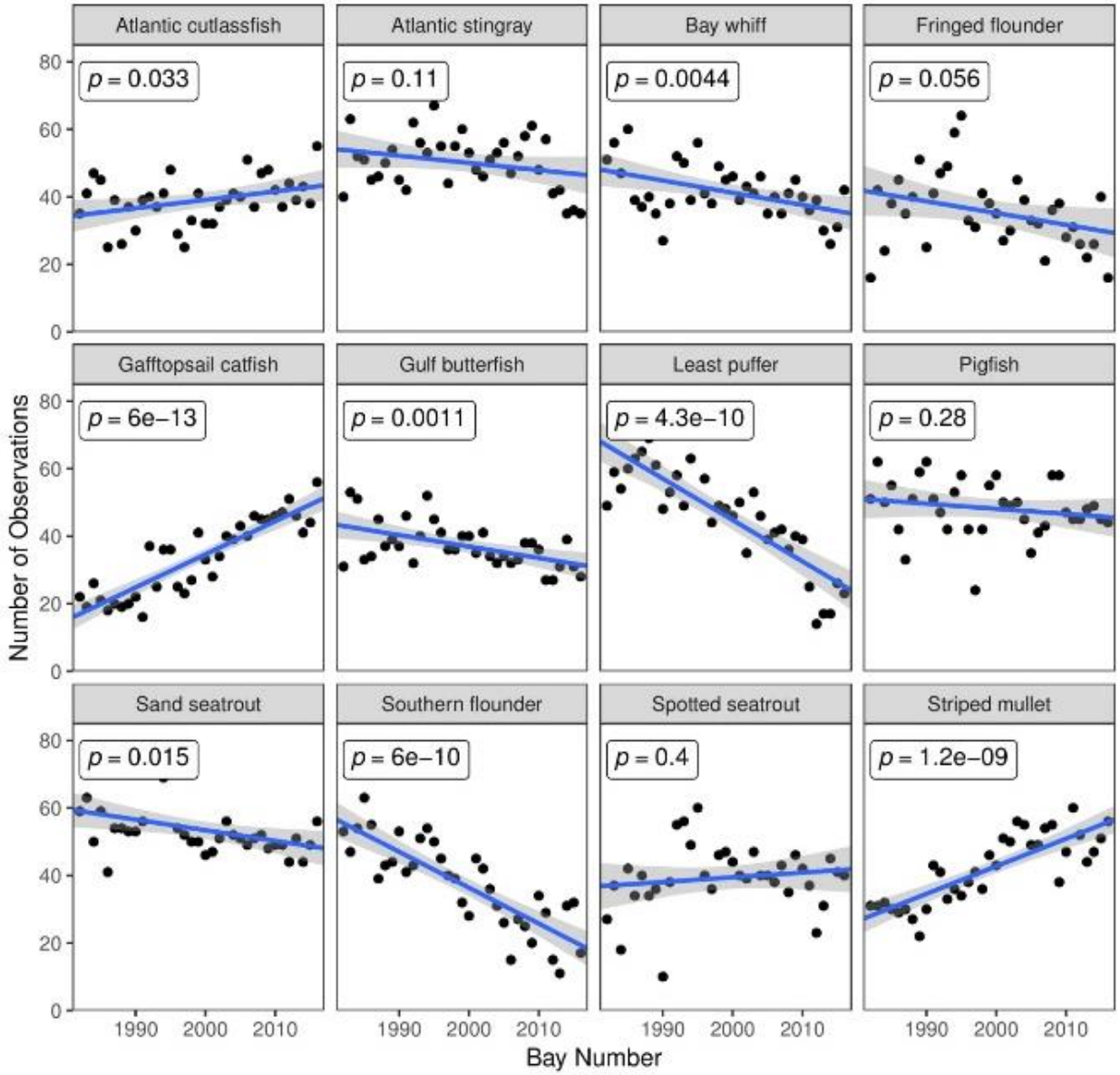


Figure 2. Linear regression of the number of observations (presences) of fish species plotted against time (in Years) for bay trawl data.

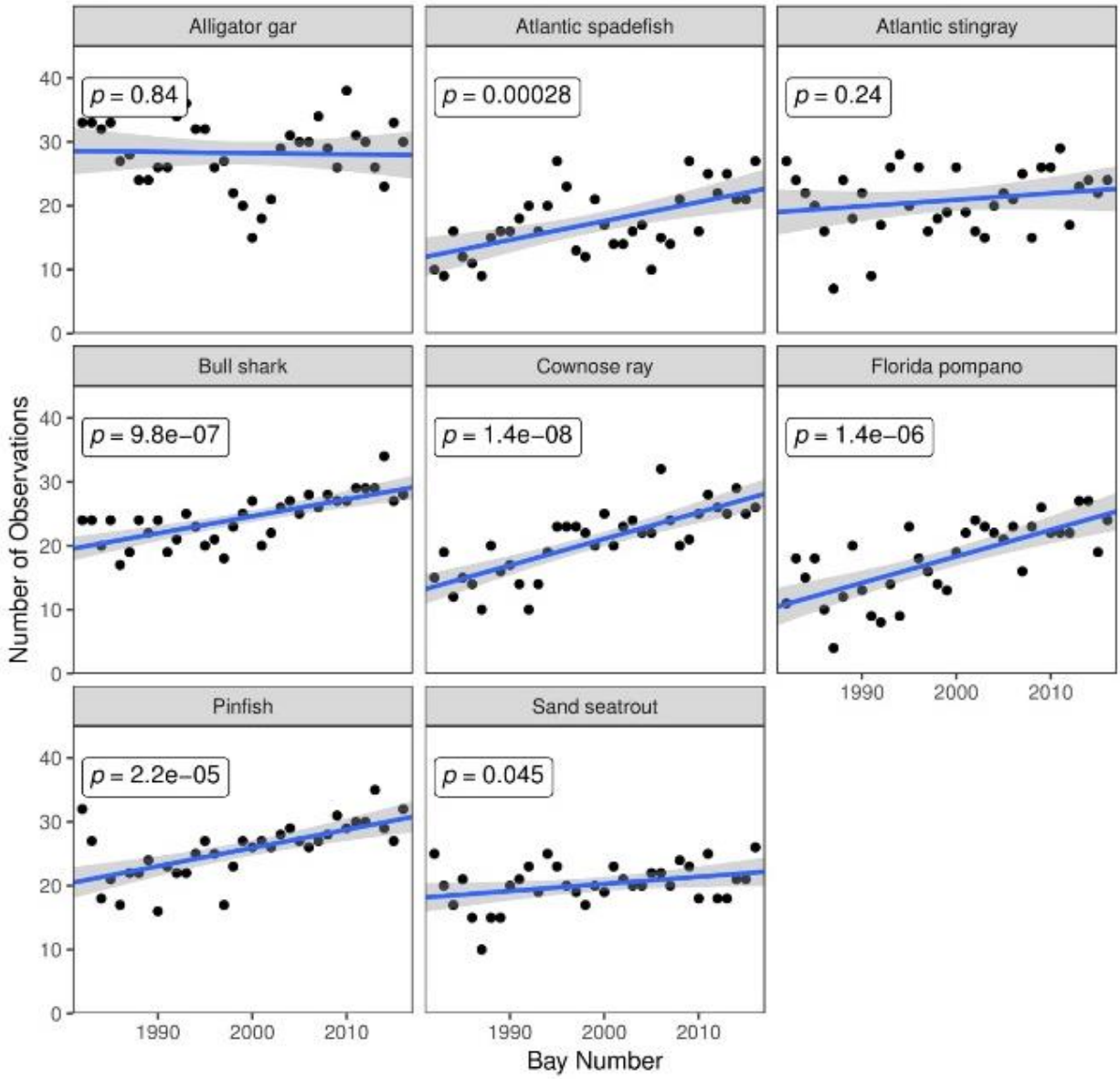


Figure 3. Linear regression of the number of observations (presences) of fish species plotted against time (in Years) for gill net data.

*Number of Times Invertebrate Species were Observed (sampled) vs. Year*

All 5 species measured via bag seine showed a significant trend (Figure 4). Of the 7 species captured with the bay trawl (Figure 5), four species showed significant trends (phosphorous jellyfish, lesser blue crab, and sea nettle), and three species showed insignificant trends (Eastern oyster, pink shrimp, and thinstripe hermit). The Gulf stone crab, the only species captured with the gill net, did not show a significant trend (Figure 6). Overall, 8 of the 13 invertebrate species showed significant trends.

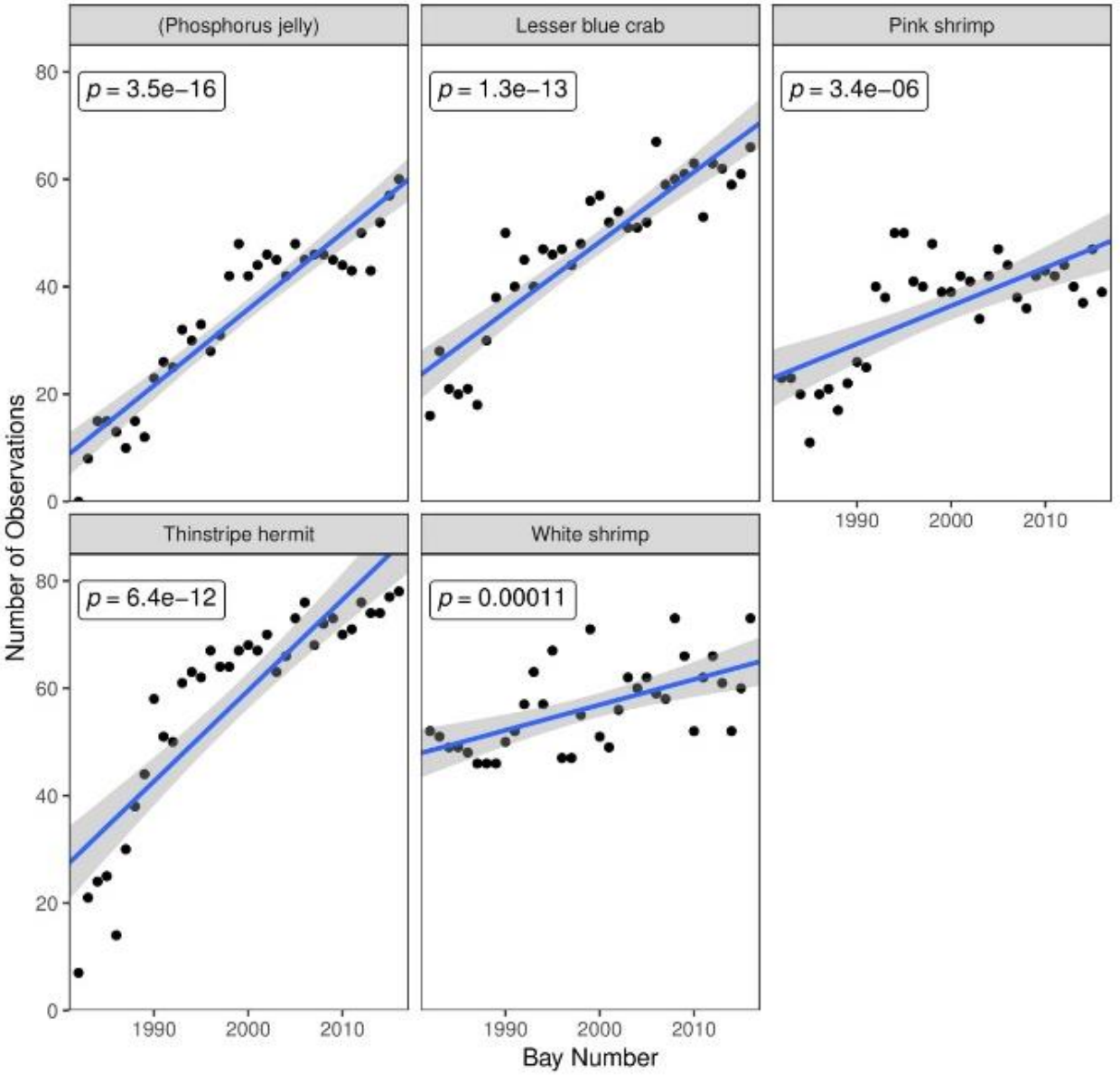


Figure 4. Linear regression of the number of observations (presences) of invertebrate species plotted against time (in Years) for bag seine data.

