A framework for evaluating complex environmental monitoring data and application to state-wide, long-term nongame fish and algae monitoring in Kansas.

by

Victoria Reed

B.S., Kansas State University, 2023

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Carl and Melinda Helwig Department of Biological and Agricultural Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2023

Approved by:

Co-Major Professor
Dr. Martha Mather

Approved by:

Co-Major Professor
Dr. Trisha Moore

Copyright

© Victoria Reed 2023.

Abstract

Environmental monitoring data provide resource managers with broad snapshots of environmental quality and/or species distributions and are essential to understand the current state and trends in ecosystems and natural resources they provide. However, the very nature of environmental monitoring data, with their breadth in time and space, makes them notoriously difficult to statistically analyze. In this thesis, two environmental monitoring datasets from the State of Kansas are considered: algae and broader water quality in Kansas' reservoirs and nongame fish and habitat in Kansas streams. The aim of this research is to introduce and apply a framework for systematically selecting, exploring, analyzing, and testing environmental datasets such as these. More specifically, this research uses long-term, state-wide datasets documenting water quality conditions, fish presence and absence, and fish habitat data in Kansas streams, lakes, and reservoirs to characterize the current conditions and relationships between water quality and cyanobacterial abundance in reservoirs and as well as fish presence with stream habitat characteristics. Ultimately, this thesis will contribute to understanding of the conditions that drive cyanobacterial blooms and subsequent effects on reservoir ecosystems. In addition to this, it will provide an understanding of the habitat variables that are associated with fish presences compared to absences. The goal of this data analysis was not to come to a single significant conclusion, but more so an approach to process and review variables in state-wide databases to gain a better understanding of what the monitoring data can provide. Then, using this information, provide a systematic framework for a comprehensive, responsible, and a useful approach to monitoring data.

Table of Contents

List of Figures

List of Tables

Acknowledgements

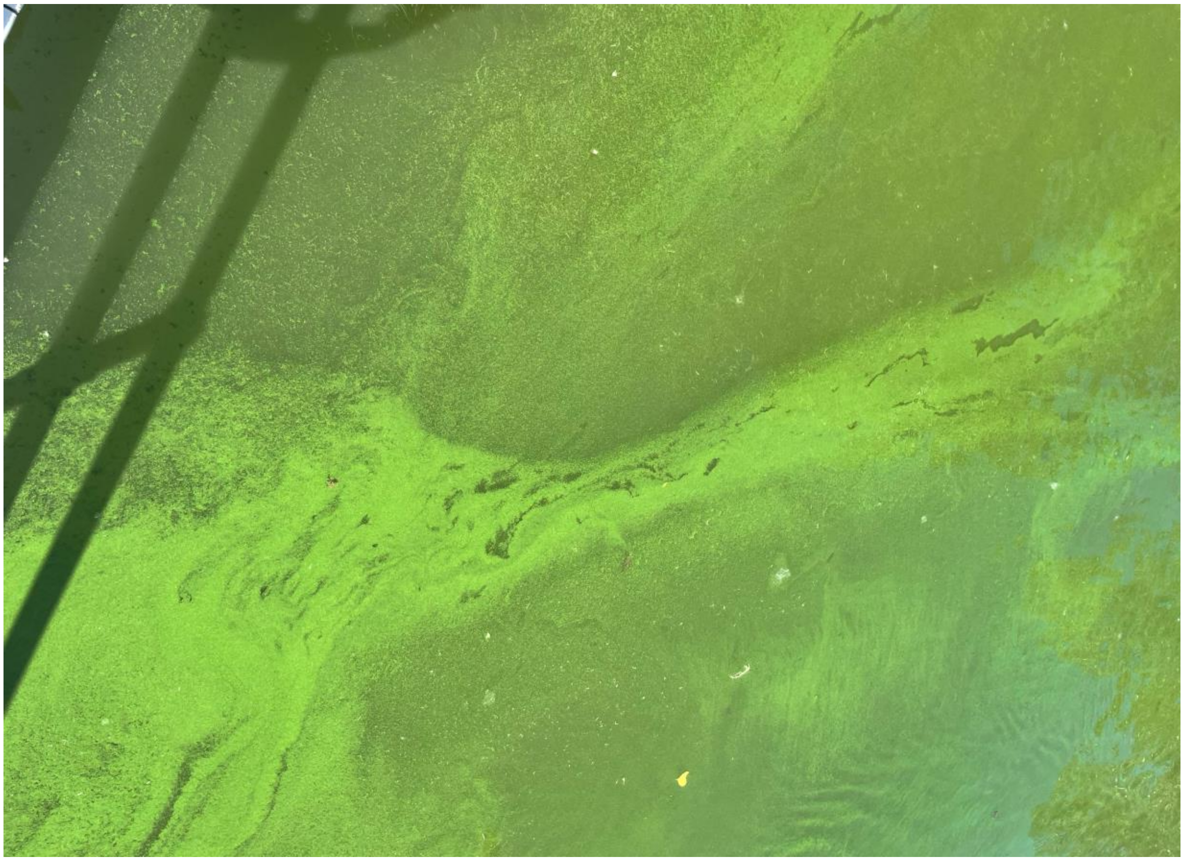
I cannot express enough thanks to my advisors and committee members, Dr. Trisha Moore, my main advisor, Dr. Martha Mather, my co-advisor, and Dr. Kate Nelson, my committee chair, for their continued support, guidance, and encouragement.