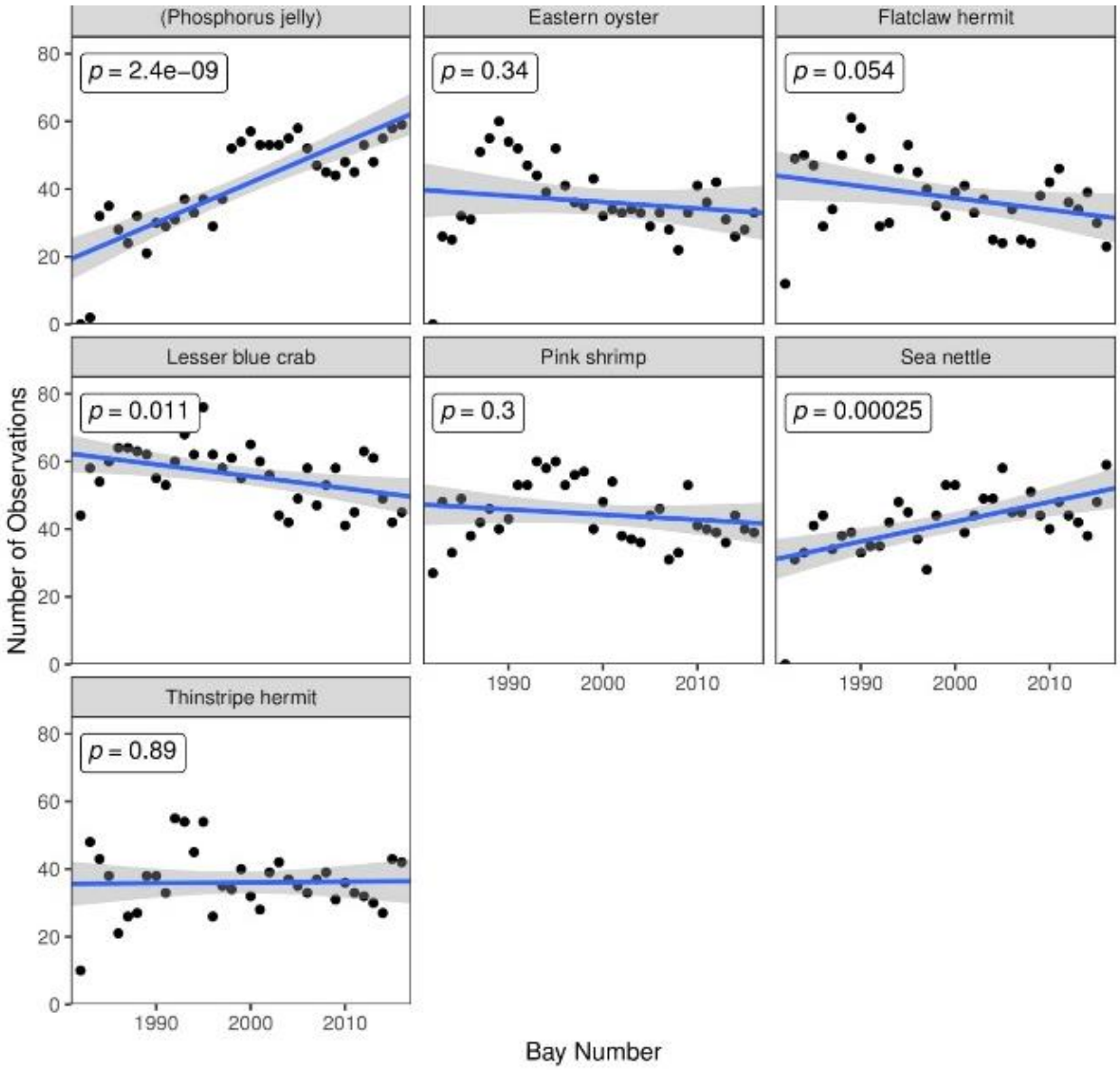


Figure 5. Linear regression of the number of observations (presences) of invertebrate species plotted against time (in Years) for bay trawl data.

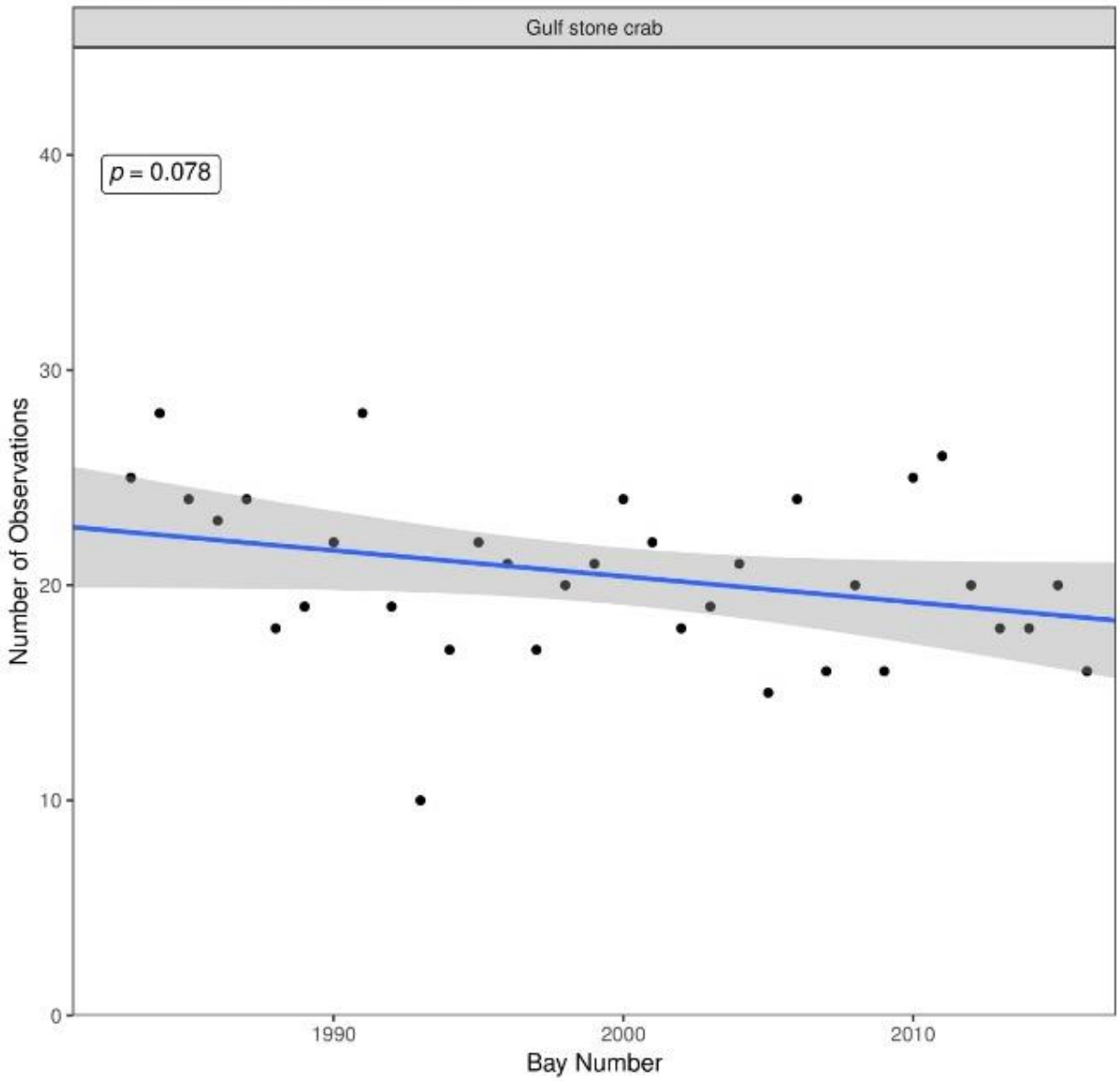


Figure 6. Linear regression of the number of observations (presences) of invertebrate species plotted against time (in Years) for gill net data.

*Number of Times Fish species were Observed (sampled) vs. Bay Number*

Of the 11 species graphed for bag seine (Figure 7), only 2 species showed a significant trend (Gulf menhaden and southern flounder). None of the 12 species graphed for bay trawl showed a significant trend (Figure 8). Of the 8 species graphed for gill net (Figure 9), only 2 species showed a significant trend (Alligator gar and bull shark). Overall, 4 of the 31 fish species showed significant trends.

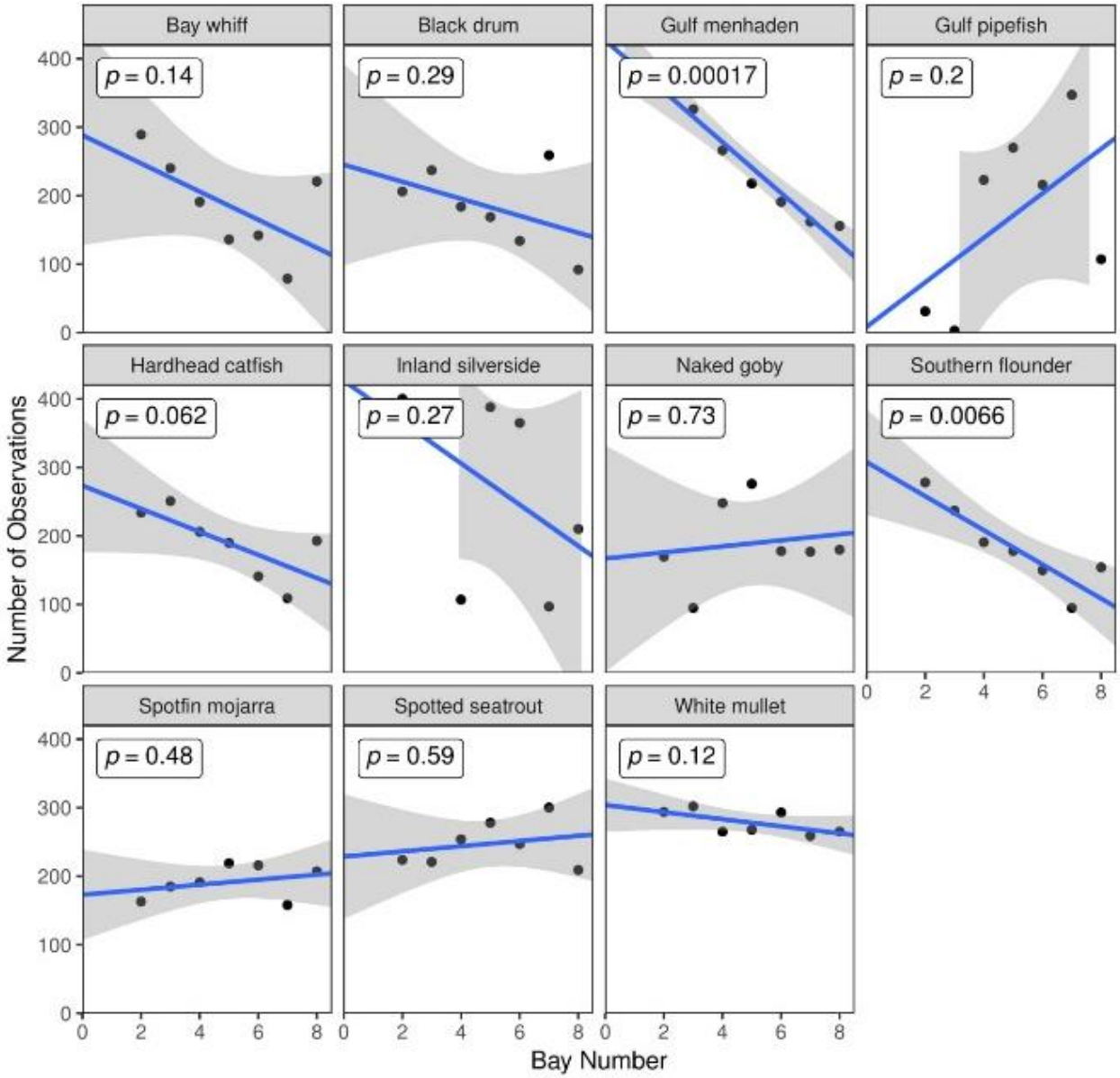


Figure 7. Linear regression of the number of observations (presences) of fish species plotted against the bay number (numbered 2-8 as indicated by Table 1) for bag seine data.

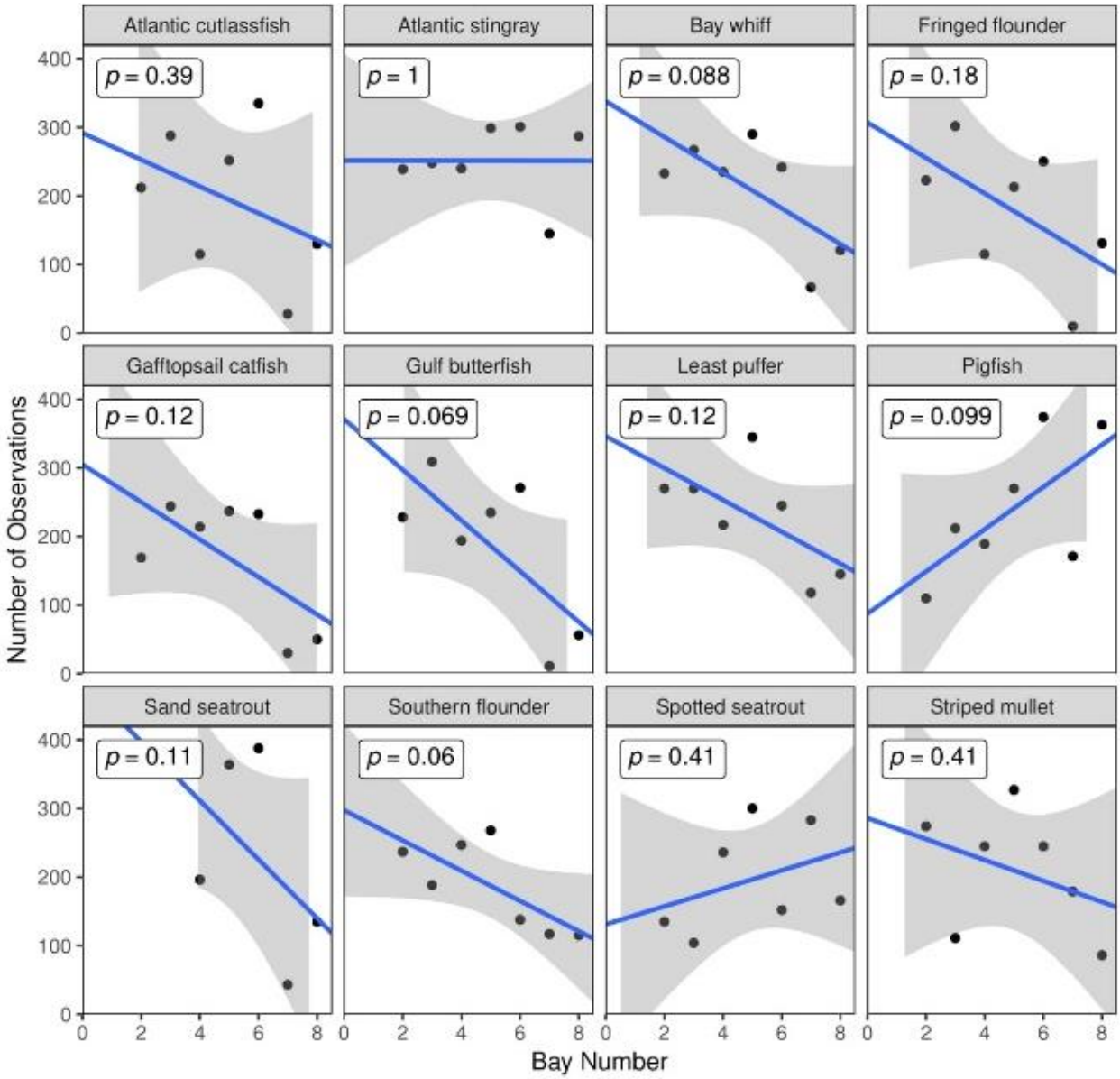


Figure 8. Linear regression of the number of observations (presences) of fish species plotted against the bay number (numbered 2-8 as indicated by Table 1) for bay trawl data.

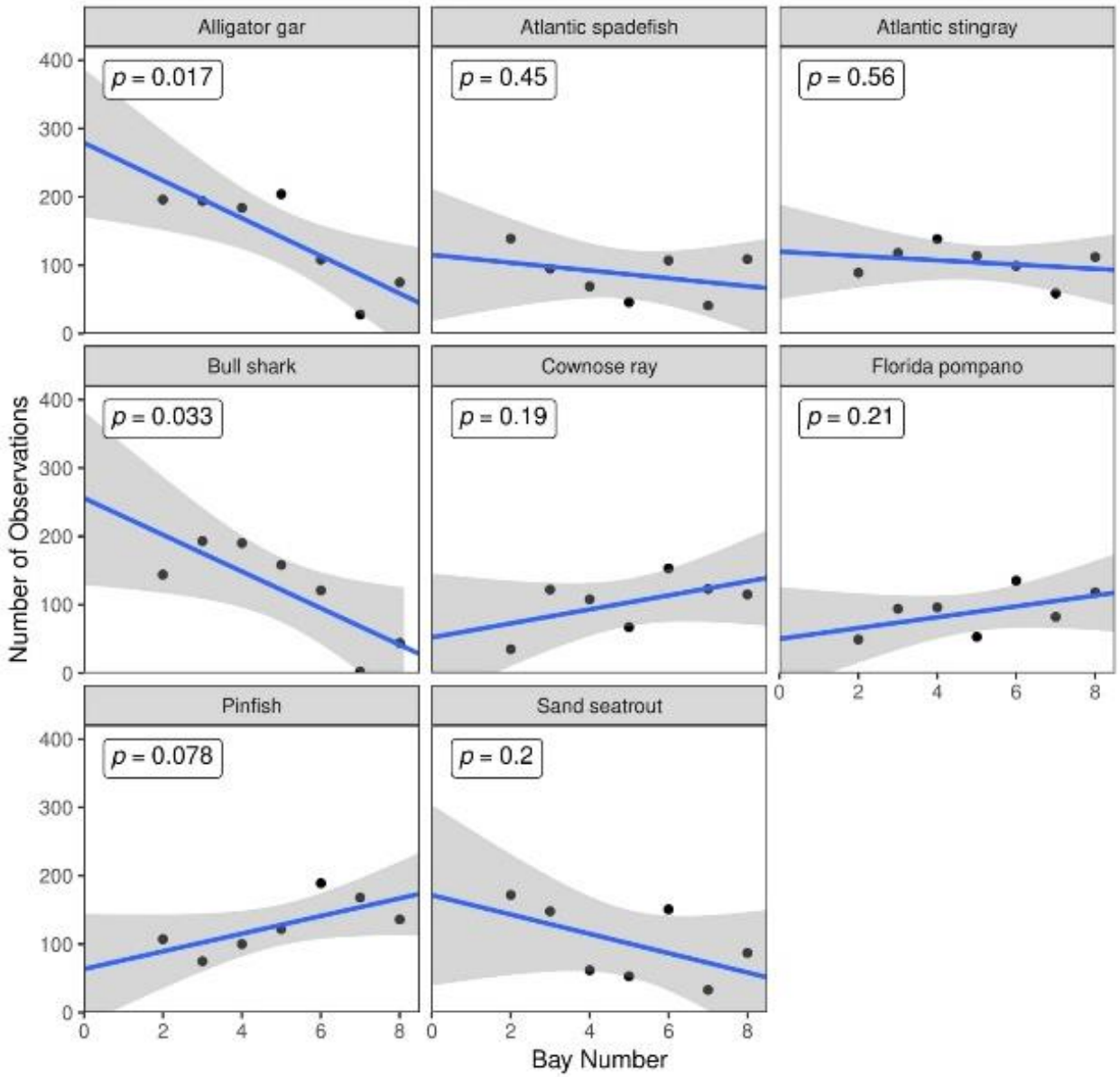


Figure 9. Linear regression of the number of observations (presences) of fish species plotted against the bay number (numbered 2-8 as indicated by Table 1) for gill net data.

*Number of Times Invertebrate species were Observed (sampled) vs. Bay Number*

For the bag seine (Figure 10), only pink shrimp showed a significant trend. Of the 7 species captured with the bay trawl (Figure 11), four species showed significant trends (phosphorous jellyfish, lesser blue crab, and sea nettle) and three species showed insignificant trends (Eastern oyster, pink shrimp, and thinstripe hermit). The Gulf stone crab, the only species captured with the gill net, showed a significant trend (Figure 12). Overall, 4 of the 13 invertebrate species graphed showed significant trends.

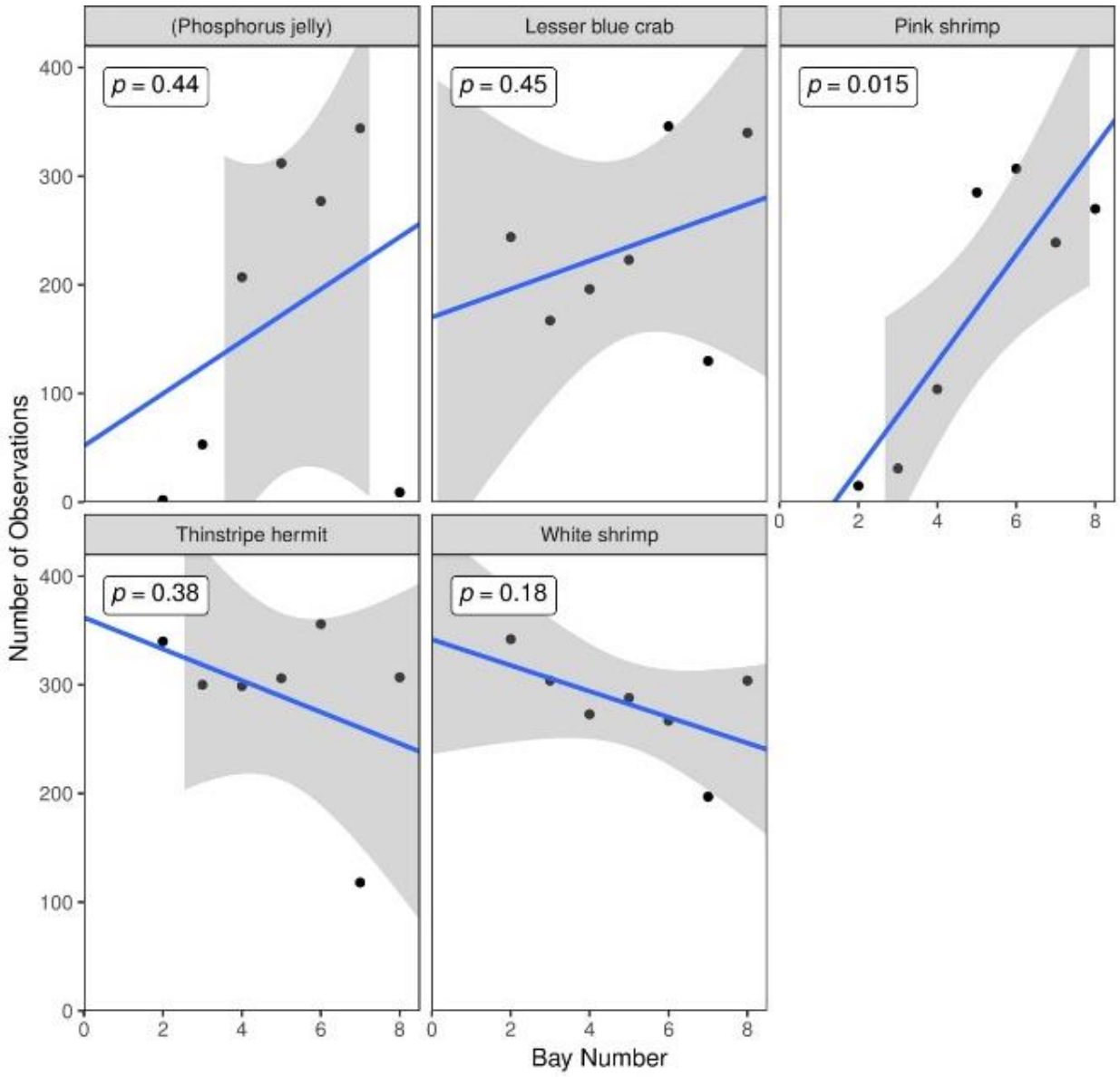


Figure 10. Linear regression of the number of observations (presences) of invertebrate species plotted against the bay number (numbered 2-8 as indicated by Table 1) for bag seine data.