Chapter 1 - General Introduction

1.1. Introduction

Cyanobacteria, more commonly known as blue-green algae, are photosynthetic bacteria that are found in both fresh waters and salt waters. They are millions of years old and are thought to be responsible for the oxygenation of the Earth's atmosphere (Garcia-Pichel, 2009). There are thousands of species of cyanobacteria, as well as a variety of cell types and cellular structures (Vincent, 2009). Cyanobacteria use CO_2 as their primary carbon source, but do not rely on a fixed source of carbon. This enables the bacteria to be dispersed throughout the body of water they occupy (Percival, et al, 2014). Cyanobacteria are very resilient, and can survive in extreme environments, such as deserts and hot springs. There are known conditions that cyanobacteria do thrive in. These conditions include warm, slow-flowing water, sunlight, and an excess of nutrients, such as nitrogen and phosphorus (Bouma-Gregson, et al., 2013). Due to these conditions and Kansas weather, algae blooms are more likely to occur during the late spring to early fall months, when temperatures are high and there is plenty of sunlight (KDHE Bureau of Water, 2011). Cyanobacteria produce chlorophyll-*a* by synthesis, which is a measured parameter when testing for the bacteria in waterbodies (Garcia-Pichel, 2009). Chlorophyll-*a* is a compound that allows algae to photosynthesize (EPA, 2022). In addition to testing for chlorophyll-*a*, blue-green algae can be visually confirmed. Blue-green algae are characterized by the thick green material that is found along shorelines (Environmental Protection Agency, 2022). The algae can be seen by the naked eye when many cells form a colony, though a single cell is only identifiable under a microscope. A blue-green algae bloom can be seen below in Figure 1.

Figure 1.1 Blue-green algae in Marion Reservoir, July 2021 (by author)



At low concentrations blue-green algae are a standard part of lake and reservoir ecosystems. However, cyanobacteria can produce harmful toxins when they bloom or experience cell rupture. The cause of such algae blooms is still being explored but is believed to be a result of both natural and human factors. Natural factors include precipitation, wind, and increasing temperatures, while human factors include cultural eutrophication (nutrient runoff from agricultural fields and urbanization), pollution, and global warming. Diminished water quality leads to eutrophication. Eutrophication is the result of excess nutrients, usually nitrogen and phosphorus, entering the body of water through stormwater runoff and depleting the available dissolved oxygen content in the water. Depleted oxygen levels are a result of the increase in plant and algae growth due to the increase in available nutrients (NOAA, 2019). Depleted oxygen and an increase in cyanobacteria have impacts on both aquatic life and humans. Common side effects of encountering cyanotoxins for humans include neurotoxin poison and hepatotoxin poisoning. Neurotoxin poisoning affects the nervous system. Human symptoms of neurotoxin poisoning include numbness and dizziness, while animal symptoms include staggering, difficulty breathing, and death. Hepatotoxin poisoning affects the liver. Human symptoms of hepatotoxin poisoning include abdominal pain, diarrhea, and vomiting, while the primary animal symptom is death (Washington State Department of Health, n.d.). Due to the life-threatening impacts of blue-green algae blooms, it is essential to maintain, conserve, and protect freshwater.

1.2. Cyanobacterial toxins

It is important to note that not all species of cyanobacteria produce toxins, nor do all toxin producing species produce toxins under all conditions. There are three common genera of cyanobacteria that produce toxins: *Microcystis*, *Dolichospermum*, and *Planktothrix* (Environmental Protection Agency, 2022). The most common genus is *Microcystis*. *Microcystis* is essential to monitor as most species within these genera are capable of producing toxins and are commonly found in Kansas lakes and reservoirs. There are many different toxins that cyanobacteria can produce, but the most common toxins produced in the United States are microcystin, cylindrospermopsin, anatoxin, saxitoxin, and β-methylamino alanine (BMAA) (USGS, 2016). Microcystin is the toxin that affects the liver and anatoxin is the toxin that affects the nervous system (Bouma-Gregson, et al., 2013).

1.3. Fish populations in Kansas lakes

Not only do blue-green algae blooms affect humans and animals, but they also affect the entire ecosystems in which they are present. At high enough concentrations, algae blooms can lead to large fish kills. Since fishing is a major recreational activity in Kansas, it is essential to preserve the local fish populations. However, blue-green algae are not the only major threat. There are multiple habitat variables that affect where fish are located and whether they are being threatened. In recent years, many native Kansas fish species have become threatened and endangered or in need of conservation efforts. Therefore, further research is necessary to establish the types of conservation advantages that each habitat can provide.

1.4. History of cyanobacteria blooms in Kansas lakes

The state of Kansas has between 400 and 600 public lakes, most artificial, known as reservoirs. The terms lake and reservoir are commonly used interchangeably and will be for the purpose of this thesis. There is evidence to suggest some Kansas lakes and reservoirs have experienced consistent blue-green algae occurrences for many decades now (Carney, 2012); however, regular monitoring programs targeted to detecting and tracking cyanobacteria blooms was not established until 2010 (KDHE, n.d.). Before the general population was educated on algae and its appearance, many were unsure of what was in the water and exposed to the potential negative health effects. Blue-green algae was often mistaken for paint or oil spills in lakes (Carney, 2012). In 2010, KDHE established its lake monitoring program. More knowledge about blue-green algae and their respective characteristics has led to an increase in reports over the years, which can be seen in the table below.

Table 1.1 KDHE HAB Reports (from KDHE, 2022)

Year	Number of KS public water bodies with confirmed HAB	Number of KS public water bodies with confirmed toxic* HAB
2010	5	2
2011	20	7
2012	27	9

2013	15	4
l2014	19	9
2015	17	10
2016	16	6
2017	24	9
2018	28	8
2019	31	12
2020	7	3
2021	33	8
2022	31	8
***toxic defined as microcystin concentration greater than 4 ug/L or greater**		

The state of Kansas is responsible for taking proper measures when high cyanobacteria or toxin concentrations are detected to keep lake goers safe, including the closure of lakes at times.

1.5. Objectives of this study

1.5.1. Aim of cyanobacteria study

The aim of this research is to use a long-term, state-wide dataset documenting water quality conditions in Kansas lakes and reservoirs to characterize

(1) Current conditions and trends in water quality cyanobacteria abundance.

(2) Relationships between cyanobacteria abundance and physical and chemical features of these waterbodies.

Ultimately, this thesis will contribute to understanding of the conditions that drive cyanobacterial blooms and subsequent effects on reservoir ecosystems. The goal of this data analysis was not to come to a single significant conclusion, but more so an approach to process and review variables that might affect blue-green algae concentrations and determine if common variables thought to be of importance are associated with an increase in algal blooms. To achieve this for the blue-green algae dataset, an extensive review of previous published literature on the topic was performed. This literature review was used to inform hypotheses regarding biological, physical, and chemical parameters that increase the likelihood of a cyanobacteria bloom. Persistent factors that lead to increases in cyanobacteria blooms were hypothesized to be an increase in temperatures, nutrient input, and turbidity. After this, Kansas lake and reservoir phytoplankton community composition and chemistry data were obtained from the Kansas Department of Health and Environment (KDHE) via their Annual Lake Reports. The information on cyanobacteria and potential drivers was extracted from the reports and compiled into Microsoft Excel. Next, to analyze the data, it was put into JMP, where the data were visually inspected and then analyzed using correlation and regression analyses.

1.5.2. Aim of fish habitat study

The aim of this research is to gain insight on fish species distributions from local-scale habitat data. For the fish habitat and population dataset, field data that included habitat variables and the presence or absence of fish in streams was obtained from the Kansas Department of Wildlife and Parks (KDWP). Fifteen fish species were selected for evaluation based on their conservation status and habitat preferences. The analyses focused on the state of Kansas. The goal was to connect fish habitat variables to fish presence and absence data to understand distribution. Then, using this information, develop a prototype of fish species and their respective ideal habitats to provide a systematic framework for a comprehensive, responsible, and a useful approach to monitoring data.

Chapter 2 - Data Analysis of Cyanobacteria Occurrence and Potential Bloom Drivers in Kansas Lakes and Reservoirs

2.1. Introduction and objectives

The Kansas Department of Health and Environment (KDHE) began a lake and wetland monitoring program in 1975 as part of its efforts to uphold the Clean Water Act. Through this program, chemical, physical, and biological properties are monitored in hundreds of publicly owned lakes, including the state's 24 federal reservoirs. Each of these water bodies is monitored on a rotational basis every 3 or 6 years. This includes monitoring for algae and algae blooms, as well as potential bloom drivers. Some of the monitored parameters include temperature, chlorophyll-*a*, nitrate, kjeldahl nitrogen, phosphorus, water clarity (secchi depth), total algae cell count and biovolume, and enumeration of phytoplankton classes (green algae, blue-green algae or cyanobacteria, diatoms, and "other") by percent, cell count, and biovolume. Here, the "other" category lumps count of euglenoids, cryptophytes, dinoflagellates, and other single-cell flagellates. The stratification status of the water body at the time of monitoring is also determined from dissolved oxygen and temperature profiles measured at the deepest part of the water body (KDHE, 2022). From these data, water quality characterizations such as total nitrogen and total phosphorus ratio (TN/TP), and trophic state index (TSI) are calculated.