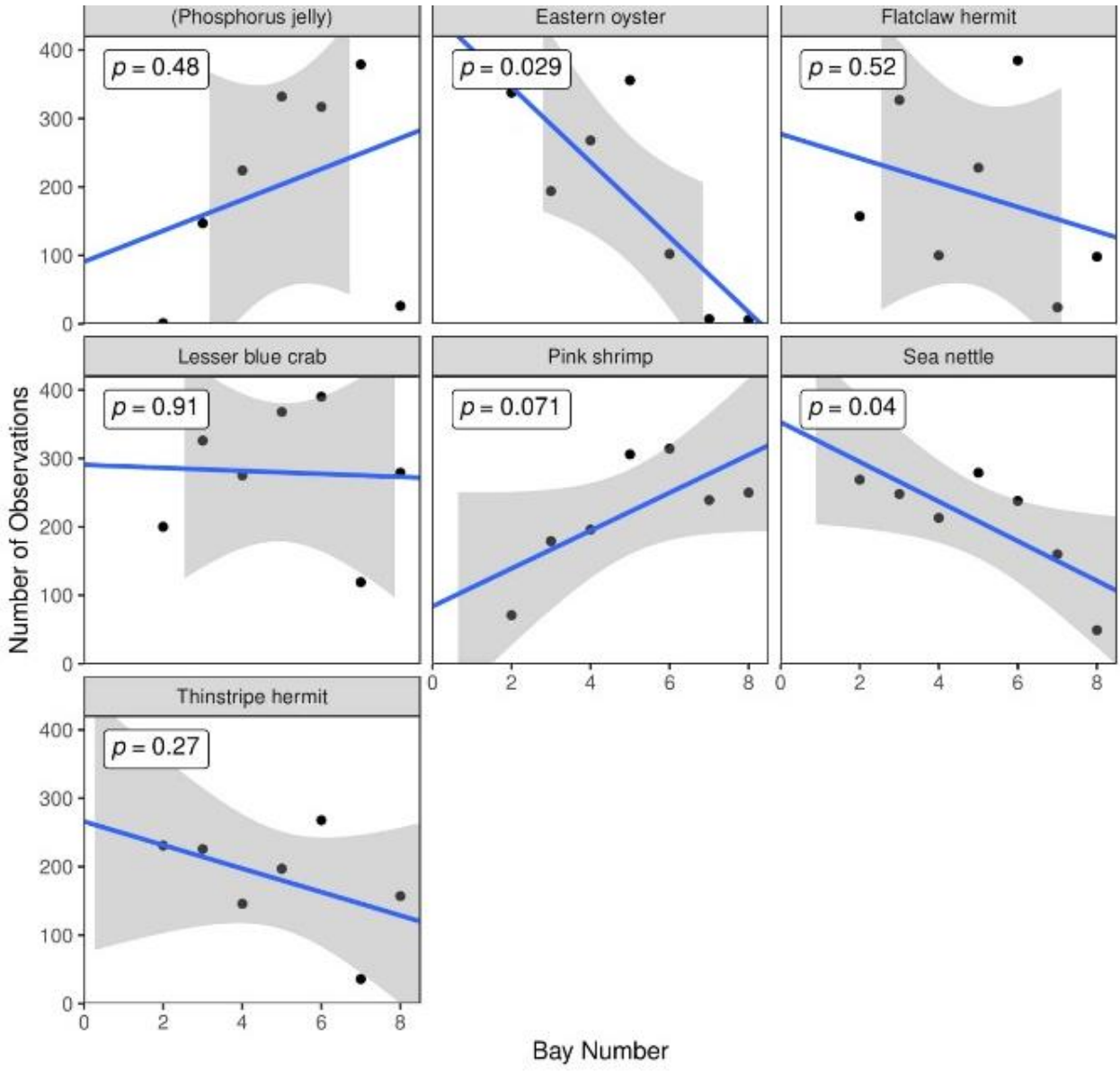


Figure 11. Linear regression of the number of observations (presences) of invertebrate species plotted against the bay number (numbered 2-8 as indicated by Table 1) for bay trawl data.

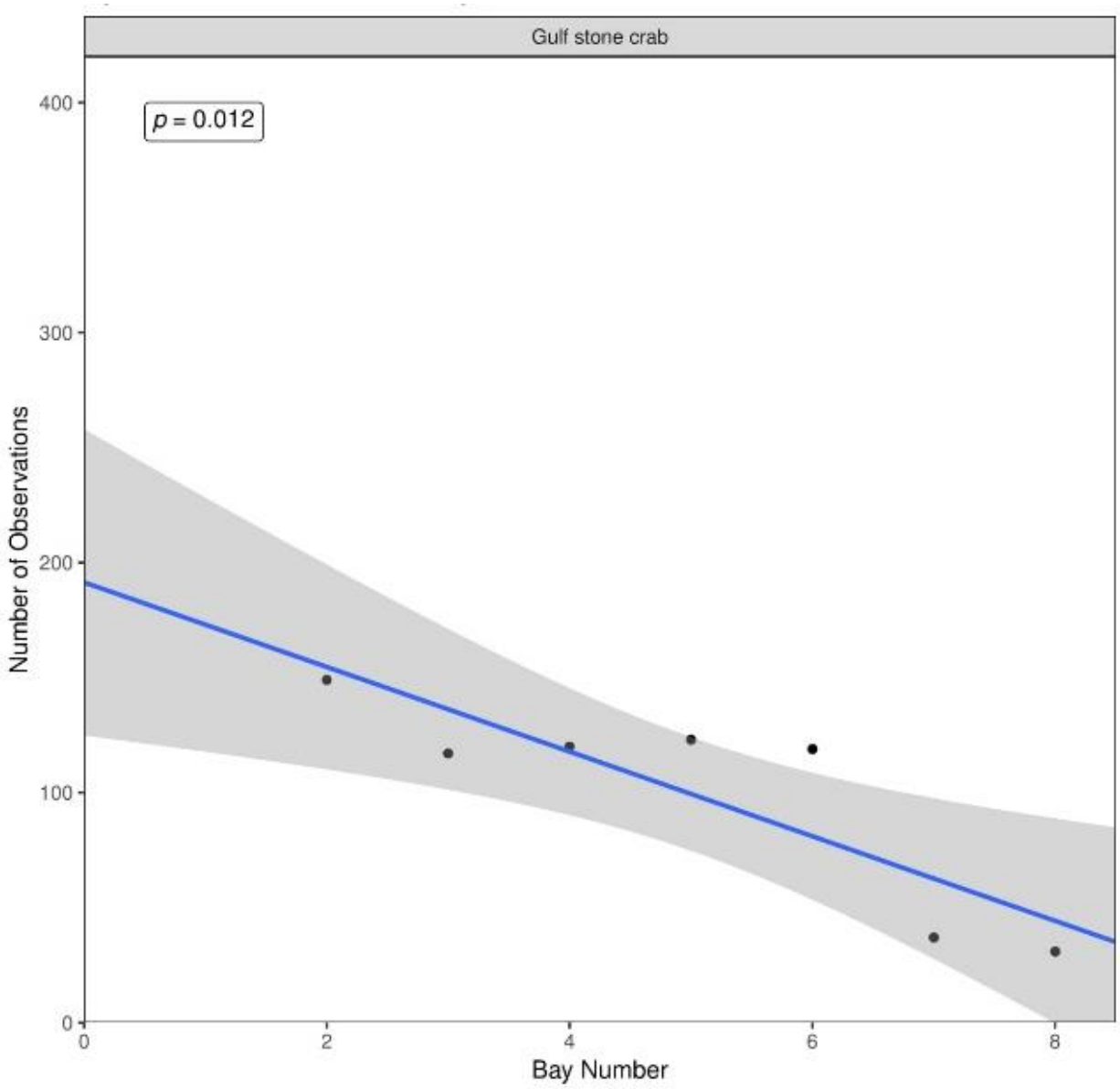


Figure 12. Linear regression of the number of observations (presences) of invertebrate species plotted against the bay number (numbered 2-8 as indicated by Table 1) for gill net data.

*Overall*

In total, 24 fish and 8 invertebrate species, as shown in Figures 1-6, had significant temporal correlations (p-value < 0.05); only 4 fish and 4 invertebrate species had significant trends (p-value < 0.05) which correlated with space (Figures 7-12). More species having a significant temporal correlation than spatial correlation suggest that time has a higher impact on species distribution than spatial/bay-to-bay differences for the ones investigated. This is also indicated by the confidence intervals on the graph (the grey zone). The confidence intervals for the temporal graphs are generally far narrower than those on the spatial graphs. This indicates that the plotted temporal relationships are more precise and representative of the means of the plotted values than those of the spatial relationships.

**Discussion**

More species were found to have significant temporal trend over spatial trend. 32 species out of 44 had significant temporal trend, while only 8 species out of 44 had significant spatial trend. Four times as many species were found to have significant association with time as compared to space. This may be attributed to changes in fishery management or environmental conditions.

Management has grown more centralized and, therefore, far-reaching. Before 1983, management of Texas marine species were overseen by different counties. In 1983, the Wildlife Conservation Act was passed, and it gave TPWD central authority to manage marine species (Bengston, Blankinship, & Bonds, 2003). Centralizing management of species under one governing body has allowed managers to collect more data and therefore make more management actions that have implications across the spatial boundaries of bays. This may have helped some species to increase in abundance.

TPWD has gained more authority and passed more regulations to help manage marine species over the years. Because they are the central body governing fisheries across the entirety of Texas, they are also tasked with stocking bays and determining catch limits in all of the bays (Bengston et al., 2003; Texas Parks and Wildlife Department). They have established hatcheries to help with supplementing recreationally important species (Texas Parks and Wildlife Department). TPWD's stock supplementing practices may explain some of the species' population trends being less variable across different bays despite differences in bay environmental conditions.

Overall, we hope temporal and spatial analysis of marine species distributions within the Gulf of Mexico can help fisheries managers recognize trends in population dynamics, take precautionary action to prevent biodiversity loss, and better manage ecologically and economically important marine fish and invertebrate species.

# CHAPTER IV

# SELF-ORGRANIZING MAPPING OF THE UPPER AND LOWER LAGUNA MADRE BAY SYSTEMS' ENVIRONMENTAL CONDITIONS

## Introduction

We hope to determine if there is spatial variability in the distribution of seagrass among two of Texas' bays: the Upper Laguna Madre (ULM) and the Lower Laguna Madre (LLM). This study will be conducted using Self-organizing mapping software. No extensive studies on changes in marine species distribution have been conducted in the GOM region using Self-Organizing Maps (SOMs) as in our case. This study will allow for the exploration of the use of SOM as a tool in marine species distribution analyses.

## Methods

SOM is an unsupervised neural network machine learning method used to reduce dimensionality. Nodes that represent a fixed position on the map are positioned into a grid shape that will be the final shape of the map. Then, the nodes are moved closer to a data point that it is closest in value to, while still maintaining the shape of the original grid. Neighboring nodes are then moved closer together or further apart in this manner until the nodes cannot move anymore. This process produces a two-dimensional map that represents multi-dimensional data and groups the data into clusters based on their similarity. By reducing the dimensional complexity of the data and clustering similar data points together, it increases readability of the data and allows for more ease in finding relationships between its variables.

Data from all 35 years for the ULM and LLM bays were analyzed. Of the three gear types, we chose to analyze the data collected using gill nets as it had the most complete data set for the bays of interest, with 3155 observations recorded for the ULM and 3153 observations recorded for the LLM. We kept all 45 samples taken per season for our analysis, totaling 90 samples per year. We hoped to compare the differences in five environmental variables for the two bays: salinity, dissolved oxygen, temperature, depth, and turbidity using the Viscovery SOMine SOM tool.

Data for the two bays were grouped in relation to the five environmental variables they were observed under, producing a separate map for each variable. The analysis was run twice using environmental data samples collected for two species that were the least sensitive to environmental change in both bays (>90% observance rate). Our first analysis was done on the hardhead catfish (*Ariopsis felis)*, referred to as SC 610. It had 2889 recorded observances in the ULM and 2960 recorded observances in the LLM. Our second analysis was done on the red drum (*Sciaenops ocellatus)*, referred to as SC 629. It had 2916 recorded observances in the ULM and 2965 recorded observances in the LLM.

A preliminary SOM analysis was conducted on the data to compare if there were any distinct differences in the environmental conditions found within the two bays or recorded between the fish species. If a difference was established for either measure, SOM analysis will be further continued to determine which environmental variables varied. Since each variable has a different measurement unit and scale, the data was first normalized in R using the "som" package (Yan, 2016). This package normalizes the data with the standardization method, in which the mean of each column is subtracted from a datum and divided by the standard deviation for that column, creating a new column of data with a mean of 0 and a standard deviation of 1.

## Self-Organizing Map Results

*Comparisons of Upper Laguna Madre vs. Lower Laguna Madre*

In Figure 13, blue marks the ULM bay and red marks the LLM bay. ULM is split into 2 clusters, with the smaller cluster centered around a concentration of high salinity, not found in the other ULM cluster or in LLM. This indicates that salinity may be an important and highly significant environmental variable that differs between the two bays. Higher values of depth and turbidity are observed in LLM, with dissolved oxygen values being only slightly skewed towards LLM as well. However, these variables do not form a cluster like with salinity, indicating the differences to the environment caused by these variables are not significant while salinity is. Temperature values did not show any discernable differences.