The information collected from these lakes is published every year by the KDHE in the Lake and Wetland Monitoring Annual Reports. A synthesis done by Dodds, Carney, and Angelo was used to set ecoregional reference conditions (2006). The evaluated conditions include total phosphorus, total nitrogen, secchi depth, and planktonic chlorophyll-*a*. To establish reference conditions in Kansas lakes and reservoirs, three methods were applied during the study to ensure that the values for all the parameters were true. Through this study, the Kansas lakes and reservoir reference conditions were established (Dodds, W., et al., 2006). The reference conditions found are listed below in Table 2.1. The values found for Kansas fall in the M-E (mesotrophic-eutrophic) range.

Table 2.1 Trophic thresholds for lakes and reservoirs worldwide (from Dodds, et al.)

Attribute	Source	O-M	M-E	E-H
Total Phosphorus (microgram*L^{-1})	Recommended	10	30	100
	Literature minimum	10	18	50
	Literature maximum	15	45	1500
Total Nitrogen (microgram*L^{-1})	Recommended	350	650	1200
	Literature minimum	140	180	1200
	Literature maximum	499	920	2940
Chl-*a* (microgram*L^{-1})	Recommended	3.5	9	25
	Literature minimum	2	5	18
	Literature maximum	4.3	10	40
Secchi (m)	Recommended	4	2	1
	Literature minimum	3.9	2	1
	Literature maximum	6	3	1.5

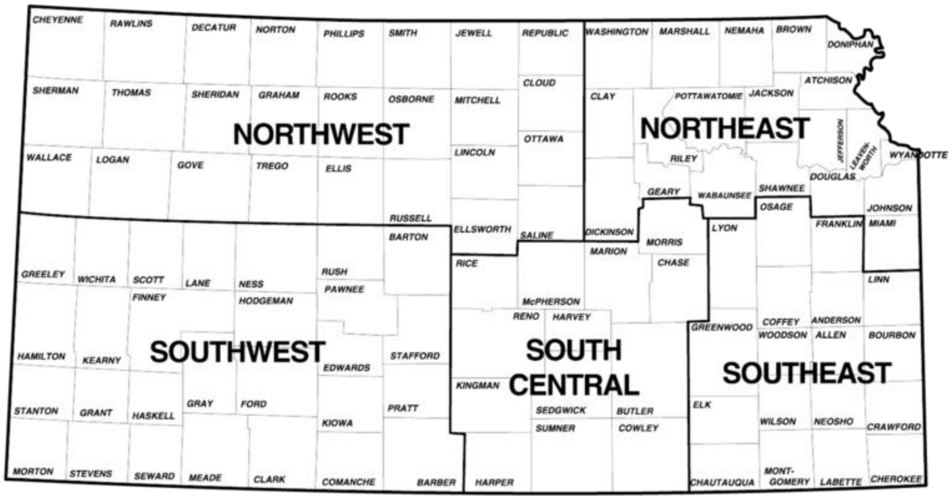
Using the values listed above as reference, water chemistry, physical properties, and phytoplankton communities from the KDHE dataset were evaluated to identify correlations between water quality conditions and cyanobacterial concentrations and associated blooms.

2.2. Data collection and analysis methods

2.2.1. Data collection

Before analysis of the data from the KDHE data, several variables that were not readily available in the Annual Reports were required. Since Kansas has a diverse weather pattern spanning the state from east to west, the specific region of the lake was thought to be of some importance. The state of Kansas is split into five regions shown below in Figure 2.1.

Figure 2.1 Regions Map (from ksoutdoors.com)



In addition to region, lakes and reservoirs are known to accumulate sediment over time, which increases the chance of eutrophication, so the age of the lakes at their sampling time was also deemed an important component for analysis. Once the year built for each reservoir was determined, the age of the lakes at the time they were sampled was calculated.

It is important to note what is considered a high or low concentration of blue-green algae for a body of water. The state of Kansas categorizes algae blooms into three thresholds: a watch, a warning, and a hazard. A watch is issued when blue-green algae cell counts are 80,000 cells/mL and/or the microcystin toxin level is 4 microgram/L. A warning is issued when blue-green algae cell counts are 250,000 cells/mL and/or the microcystin toxin level is 8 microgram/L. A hazard is issued when blue-green algae cell counts are 10,000,000 cells/mL and/or the microcystin toxin level is 2,000 microgram/L (KDHE, 2022). These thresholds were considered during data analysis.

The trophic status index, TSI, of a lake is an important characterization factor. This index was originally proposed by Carlson (1977) and has been established as a useful indicator on this to quickly understand the condition of a body of water. The trophic status index of a lake is found by using the following formula (Equation 1):

Equation 1 Trophic Status Index

$$TSI=10(6-(2.04-0.68\log_e C))/\log_e(2))$$

where C is the chlorophyll-*a* level in micrograms per liter.

A TSI score of 0-39 categorizes a lake as oligo-mesotrophic (OM), meaning it has a low level of planktonic algae and has a high level of water clarity. A TSI score of 40-49 categorizes a lake as mesotrophic (M), meaning it has a moderate level of planktonic algae, but still has a high level of water clarity. A TSI score of 50-64 categorizes a lake as eutrophic (E), meaning it has a moderate-to-large algae community. The eutrophic category is split into three subcategories: 50-54 is slightly eutrophic (SE), 55-59 is fully eutrophic (E), and 60-64 is very eutrophic (VE). The final category is hypereutrophic (H), which is split into lower hypereutrophic (LH), with a TSI score of 64-69.9 and upper hypereutrophic (UH), with a TSI score greater than 70. Hypereutrophic lakes have a very large phytoplankton population and have low water clarity (KDHE, 2022).

Total nitrogen and total phosphorus are commonly measured parameters as they are an indicator of excess nutrients in a reservoir. An excess of nutrients under specific conditions will result in increased plant and algae growth. The total nitrogen and total phosphorus ratio can be used to determine which nutrient is limiting or accelerating plant growth (KDHE, n.d.). Lakes that are nitrogen limited have a water column TN:TP ratio of less than 8, lakes that are co-limited by nitrogen and phosphorus have a TN:TP ratio between 9 and 21, and lakes that are phosphorus limited have a water column TN:TP ratio that is greater than 29 (Dzialowski et. al, 2005).

The most common and feasible way to indicator blue-green algae biomass is by measuring the amount of chlorophyll-*a* in the water. While total nitrogen and total phosphorus are variables known to lead to an increase in algae blooms, the excess chlorophyll-*a* in the water is caused by an algae bloom. Therefore chlorophyll-*a* is known as the response variable. The threshold set by the Kansas Department of Health and Environment is that the chlorophyll-*a* concentration should not exceed 10 micrograms/L for lakes and reservoirs. At a concentration above 10 micrograms/L, cyanobacteria dominance is assumed (KDHE Bureau of Water, 2011).

Finally, phytoplankton community data were obtained and examined from KDHE's Lake and Wetland Monitoring Program. The phytoplankton classes include green algae, blue-green algae or cyanobacteria, diatoms, and "other" by percentage, cell count, and biovolume. Here, the "other" category lumps count of euglenoids, cryptophytes, dinoflagellates, and other single-cell flagellates.

2.2.2. Data analyses and results

To compare different harmful algae bloom drivers, a statistical software, JMP, was used in combination with Microsoft Excel. Microsoft Excel was used to combine and clean data, while JMP was used for data exploration and visualization. JMP was used to analyze multivariate correlations, distributions, and to model the relationship between different driver variables.

The multivariate correlation analysis was performed using Spearman's Approach. A Spearman's rho coefficient (Ω) is given for each set of variables. A Ω value of 0.6 to 1 indicates a strong positive correlation between the two variables while a value of –0.6 to –1 indicates a strong negative correlation between the two variables. The variables assessed using multivariate correlation were TSI, TN/TP ratio, the age of lake at sampling time, total algae cell count, blue-green algae cell count, green algae cell count, diatom algae cell count, and other algae cell count. Both cell count and biovolume are collected for algae measurements, however biovolume was not collected until 1995, whereas the other data begins in 1990. Both biovolume and cell count are important measurements, as they show different characteristics of the algae. Cell count is useful when wanting to see the trends in algae concentrations in an area. It is also a better indicator of structural changes of the algae and reproduction time. Due to this, and the lack of complete biovolume measurements, the cell count was chosen for the analyses on this dataset. Table 2.2 shows the strong positive and strong negative results of the correlation analysis.