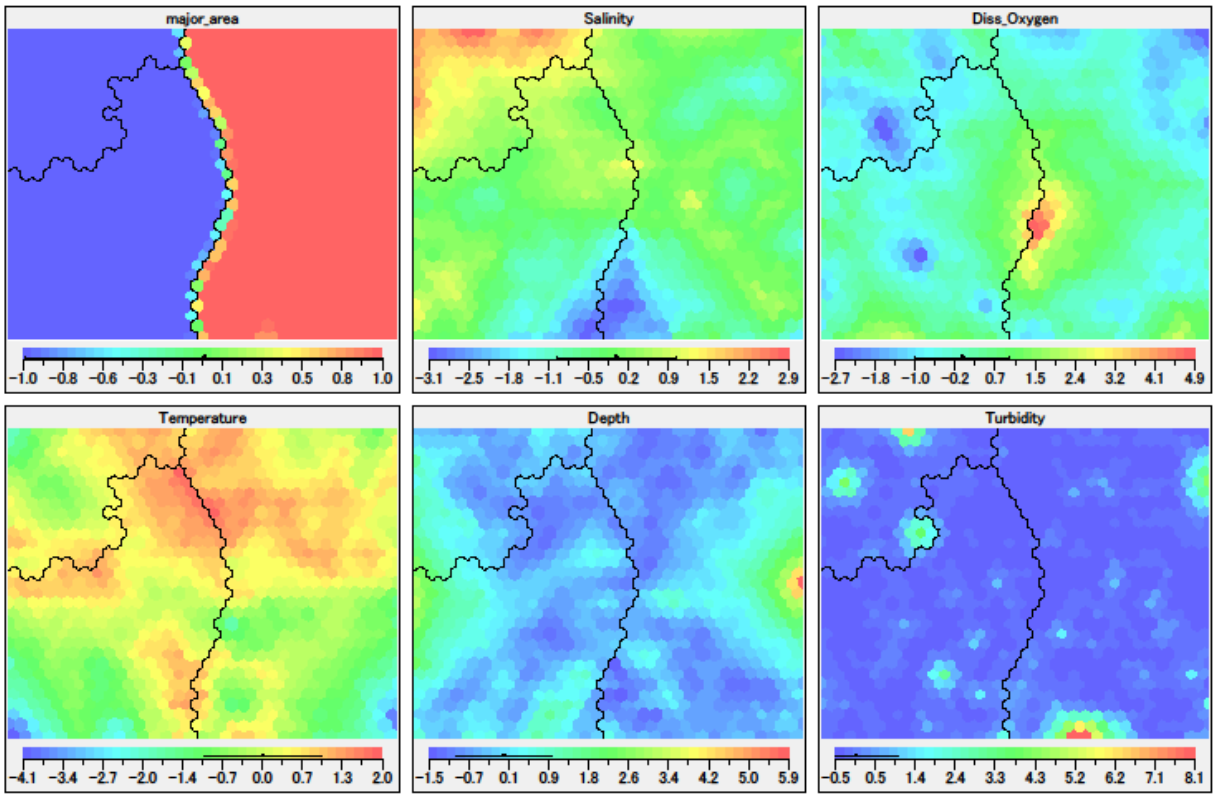


Figure 13. SOM comparing the environmental conditions observed between ULM and LLM. Measures of environmental variables range from blue to red, with blue indicating a low value for that variable and red indicating a high value for that variable. Note that the colors and scale are independent for each variable. The major_area panel represents the two bays mapped (blue for ULM and red for LLM).

Figures 14 and 15 break down the differences in environmental variables further. The focus of these two graphs is to see the difference in the range of values for the environmental variables observed. B7 had a wider range of observed temperatures (+/- 0.5 in both ends of the axis), but this was not a significant amount. Interestingly, the range of the turbidity and dissolved oxygen values did not differ between the bays, despite earlier observations that they were skewed towards LLM. ULM had double the higher salinity range of LLM (6.2 vs 3.5), while LLM had double the higher depth range of ULM (3.1 vs 1.4). This indicate that of the five environmental variables, perhaps only salinity and depth are the only two that are significant.

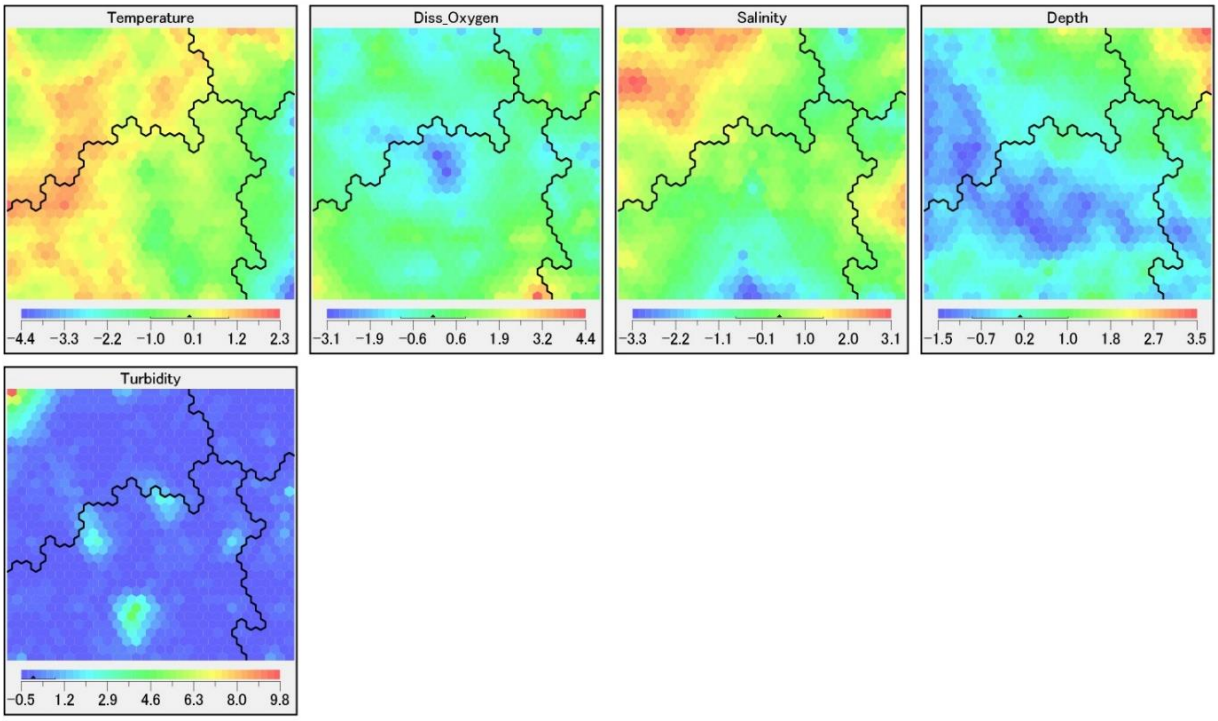


Figure 14. SOM of the environmental conditions observed in ULM. Measures of environmental variables range from blue to red, with blue indicating a low value for that variable and red indicating a high value for that variable. Note that the colors and scale are independent for each variable.

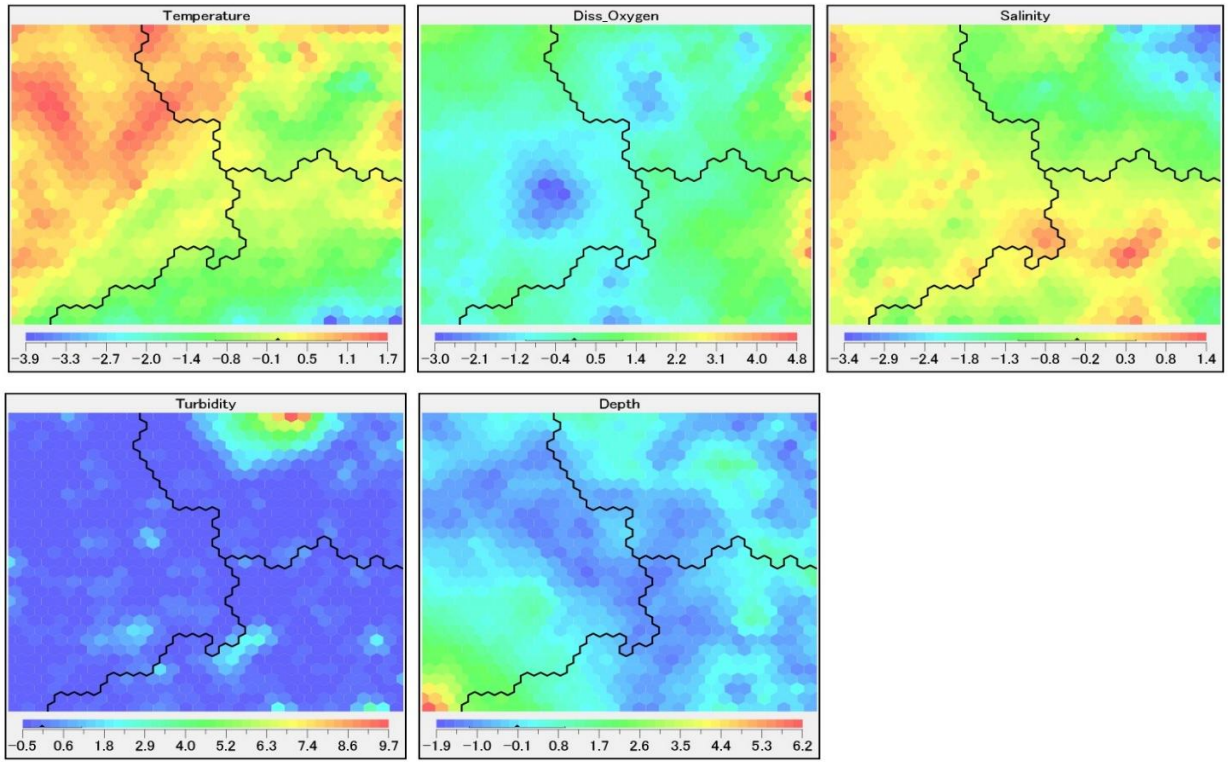


Figure 15. SOM of the environmental conditions observed in LLM. Measures of environmental variables range from blue to red, with blue indicating a low value for that variable and red indicating a high value for that variable. Note that the colors and scale are independent for each variable.

*Comparisons of Environmental Variables Observed for Hardhead Catfish vs. Red Drum*

Figures 16 and 17 compared the fish data (species_code panel) within each bay, with red marking red drum and blue marking hardhead catfish. The differences in environmental variables measured could have been caused by skew in habitat preferences by one of the two species. If the fish data showed a clean split when observed within each bay (no skewing towards certain habitats or clustering within one species' observation), it would indicate they have no effect towards the SOM division of the environmental variables into clusters. The data were split into 8 clusters for ULM (Figure 16) and 3 clusters for LLM (Figure 17). However, the clusters mirror each other vertically across the middle, split evenly between the two species, indicating that the species were not the variable that influenced the creation of that cluster. No environmental variable, except turbidity, was more prevalent in one half of the graph than the other. Red drum seem to be more tolerant of higher turbidities as compared to hardhead catfish, as indicated by the majority of red colors (highest value of observances) being found in the parts of the map associated with red drum but not under hardhead catfish. This pattern is most clearly seen in Figure 17. This indicates that the species most likely did not bias the resulting environmental comparison for the two bays, asides for turbidity.

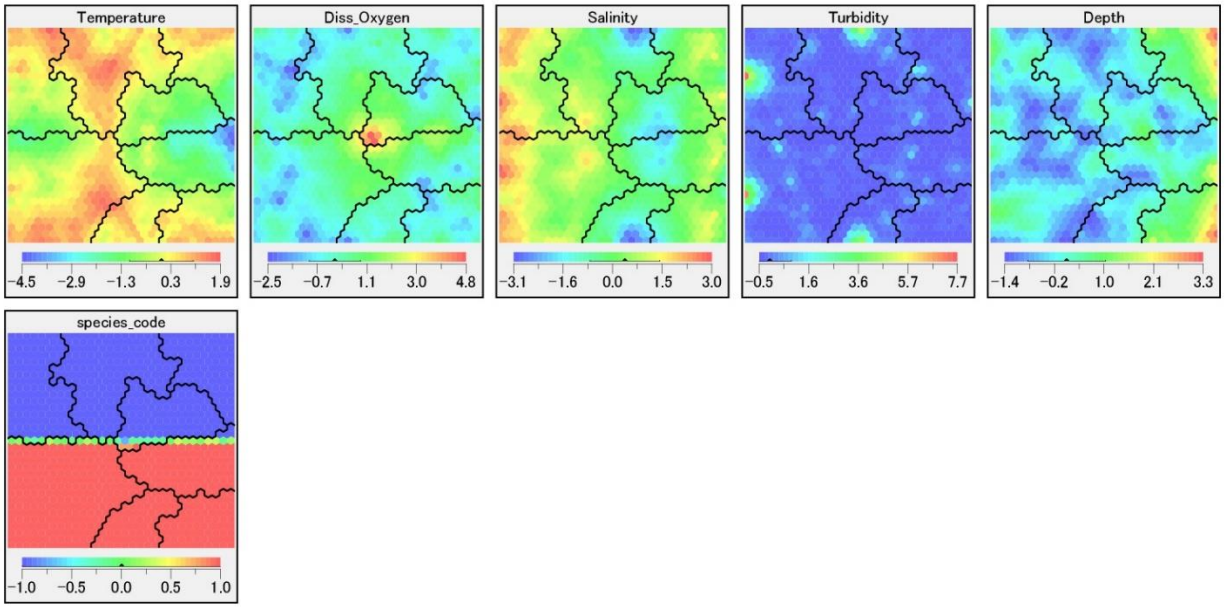


Figure 16. SOM of ULM comparing environmental conditions observed under both fish species. Measures of environmental variables range from blue to red, with blue indicating a low value for that variable and red indicating a high value for that variable. Note that the colors and scale are independent for each variable. The species_code panel represents the two species mapped (red for red drum and blue for hardhead catfish).