Table 2.2 Multivariate Correlation Analysis Results

Variable	By Variable	Spearman's Ω	Correlation Results
TSI	Age of lake	0.2423	No strong correlation
Total cell count	Age of lake	0.265	No strong correlation
Total cell count	TSI	0.6105	Strong positive
Green algae cell count	Age of lake	-0.0665	No strong correlation
Green algae cell count	TSI	-0.2952	No strong correlation
Green algae cell count	Total cell count	-0.4839	No strong correlation
Blue-green algae cell count	Age of lake	0.089	No strong correlation
Blue-green algae cell count	TSI	0.322	No strong correlation
Blue-green algae cell count	Total cell count	0.62	Strong positive
Blue-green algae cell count	Green algae cell count	-0.8267	Strong negative
Diatom algae cell count	Age of lake	-0.0737	No strong correlation
Diatom algae cell count	TSI	-0.1678	No strong correlation

Diatom algae cell count	Total cell count	-0.3792	No strong correlation
Diatom algae cell count	Green algae cell count	0.2935	No strong correlation
Diatom algae cell count	Blue-green algae cell count	-0.5908	Strong negative
Other algae cell count	Age of lake	-0.0206	No strong correlation
Other algae cell count	TSI	-0.1849	No strong correlation
Other algae cell count	Total cell count	-0.4263	No strong correlation
Other algae cell count	Green algae cell count	0.4687	No strong correlation
Other algae cell count	Blue-green algae cell count	-0.6353	Strong negative
Other algae cell count	Diatom cell count	0.3175	No strong correlation
TN/TP ratio	Age of lake	-0.1538	No strong correlation
TN/TP ratio	TSI	-0.2111	No strong correlation
TN/TP ratio	Total cell count	-0.0193	No strong correlation
TN/TP ratio	Green algae cell count	-0.0277	No strong correlation
TN/TP ratio	Blue-green algae cell count	0.1036	No strong correlation
TN/TP ratio	Diatom cell count	-0.1349	No strong correlation
TN/TP ratio	Other algae cell count	-0.0783	No strong correlation

From this analysis, there were not many strong positive or strong negative correlations. The strong positives were a correlation between total algae cell count and TSI, and blue-green algae cell count and total cell count. These are still important discoveries. This shows that the trophic status of a lake or reservoir does influence the total amount of algae in the water. It also shows that in Kansas lakes and reservoirs where algae are present, it is most likely blue-green algae. The strong negatives are blue-green algae and all the other algal types, green, diatom, and others, further showing that blue-green algae are probably the dominant algae in Kansas lakes and that it probably outcompetes the other types of algae (keeping their counts low) . The other variables did not show any strong correlations.

The JMP software was also used to show the distribution of the trophic status index, and the distribution of the age of lakes at their respective sampling times across all lakes sampled. Such analysis was performed using every lake that had been sampled by the KDHE. This includes 323 lakes in the Northeast region, 101 lakes in the Northwest region, 147 lakes in the South-central region, 351 lakes in the Southeast region, and 100 lakes in the Southwest region. To better analyze the data, a local filter was applied which divided the results into the five different regions in Kansas. Table 2.3 shows the average mean, average minimum, and average maximum of blue-green algae cell count by region. Table 2.4 below shows the average mean, average minimum, and average maximum TSI for the lakes in each region. Table 2.5 shows the distribution of age of the lakes at sampling time. Table 2.6 shows the distribution of TN/TP ratios by region.

Table 2.3 Distribution of Blue-Green Algae Cell Count by Region

Region	Avg. Minimum BG Cell Count (cells/mL)	Average BG Cell Count (cells/mL)	Avg. Maximum BG Cell Count (cells/mL)

Northeast	0	6,379	13,207,950
Northwest	0	4,259	3,754,009
South-central	0	1,151	544,485
Southeast	0	1,286	477,567
Southwest	0	2,054	1,444,275

From Table 2.3, the average blue-green algae cell counts for all regions in Kansas were well under the lowest threshold level, a watch, of 80,000 cells/mL. The average maximum blue-green cell count, however, is over both the hazard level and warning level for northeast Kansas, and is over the warning level, but does not quite reach the hazard level for northwest, south-central, southeast, and southwest Kansas. This indicates that on average, during the spring and summer seasons, lakes, and reservoirs in all regions of Kansas will experience at least one warning for blue-green algae.

Table 2.4 Distribution of TSI by Region

Region	Avg. Minimum TSI	Average TSI	Avg. Maximum TSI
Northeast	33	58	106
Northwest	35	60	88
South-central	35	59	83
Southeast	19	54	93
Southwest	32	62	88

From Table 2.4, the minimum average TSI for lakes and reservoirs in all regions of Kansas fall into the oligo-mesotrophic range. This indicates a low concentration of algae and increased water clarity. The average TSI for lakes and reservoirs in Kansas falls into the mesotrophic range. This indicates a moderate algae concentration, but most likely low water clarity. The average maximum TSI for lakes and reservoirs in Kansas falls into the hypereutrophic range. This indicates a very high algae concentration, a high concentration of excess nutrients and low water clarity. This analysis confirms the assumption that Kansas lakes and reservoirs generally fall into the mesotrophic range.