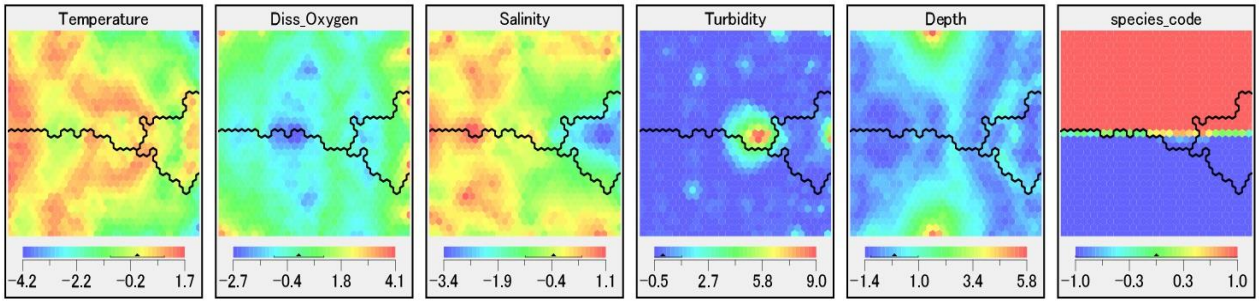


Figure 17. SOM of LLM comparing environmental conditions observed under both fish species. Measures of environmental variables range from blue to red, with blue indicating a low value for that variable and red indicating a high value for that variable. Note that the colors and scale are independent for each variable. The species_code panel represents the two species mapped (red for red drum and blue for hardhead catfish).

*Overall*

Finally, all four factors (red drum, hardhead catfish, ULM, LLM) were used to create a consolidated SOM in Figure 18. There are 4 clusters, correlating to each species (species_code) and bay (major_area). However, the spread of environmental variables along the species' codes split (horizontal split) mirror each other almost completely, while the split along the bays (vertical split) shows some variation. Along the bay split (vertical split), we once again observe a higher salinity concentration in ULM not found in LLM as seen in Figure 13. However, there is no cluster formed around it this time, indicating the salinity difference was not as big as the difference between the bays and fish species. This indicates that the significance of the differences between the fish species data is perhaps higher than the difference in environmental variables, but not as significant as the differences observed between the difference between the bays.

The difference in concentration of depth and turbidity is almost unobservably small, indicating once again these are not significant variables. Interestingly, the skew of dissolved oxygen towards LLM is clearer in this map than previous ones, such as Figures 14 and 15. The depth of the bays also seem slightly skewed towards B8. The difference in depth and dissolved oxygen measures were not as projected as the difference in salinity, indicating that they were not as significant. This may also explain why the difference in dissolved oxygen measures were not as easily observed in previous figures. The significance of the salinity measures may be so great as that the SOM algorithm covered it up.

Overall, salinity is the only environmental variable that showed a consistently large difference between the bays and both fish samples.

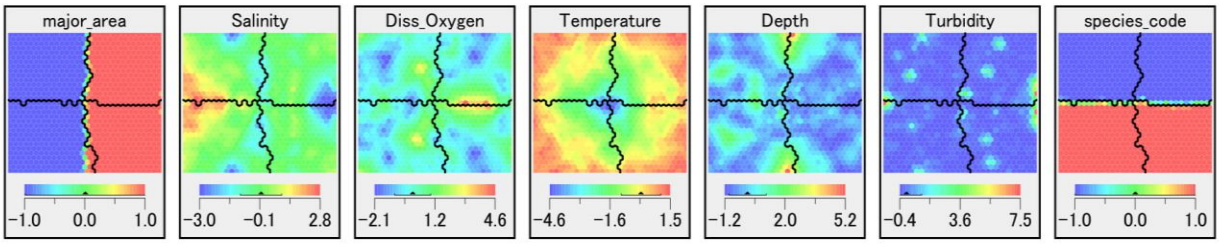


Figure 18. SOM comparing the environmental conditions observed between ULM, LLM, red drum, and hardhead catfish. Measures of environmental variables range from blue to red, with blue indicating a low value for that variable and red indicating a high value for that variable. Colors and scale are independent for each variable. The major_area panel represents the two bays (blue for ULM and red for LLM) and the species_code panel represents the two species (red for red drum and blue for hardhead catfish).

**Discussion**

Even when comparing species with high observation rates, SOM tools are able to discern a difference in environmental conditions that the species are found under. Both red drum and hardhead catfish are highly environmentally tolerant species that had greater than 90% observation rate in the two bays. Even under such high tolerances, the tool was able to measure differences in preferred environmental conditions of the two species. Red drum was only observed 32 more times than the hardhead catfish (5881 observations for the red drum to 5849 observations for the hardhead catfish), but this indicates that it may be able to persist under a slightly higher range of environmental conditions as compared to hardhead catfish, at least during the periods of sample collection. Indeed, the SOM analysis showed that red drum could tolerate higher turbidities as compared to hardhead catfish. We suggest SOM to be a potentially effective tool in studying marine species distributions and habitats.

# CHAPTER V

# SUMMARY AND CONCLUSION

TPWD has collected a large data sets on fish distribution. At the same time, various management techniques become available for them to choose. As time passes, more data will become available through their Marine Resources Monitoring Program. New analytical techniques, such as SOM, allow for this data set to be analyzed in new ways to increase the knowledge gained from the data further. We hope that as new knowledge and techniques are developed, TPWD's management action would improve its effectiveness further.

## REFERENCES


Adams, C. M., Hernandez, E., & Cato, J. C. (2004). The economic significance of the Gulf of Mexico related to population, income, employment, minerals, fisheries and shipping. *Ocean & Coastal Management, 47*(11-12), 565-580.

Bengston, S., Blankinship, R., & Bonds, C. (2003). *Texas Parks and Wildlife Department History: 1963-2003*. Retrieved from https://tpwd.texas.gov/publications/pwdpubs/media/pwd_rp_e0100_1144.pdf.

Lallo, E. (2017). NOAA Releases Fisheries Economics Report; 253,000 Jobs Good for Gulf,. Retrieved from http://gulfseafoodnews.com/2017/05/10/noaa-releases-2015-fisheries-economics-report-253000-jobs-good-gulf/

Martinez-Andrade, F. (2015). *Marine Resource Monitoring Operations Manual*. Instructional Manual. Texas Parks and Wildlife Department, Coastal Fisheries Division.

National Marine Fisheries Service. (2017). *Fisheries Economics of the United States, 2015*. Retrieved from https://www.fisheries.noaa.gov/resource/document/fisheries-economics-united-states-report-2015

Redwine, A. (1997). The economic value of the Texas Gulf Coast. In: Texas Natural Resource Conservation Commission.

Texas Parks and Wildlife Department. Fisheries Management at TPWD.

Yan, J. (2016). som: Self-Organizing Map. Retrieved from https://CRAN.R-project.org/package=som

# APPENDIX A

# PACKAGES INSTALLED

**List of Installed Packages**

```
library("readxl")

library("zoo")

library("data.table")

library("dplyr")

library("ggplus")

library("scales")

library("ggplot2")

library("grid")

library("plyr")

library("gridExtra")

library("ggforce")
```

**References**


Auguie B (2017). _gridExtra: Miscellaneous Functions for "Grid" Graphics_. R package version 2.3, <URL: https://CRAN.R-project.org/package=gridExtra>.

Bates D, Mächler M, Bolker B, Walker S (2015). “Fitting Linear Mixed-Effects Models Using l me4.” _Journal of Statistical Software_, *67*(1), 1-48. doi: 10.18637/jss.v067.i01 (URL: http://d oi.org/10.18637/jss.v067.i01).

Bates D, Maechler M (2018). _Matrix: Sparse and Dense Matrix Classes and Methods_. R packa ge version 1.2-14, <URL: https://CRAN.R-project.org/package=Matrix>.

Dowle M, Srinivasan A (2018). _data.table: Extension of `data.frame`_. R package version 1.11. 8, <URL: https://CRAN.R-project.org/package=data.table>.

Guiastrennec B (2018). _ggplus: Set of additional functions for ggplot2_. R package version 0.1, <URL: https://github.com/guiastrennec/ggplus>.

Henry L, Wickham H (2018). _purrr: Functional Programming Tools_. R package version 0.2.5, <URL: https://CRAN.R-project.org/package=purrr>.

Müller K, Wickham H (2018). _tibble: Simple Data Frames_. R package version 1.4.2, <URL: ht tps://CRAN.R-project.org/package=tibble>.

Pedersen TL (2018). _ggforce: Accelerating 'ggplot2'_. R package version 0.1.3, <URL: https:// CRAN.R-project.org/package=ggforce>.

RStudio Team (2016). _RStudio: Integrated Development Environment for R_. RStudio, Inc., Bo ston, MA. <URL: http://www.rstudio.com/>.

R Core Team (2018). _R: A Language and Environment for Statistical Computing_. R Foundatio n for Statistical Computing, Vienna, Austria. <URL: https://www.R-project.org/>.

Robinson D, Hayes A (2018). _broom: Convert Statistical Analysis Objects into Tidy Tibbles_. R package version 0.5.1, <URL: https://CRAN.R-project.org/package=broom>.

Wickham H (2011). “The Split-Apply-Combine Strategy for Data Analysis.” _Journal of Statisti cal Software_, *40*(1), 1-29. <URL: http://www.jstatsoft.org/v40/i01/>.

Wickham H (2016). _ggplot2: Elegant Graphics for Data Analysis_. Springer-Verlag New York. ISBN 978-3-319-24277-4, <URL: http://ggplot2.org>.

Wickham H (2017). _tidyverse: Easily Install and Load the 'Tidyverse'_. R package version 1.2.1 , <URL: https://CRAN.R-project.org/package=tidyverse>.

Wickham H (2018). _forcats: Tools for Working with Categorical Variables (Factors)_. R packa ge version 0.3.0, <URL: https://CRAN.R-project.org/package=forcats>.

Wickham H (2018). _stringr: Simple, Consistent Wrappers for Common String Operations_. R p ackage version 1.3.1, <URL: https://CRAN.R-project.org/package=stringr>.

Wickham H (2018). _scales: Scale Functions for Visualization_. R package version 1.0.0, <URL : https://CRAN.R-project.org/package=scales>.

Wickham H, Bryan J (2018). _readxl: Read Excel Files_. R package version 1.1.0, <URL: https:/ /CRAN.R-project.org/package=readxl>.

Wickham H, Bryan J (2018). _usethis: Automate Package and Project Setup_. R package version 1.4.0, <URL: https://CRAN.R-project.org/package=usethis>.

Wickham H, François R, Henry L, Müller K (2018). _dplyr: A Grammar of Data Manipulation_. R package version 0.7.8, <URL: https://CRAN.R-project.org/package=dplyr>.

Wickham H, Henry L (2018). _tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions_. R package version 0.8.2, <URL: https://CRAN.R-project.org/package=tidyr>.

Wickham H, Hester J, Chang W (2018). _devtools: Tools to Make Developing R Packages Easie r_. R package version 2.0.1, <URL: https://CRAN.R-project.org/package=devtools>.

Wickham H, Hester J, Francois R (2018). _readr: Read Rectangular Text Data_. R package versi on 1.3.0, <URL: https://CRAN.R-project.org/package=readr>.

Zeileis A, Grothendieck G (2005). “zoo: S3 Infrastructure for Regular and Irregular Time Series. ” _Journal of Statistical Software_, *14*(6), 1-27. doi: 10.18637/jss.v014.i06 (URL: http://doi.or g/10.18637/jss.v014.i06).