Table 2.5 Distribution of Age of Lakes at Sampling Time

Region	Avg. Minimum Age (years)	Average Age (years)	Avg. Maximum Age (years)
Northeast	1	42	106
Northwest	4	52	106
South-central	4	52	94
Southeast	7	45	137
Southwest	4	55	111

From Table 2.5, the average age of lakes in all regions of Kansas was relatively similar. This makes it hard to compare any difference in sedimentation or eutrophic state of the water due to the similarity in ages among Kansas lakes or reservoirs.

Table 2.6 TN/TP Ratio of Lakes at Sampling Time

Region	Avg. Minimum TN/TP Ratio	Average TN/TP Ratio	Avg. Maximum TN/TP Ratio
Northeast	0.5	16.6	365
Northwest	1.4	13.9	155
South-central	0.51	11	201
Southeast	0.52	16.5	329
Southwest	1.64	16.9	158

From Table 2.6, the average TN/TP ratio for all regions in Kansas shows a co-limitation of both nitrogen and phosphorus nutrients. The co-limitation of nitrogen and phosphorus is a common response of cyanobacteria to nutrient inputs in slow-flowing surface water (Savic, R, et al., 2022). This is most likely due to the cyanobacteria consumption of these nutrients. This shows that there is a need to limit nutrient input into lakes and reservoirs across Kansas.

2.3. Conclusions

The goal of this data analysis was not to come to a single significant conclusion, but more so an approach to process and review variables that might affect blue-green algae concentrations and determine if common variables thought to be of importance are associated with an increase in algal blooms. While some variables do follow the hypothesized trends, others do not. Monitoring data is essential for maintaining lakes and reservoirs in Kansas and it is usually collected with justifiable methodology. However, there are challenges to understanding and analyzing monitoring data. One of the main challenges, which was echoed through this data analysis is that one dataset cannot answer all questions. This dataset focuses on a few variables, with the awareness that cyanobacteria blooms can be influenced by many other variables not assessed. There is also a struggle to identify significance with such a broad set of data. While this analysis does work to identify possible bloom drivers, it will continue to be improved with each model created and evaluated.

Chapter 3 - Kansas Fish Population and Habitat

3.1. Introduction

Habitat influences fish species distribution and habitat variables have been widely used to examine patterns of fish presence and absence. Gerken et al. (2013) used monitoring data containing habitat variables such as substrate, stream size (width and depth), and land use variables to draw a correlation between habitat factors and the distribution of Topeka Shiners in Kansas streams. Similarly, using habitat variables such as stream size, slope, and land use, Wall et al. (2004) created multivariate logistic regression model for Topeka Shiner distribution that classified presences or absences. The model created accurately classified presences or absences considering the specific variables. Another study used water velocity, water depth, substrates, and channel units to understand the distribution of Freshwater Drum and Blue Catfish in the Middle Mississippi River for modifying and restoring fish habitats (Love et al., 2016). It is apparent that habitat has been used to draw a correlation to fish presence and absence for many years. Future efforts to connect habitat and fish species distributions would benefit from a standard procedure, or framework, for monitoring data that laid out specific steps and questions to guide data analysis.

In Kansas, the Kansas Department of Wildlife and Parks (KDWP) is charged with conserving and enhancing the state's natural wildlife and habitats, including native non-game fish. In line with this mission, KDWP maintains a monitoring program to document fish species distributions in streams and rivers throughout the state. In many of the locations where fish are sampled, habitat data are also collected. The monitoring of Kansas streams is essential to provide context for conservation and management. However, monitoring data can be difficult to analyze statistically, and thus to draw meaningful conclusions from which to form management plans to conserve species or critical habitat. In an effort to assist KDWP's efforts to conserve and manage nongame fish, a collaborative effort between the Kansas State University Department of Biology, Department of Geography and Geospatial Sciences, and Department of Biological and Agricultural Engineering, and KDWP was initiated. The goal of the project was to connect fish habitat variables to fish presence and absence data to understand statewide fish population. The specific objectives of the study include the development of a prototype model of fish species and their respective habitats in Kansas. With the development of such a model, a systematic framework is established for a comprehensive, responsible, and a useful approach to further understand and analyze monitoring data.