# APPENDIX B

## IMPORTING AND CLEANING OF THE ORIGINAL DATA SETS

*Note: The process for data import and clean-up is the same for all sampling methods (bag seine, bay trawl, and gill net). Only the code used for bag seine samples will be included in this abstract.

```
#ORIGINAL DATA FILES IMPORTED FROM EXCEL
species_codes <- read_excel("TPWD_Spp_codes.xlsx")  #Species code
datacolnames(species_codes) <- c("species_code", "latin_name", "common_name")

bag1wNA <- read_excel("TPWD Bag seine 82-00 all spp.xlsx")
          #BAG SEINE DATA FROM 1982-2000
bag2wNA <- read_excel("TPWD Bag seine 01-16 all spp.xlsx")
          #BAG SEINE DATA FROM 2001-2016 UNEDITED

#COMBINING FILES TO BE 1982-2016
bagwNA <- rbind(bag1wNA,bag2wNA)   #COMBINED BAG SEINE 1 AND 2 DATA

#CHANGE ALL NA DATA COLUMNS TO 0
bagw0 <- as.data.frame (lapply(bagwNA, function(d) { d[is.na(d)] <- 0; d }))

#COMBINE YEAR AND MONTH COLUMN TO CREATE A DATE COLUMN
bagdates <- as.data.frame (paste(bagw0$Year, bagw0$Month, sep="-"))
```

```
colnames(bagdates) <- "Date"  #CHANGE COLUMN NAME TO BE DATE.


#DATA FRAMES W NA VALUES FIXED, AND DATE COLUMN ADDED

bagwdate <- cbind(bagw0,bagdates)


#FINAL DATA FRAMES WITH NA VALUES FIXED, A DATE COLUMN, AND THE
SPECIES' SCIENTIFIC AND COMMON NAMES.

bag <- merge(bagwdate, species_codes, by = c("species_code"))
```

# APPENDIX C

## CHAPTER III FISH DATA TRANSFORMATIONS

*Note: The process of data transformations is the same for both species types (fish and invertebrate) and for all sampling methods (bag seine, bay trawl, and gill net). Only the code used for bag seine fish samples will be included in this abstract.

```
#DATA WITH DUPLICATE STATIONS REMOVED
bag_individual_stations <- bag[!duplicated(bag$station_id),]

#DATA FRAME WITH INVERTEBRATES AND VEGETATION REMOVED
bag_fish <- subset(bag, species_code <= 1800 & major_area > 1& major_area < 9 )

#DATA FRAME OF ONLY ONE FISH OBSERVATION RECORDED PER MONTH
bag_fish_one_observation <-
bag_fish[!duplicated(bag_fish[c("species_code","Date","major_area")]),]

#DATA FRAME OF THE NUMBER OF FISH OBSERVATIONS MADE EACH YEAR ACROSS ALL BAYS
bag_fish_nObservationsPerYear <- as.data.frame(bag_fish_one_observation %>%
  group_by(Year, species_code, latin_name, common_name) %>%
  tally)
```

```
#DATA FRAME OF THE NUMBER OF DATES FISH OBSERVATIONS WERE MADE PER BAY
bag_fish_nDates <- as.data.frame(bag_fish_one_observation %>%
  group_by(major_area, species_code, latin_name, common_name) %>%
  tally)

#DATA FRAME OF THE TOTAL NUMBER OF DATES FISH OBSERVATIONS WERE MADE
bag_fish_nObservationsTotal <- as.data.frame(bag_fish_one_observation %>%
  group_by(species_code, latin_name, common_name) %>%
  tally)

#PROPORTION OF FISH OBSERVATIONS ACROSS ALL BAYS AND YEARS
bag_fish_proportion <- as.data.frame(bag_fish_nObservationsTotal$n/2940)
colnames(bag_fish_proportion) <- "Proportion"
bag_fish_species_proportions <- cbind(bag_fish_nObservationsTotal, bag_fish_proportion)
```

# APPENDIX D

# CHAPTER III FISH DATA LINEAR REGRESSIONS

*Note: The process of data transformations is the same for both species types (fish and invertebrate) and for all sampling methods (bag seine, bay trawl, and gill net). Only the code used for bag seine fish samples will be included in this abstract.

```
#DATA FRAME WITH ONLY FISH OBSERVED BETWEEN 40%-70% OF THE TIME
bag_fish_nDateswProportions <- merge(bag_fish_nDates, bag_fish_species_proportions[ , c("species_code", "Proportion")], by = "species_code")

#SUBSET ONLY FISH OBSERVED BETWEEN 40%-70% OF THE TIME BUT WITH BAYS
bag_fish_subsetted_nDates <- subset(bag_fish_nDateswProportions, Proportion >= 0.4 & Proportion <= 0.7)

#SUBSETTED SPECIES NAMES AND NUMBER ONLY
bag_fish_subsetted_names_partial <- bag_fish_subsetted_nDates[!duplicated(bag_fish_subsetted_nDates[c("species_code")]),]
bag_fish_subsetted_names <-  bag_fish_subsetted_names_partial[,c("species_code", "common_name", "latin_name")]

#DATA TABLES FOR N DATES OF FISH OBSERVATION ACROSS BAYS
bag_fish_subsetted_nDates_table <- data.table(bag_fish_subsetted_nDates)[,list(n), keyby = c('major_area', 'species_code')]
```

```
#ADD IN DATA POINTS FOR MISSING FISH OBSERVATIONS IN EACH BAY
bag_fish_subsetted_nDates_nomissingvalues <-
bag_fish_subsetted_nDates_table[CJ(unique(major_area), unique(species_code))]

#CHANGE NA VALUES TO 0
bag_fish_subsetted_nDates_noNA <- as.data.table
(lapply(bag_fish_subsetted_nDates_nomissingvalues, function(d) { d[is.na(d)] <- 0; d }))

#LINEAR REGRESSION OF N DATES OF FISH OBSERVATION ACROSS BAYS
bag_fish_subsetted_nDates_linear_regression <- bag_fish_subsetted_nDates_noNA %>%
  group_by(species_code) %>%
  do(tidy(lm(n ~ major_area, data= .)))
        #NOTE: THE ESTIMATES FOR THE TERM MAJOR_AREA IS THE VALUE OF
THE SLOPE.

#SAVE THE N DATES OF FISH OBSERVATION ACROSS BAYS LINEAR REGRESSION
TABLES IN EXCEL
write.csv(x=bag_fish_subsetted_nDates_linear_regression,
file="C:/Users/arespostale/Dropbox/Occupancy/Linear_Regressions/Subsetted/Regression_Tren
d/bag_fish_subsetted_nDates_linear_regression.csv")
```

```
#FIT THE VALUES OF THE LINEAR REGRESSION OF N DATES OF FISH
OBSERVATION ACROSS BAYS
bag_fish_subsetted_nDates_fit_linear_regression <- bag_fish_subsetted_nDates_noNA %>%
  group_by(species_code) %>%
  do(fit=lm(n ~ major_area, data= .))

#R-SQUARED AND ADJUSTED R-SQUARED VALUES FOR THE LINEAR REGRESSION
OF N DATES OF FISH OBSERVATION ACROSS BAYS
bag_fish_subsetted_nDates_fit_linear_regression_r2 <-
bag_fish_subsetted_nDates_fit_linear_regression %>%
  glance(fit)

#SAVE THE R-SQUARED OUTPUTS FOR THE LINEAR REGRESSION OF N DATES OF
FISH OBSERVATION ACROSS BAYS
write.csv(x=bag_fish_subsetted_nDates_fit_linear_regression_r2,
file="C:/Users/arespostale/Dropbox/Occupancy/Linear_Regressions/Subsetted/R2_Values/bag_f
ish_subsetted_nDates_fit_linear_regression_r2.csv")

#N DATES OF FISH OBSERVATION ACROSS BAYS: SUBSET ONLY INTERCEPT AND
SLOPE VALUES
bag_fish_subsetted_nDates_linear_regression_3col <-
subset(bag_fish_subsetted_nDates_linear_regression, select=c(species_code, term, estimate))
```

```
#N DATES OF FISH OBSERVATION ACROSS BAYS: FIX INTERCEPT AND SLOPE TABLE
bag_fish_subsetted_nDates_linear_regression_spread <-
spread(bag_fish_subsetted_nDates_linear_regression_3col, key=term, value=estimate)
        #MAKE INTERCEPT AND SLOPE HAVE THEIR OWN COLUMNS INSTEAD OF BEING ONE COLUMN TOGETHER.
colnames(bag_fish_subsetted_nDates_linear_regression_spread) <- c("species_code",
"intercept", "slope")
        #CHANGE COLUMN TITLES TO BE INTERCEPT AND SLOPE TO AVOID CONFUSION.

#N DATES OF FISH OBSERVATION ACROSS BAYS: TABLE WITH SPECIES_CODE, INTERCEPT, SLOPE, R2, AND P VALUES
bag_fish_subsetted_nDates_linear_regression_partially_combined <- merge(x=
bag_fish_subsetted_nDates_linear_regression_spread, y=
bag_fish_subsetted_nDates_fit_linear_regression_r2[, c("species_code", "adj.r.squared",
"p.value")], by= "species_code")

#N DATES OF FISH OBSERVATION ACROSS BAYS: TABLE WITH MAJOR_AREA, SPECIES_CODE, N, INTERCEPT, SLOPE, R2, AND P VALUES
bag_fish_subsetted_nDates_linear_regression_combined <- merge(x=
bag_fish_subsetted_nDates_noNA[, c("major_area", "species_code", "n")], y=
bag_fish_subsetted_nDates_linear_regression_partially_combined, by= "species_code")
```

```
#ADDING COMMON NAMES TO THE PARTIALLY AND COMBINED TABLES
bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames <-
merge(bag_fish_subsetted_names,
bag_fish_subsetted_nDates_linear_regression_partially_combined, by= "species_code")
bag_fish_subsetted_nDates_linear_regression_combined_wnames <-
merge(bag_fish_subsetted_names, bag_fish_subsetted_nDates_linear_regression_combined, by=
"species_code")

#N DATES OF FISH OBSERVATION ACROSS BAYS: GEOM_LABEL TEXT CREATION
FUNCTION
bag_fish_subsetted_nDates_lm_eqn =
function(bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames){
  intercept <-
bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames["intercept"]
  plusminus <- ifelse(sign(intercept) >= 0,
                      paste0(" + "),
                      paste0("")   )
  eq1 <- substitute(paste(italic(y) == m* italic(x),plusminus, b),
                    list(b =
format(bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames$intercept,
digits = 3),
                         plusminus = plusminus,
```

```
                    m =
format(bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames$slope, digits
= 3)))
  eq2 <- substitute(paste(italic(r)^2~"="~r2),
                    list(r2 =
format(bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames$adj.r.square
d, digits = 4)))
  eq3 <- substitute(paste(italic(p) ~"="~p.val),
                    list(p.val =
format(bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames$p.value,
digits = 2)))
  c( as.character(as.expression(eq1)), as.character(as.expression(eq2)),
as.character(as.expression(eq3)))
}
bag_fish_subsetted_nDates_label_sc <-
ddply(bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames,.(species_cod
e),bag_fish_subsetted_nDates_lm_eqn)
bag_fish_subsetted_nDates_label_common <-
ddply(bag_fish_subsetted_nDates_linear_regression_partially_combined_wnames,.(common_na
me),bag_fish_subsetted_nDates_lm_eqn)
```

```
#N DATES OF FISH OBSERVATION ACROSS BAYS: TABLE WITH ALL VALUES
bag_fish_subsetted_nDates_linear_regression_final <- merge(x=
bag_fish_subsetted_nDates_linear_regression_combined_wnames, y=
bag_fish_subsetted_nDates_label_sc, by= "species_code")

#N DATES OF FISH OBSERVATION ACROSS BAYS: FULL GRAPH
bag_fish_nDates_n_pages <- ceiling(
  length(unique(bag_fish_subsetted_nDates_linear_regression_final$common_name))/12
)
pdf("C:/Users/arespostale/Dropbox/Occupancy/Linear_Regressions/Graphs/Subsetted/bag_fish_
nDates_common.pdf")
for(i in seq_len(bag_fish_nDates_n_pages)) {
  print(ggplot(data=bag_fish_subsetted_nDates_linear_regression_final, aes(x= major_area, y= n,
group=common_name))+
        geom_point()+
        facet_wrap_paginate(~common_name, nrow= 3, ncol= 4, page = i)+
        stat_smooth(fullrange=TRUE, method = lm)+
        geom_label(data=bag_fish_subsetted_nDates_label_common, aes(x= 0.5, y= 400,
label=V3, vjust="inward", hjust="inward"), parse = TRUE, inherit.aes=FALSE)+
        labs(title = 'Species Observation over Bays',
            x = 'Bay Number',
            y = 'Number of Observations')+
        scale_x_continuous(expand=c(0,0), limits=c(-1,9)) +
```

```
        scale_y_continuous(expand=c(0,0), limits=c(-50,450)) +
        coord_cartesian(xlim=c(0,8.5), ylim=c(0,420)) +
        theme_bw() +
        theme(panel.grid.major.x = element_blank(),
            panel.grid.minor.x = element_blank(),
            panel.grid.major.y = element_blank(),
            panel.grid.minor.y = element_blank()))
}
dev.off()

#DATA FRAME WITH ONLY FISH OBSERVED BETWEEN 40%-70% OF THE TIME
bag_fish_nObservationsPerYearwProportions <- merge(bag_fish_nObservationsPerYear,
bag_fish_species_proportions[ , c("species_code", "Proportion")], by = "species_code")

#DATA FRAME WITH ONLY FISH OBSERVED BETWEEN 40%-70% OF THE TIME BUT
WITH BAYS
bag_fish_subsetted_nObservationsPerYear <-
subset(bag_fish_nObservationsPerYearwProportions, Proportion >= 0.4 & Proportion <= 0.7)

#SUBSETTED SPECIES NAMES AND NUMBER ONLY
bag_fish_subsetted_names_partial <-
bag_fish_subsetted_nObservationsPerYear[!duplicated(bag_fish_subsetted_nObservationsPerYe
ar[c("species_code")]),]
```

```
bag_fish_subsetted_names <-  bag_fish_subsetted_names_partial[,c("species_code",
"common_name", "latin_name")]

#DATA TABLES FOR N FISH OBSERVATIONS OVER TIME (IN YEARS)
bag_fish_subsetted_nObservationsPerYear_table <-
data.table(bag_fish_subsetted_nObservationsPerYear)[,list(n), keyby = c('Year', 'species_code')]

#ADD IN DATA POINTS FOR MISSING FISH OBSERVATIONS IN EACH BAY
bag_fish_subsetted_nObservationsPerYear_nomissingvalues <-
bag_fish_subsetted_nObservationsPerYear_table[CJ(unique(Year), unique(species_code))]
#CHANGE NA VALUES TO 0
bag_fish_subsetted_nObservationsPerYear_noNA <- as.data.table
(lapply(bag_fish_subsetted_nObservationsPerYear_nomissingvalues, function(d) { d[is.na(d)] <-
0; d }))

#LINEAR REGRESSION OF N FISH OBSERVATIONS OVER TIME (IN YEARS)
bag_fish_subsetted_nObservationsPerYear_linear_regression <-
bag_fish_subsetted_nObservationsPerYear_noNA %>%
  group_by(species_code) %>%
  do(tidy(lm(n ~ Year, data= .)))
```

```
#SAVE THE N FISH OBSERVATIONS OVER TIME (IN YEARS) LINEAR REGRESSION TABLES IN EXCEL
write.csv(x=bag_fish_subsetted_nObservationsPerYear_linear_regression, file="C:/Users/arespostale/Dropbox/Occupancy/Linear_Regressions/Subsetted/Regression_Trend/bag_fish_subsetted_nObservationsPerYear_linear_regression.csv")

#FIT THE VALUES OF THE LINEAR REGRESSION OF N FISH OBSERVATIONS OVER TIME (IN YEARS)
bag_fish_subsetted_nObservationsPerYear_fit_linear_regression <-
bag_fish_subsetted_nObservationsPerYear_noNA %>%
  group_by(species_code) %>%
  do(fit=lm(n ~ Year, data= .))

#R-SQUARED AND ADJUSTED R-SQUARED VALUES FOR THE LINEAR REGRESSION OF N FISH OBSERVATIONS OVER TIME (IN YEARS)
bag_fish_subsetted_nObservationsPerYear_fit_linear_regression_r2 <-
bag_fish_subsetted_nObservationsPerYear_fit_linear_regression %>%
  glance(fit)
```

```
#SAVE THE R-SQUARED OUTPUTS FOR THE LINEAR REGRESSION OF N FISH OBSERVATIONS OVER TIME (IN YEARS)
write.csv(x=bag_fish_subsetted_nObservationsPerYear_fit_linear_regression_r2, file="C:/Users/arespostale/Dropbox/Occupancy/Linear_Regressions/Subsetted/R2_Values/bag_fish_subsetted_nObservationsPerYear_fit_linear_regression_r2.csv")

#N DATES OF FISH OBSERVATION ACROSS BAYS: SUBSET ONLY INTERCEPT AND SLOPE VALUES
bag_fish_nObservationsPerYear_linear_regression_subsetted_3col <- subset(bag_fish_subsetted_nObservationsPerYear_linear_regression, select=c(species_code, term, estimate))

#N DATES OF FISH OBSERVATION ACROSS BAYS: FIX INTERCEPT AND SLOPE TABLE
bag_fish_nObservationsPerYear_linear_regression_subsetted_spread <- spread(bag_fish_nObservationsPerYear_linear_regression_subsetted_3col, key=term, value=estimate)
    #MAKE INTERCEPT AND SLOPE HAVE THEIR OWN COLUMNS INSTEAD OF BEING ONE COLUMN TOGETHER.
colnames(bag_fish_nObservationsPerYear_linear_regression_subsetted_spread) <- c("species_code", "intercept", "slope")
    #CHANGE COLUMN TITLES TO BE INTERCEPT AND SLOPE TO AVOID CONFUSION.
```

```
#N DATES OF FISH OBSERVATION ACROSS BAYS: TABLE WITH SPECIES_CODE,
INTERCEPT, SLOPE, R2, AND P VALUES
bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined <- merge(x=
bag_fish_nObservationsPerYear_linear_regression_subsetted_spread, y=
bag_fish_subsetted_nObservationsPerYear_fit_linear_regression_r2[, c("species_code",
"adj.r.squared", "p.value")], by= "species_code")

#N DATES OF FISH OBSERVATION ACROSS BAYS: TABLE WITH YEAR,
SPECIES_CODE, N, INTERCEPT, SLOPE, R2, AND P VALUES
bag_fish_nObservationsPerYear_linear_regression_subsetted_combined <- merge(x=
bag_fish_subsetted_nObservationsPerYear_noNA[, c("Year", "species_code", "n")], y=
bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined, by=
"species_code")

#ADDING COMMON NAMES TO THE PARTIALLY AND COMBINED TABLES
bag_fish_subsetted_nObservationsPerYear_linear_regression_partially_combined_wnames <-
merge(bag_fish_subsetted_names,
bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined, by=
"species_code")
bag_fish_subsetted_nObservationsPerYear_linear_regression_combined_wnames <-
merge(bag_fish_subsetted_names,
bag_fish_nObservationsPerYear_linear_regression_subsetted_combined, by= "species_code")
```

```
#N DATES OF FISH OBSERVATION ACROSS BAYS: GEOM_LABEL TEXT CREATION FUNCTION
bag_fish_subsetted_nObservationsPerYear_lm_eqn =
function(bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined_wnames){
  intercept <-
bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined_wnames["intercept"]
  plusminus <- ifelse(sign(intercept) >= 0,
                      paste0(" + "),
                      paste0("")    )
  eq1 <- substitute(paste(italic(y) == m* italic(x),plusminus, b),
                    list(b =
format(bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined_wnames$intercept, digits = 3),
                         plusminus = plusminus,
                         m =
format(bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined_wnames$slope, digits = 3)))
  eq2 <- substitute(paste(italic(r)^2~"="~r2),
                    list(r2 =
format(bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined_wnames$adj.r.squared, digits = 4)))
```

```
  eq3 <- substitute(paste(italic(p) ~"="~p.val),
                    list(p.val =
format(bag_fish_nObservationsPerYear_linear_regression_subsetted_partially_combined_wnam
es$p.value, digits = 2)))
  c( as.character(as.expression(eq1)), as.character(as.expression(eq2)),
as.character(as.expression(eq3)))
}
bag_fish_subsetted_nObservationsPerYear_label_sc <-
ddply(bag_fish_subsetted_nObservationsPerYear_linear_regression_partially_combined_wname
s,.(species_code),bag_fish_subsetted_nObservationsPerYear_lm_eqn)
bag_fish_subsetted_nObservationsPerYear_label_common <-
ddply(bag_fish_subsetted_nObservationsPerYear_linear_regression_partially_combined_wname
s,.(common_name),bag_fish_subsetted_nObservationsPerYear_lm_eqn)

#N DATES OF FISH OBSERVATION ACROSS BAYS: TABLE WITH ALL VALUES
bag_fish_nObservationsPerYear_linear_regression_subsetted_final <- merge(x=
bag_fish_subsetted_nObservationsPerYear_linear_regression_combined_wnames, y=
bag_fish_subsetted_nObservationsPerYear_label_sc, by= "species_code")

#N DATES OF FISH OBSERVATION ACROSS BAYS: FULL GRAPH
bag_fish_nObservationsPerYear_subsetted_n_pages <- ceiling(
length(unique(bag_fish_nObservationsPerYear_linear_regression_subsetted_final$common_na
me))/12
```

```
)
pdf("C:/Users/arespostale/Dropbox/Occupancy/Linear_Regressions/Graphs/Subsetted/bag_fish_
nObservationsPerYear_common.pdf")
for(i in seq_len(bag_fish_nObservationsPerYear_subsetted_n_pages)) {
  print(ggplot(data=bag_fish_nObservationsPerYear_linear_regression_subsetted_final, aes(x=
Year, y= n, group=common_name))+
        geom_point()+
        facet_wrap_paginate(~common_name, nrow= 3, ncol= 4, page = i)+
        stat_smooth(fullrange=TRUE, method = lm)+
        geom_label(data=bag_fish_subsetted_nObservationsPerYear_label_common, aes(x=
1982, y= 80, label=V3, vjust="inward", hjust="inward"), parse = TRUE, inherit.aes=FALSE)+
        labs(title = 'Species Observation over Bays',
             x = 'Bay Number',
             y = 'Number of Observations')+
        scale_x_continuous(expand=c(0,0), limits=c(1950,2100)) +
        scale_y_continuous(expand=c(0,0), limits=c(-50,100)) +
        coord_cartesian(xlim=c(1981,2017), ylim=c(0,85)) +
        theme_bw() +
        theme(panel.grid.major.x = element_blank(),
              panel.grid.minor.x = element_blank(),
              panel.grid.major.y = element_blank(),
              panel.grid.minor.y = element_blank()))
}
```

```
dev.off()
```

# APPENDIX E

# CHAPTER IV DATA TRANSFORMATIONS

```
#DETERMINING WHAT BAYS I'LL USE
#ONLY KEEP ONE SAMPLE PER STATION
one_station_observation <- netw0[!duplicated(netw0[c("sample_id", "Year","major_area")]),]

#SHOWS THE NUMBER OF STATIONS (DATES OF SAMPLES) IN EACH YEAR (90
TIMES PER YEAR X 35 YEARS= 3150 MAX)
nstations_PerYear <- one_station_observation %>%
  group_by(major_area, Year) %>%
  tally

#SHOWS THE NUMBER OF STATIONS (DATES OF SAMPLES) FOR EACH BAY
ACROSS ALL YEARS
nstations_BayTotal <- one_station_observation %>%
  group_by(major_area) %>%
  tally

#DETERMINING WHICH FISH I'LL USE
#TABLE OF JUST BAY X AND ALL FISH
B5 <- subset(netw0, species_code <= 1800 & major_area == 5)
B6 <- subset(netw0, species_code <= 1800 & major_area == 6)
B7 <- subset(netw0, species_code <= 1800 & major_area == 7)
```

```
B8 <- subset(netw0, species_code <= 1800 & major_area == 8)
#CODE TO DETERMINE WHICH FISH IS OBSERVED EVERY SAMPLING PERIOD (90
TIMES PER YEAR X 35 YEARS= 3150 MAX; 1575 50% OBSERVATION)
nfish_B5 <- B5 %>%
  group_by(major_area, species_code) %>%
  tally

nfish_B6 <- B6 %>%
  group_by(major_area, species_code) %>%
  tally

nfish_B7 <- B7 %>%
  group_by(major_area, species_code) %>%
  tally

nfish_B8 <- B8 %>%
  group_by(major_area, species_code) %>%
  tally

                    #GROUP BY FISH
#SUBSET
SC610 <- subset(netw0, species_code == 610 & (major_area == 7 | major_area == 8))
SC629 <- subset(netw0, species_code == 629 & (major_area == 7 | major_area == 8))
```

```
#NORMALIZE (CAN COMPARE WHAT OVERALL DATA LOOKS LIKE BASED ON WHICH SPECIES DATA COMES FOR TO CHECK ACCURACY AND CAN COMPARE OVERALL SPECIES ENVIRONMENTAL PREFERENCE DIFFERENCE)
SC610_Z <- as.data.frame(normalize(SC610, byrow = FALSE))
SC629_Z <- as.data.frame(normalize(SC629, byrow = FALSE))

#Separate By Bay (Can compare what overall data looks like for each Bay based on which species data comes for to check accuracy)
SC610_Z_B7 <- subset(SC610_Z, major_area <= 0)
SC610_Z_B8 <- subset(SC629_Z, major_area >= 0)
SC629_Z_B7 <- subset(SC629_Z, major_area <= 0)
SC629_Z_B8 <- subset(SC629_Z, major_area >= 0)

#GROUP ALL DATA TOGETHER
#SUBSET
ALL <- subset(netw0, (species_code == 610 | species_code == 629) & (major_area == 7 | major_area == 8))

#NORMALIZE
ALL_Z <- as.data.frame(normalize(ALL, byrow = FALSE))

#SEPARATE BY FISH
```

```
ALL_Z_SC610 <- subset(ALL_Z, species_code <= 0)

ALL_Z_SC629 <- subset(ALL_Z, species_code >= 0)

#SEPARATE BY BAY

ALL_Z_B7 <- subset(ALL_Z, major_area <= 0)

ALL_Z_B8 <- subset(ALL_Z, major_area >= 0)


#SEPARATE BY BAY AND FISH

ALL_Z_B7SC610 <- subset(ALL_Z, species_code <= 0 & major_area <= 0)

ALL_Z_B8SC610 <- subset(ALL_Z, species_code <= 0 & major_area >= 0)

ALL_Z_B7SC629 <- subset(ALL_Z, species_code >= 0 & major_area <= 0)

ALL_Z_B8SC629 <- subset(ALL_Z, species_code >= 0 & major_area >= 0)
```