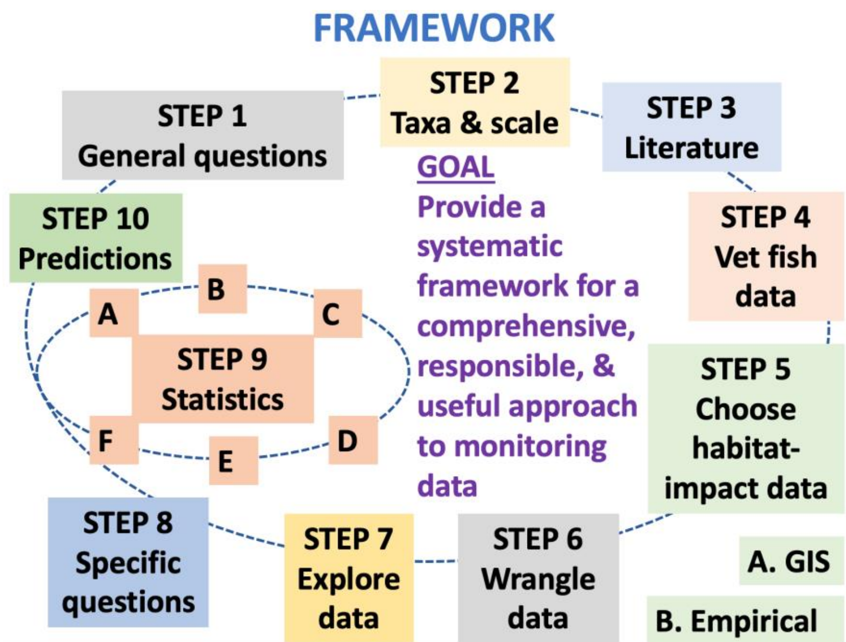
In the overall project, the collaborative team is considering both geospatial data as an indicator of habitat across a range in spatial scales and local empirical habitat data collected through KDWP's monitoring program. Both datasets have strengths and weaknesses; for example, variables such as land use best characterized in geospatial datasets while variables such as channel unit and in-stream substrate must be collected in the field. This chapter focuses only on local empirical monitoring data to gain insight into fish species distributions. The goal of the data analysis was not to come to a single significant conclusion or prediction model, but to provide and demonstrate an approach (series of steps, questions, interpretation guidelines, and connections) to integrate all available data, make testable predictions, and to suggest useful future data collections.

The primary objective of this study is the development of a framework that can be used to aid conservation of select fish species and their respective habitats in Kansas. Fish species in Kansas streams are diverse and so are their associated habitats. Human and climate impacts are constant stressors on fish habitats in Kansas. Some of these stressors include global climate change, extreme weather patterns, heavy agricultural activity, dams, and deforestation that often precedes urban development. The Kansas State Wildlife Action Plan (SWAP) addresses native fish habitat needs and related impacts (KDWP, 2022). The framework of evaluating empirical habitat variables and fish presence and absence will strengthen this plan.

3.2. Methodological framework

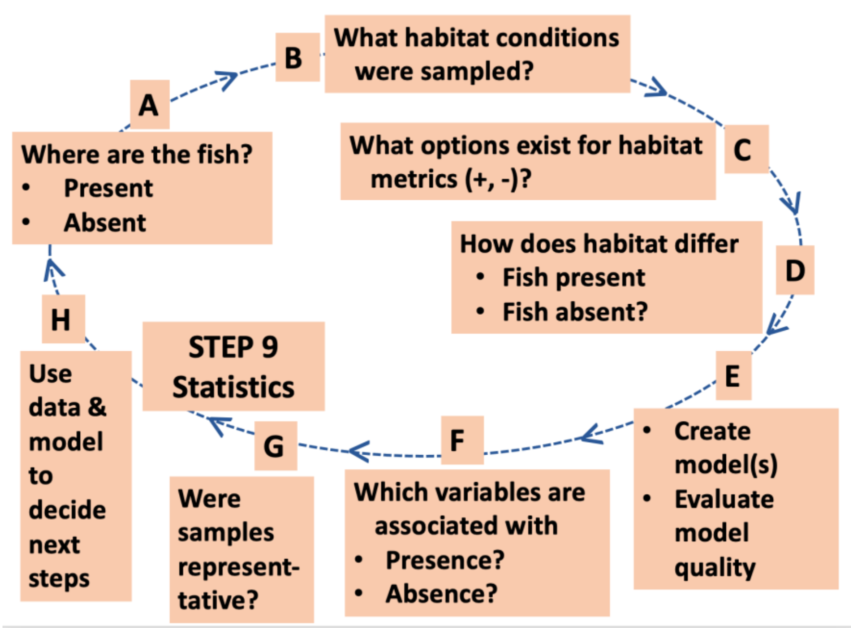
To develop a useful monitoring and data analysis approach, a framework was established that could be followed by any fisheries professional seeking to gain insights from biological monitoring datasets. The framework (Figure 3.1) is comprised of 10 major steps.

Figure 3.1 Methodological Framework (personal communication, Dr. Martha Mather)



The first step is to identify the objectives, or questions, to which an answer is needed about the dataset. In this case, what habitat variables have the greatest impact on which fish species presence? The second step is to identify the taxa, or species, to assess and the scale at which to evaluate. The third step is to identify any studies already completed, so that we can build on existing knowledge and the same approach isn't being unnecessarily repeated. This step also includes reviewing habitat-fish species associations from the literature to inform the types of habitats and/or impact data that should be considered. In step four, the monitoring data is vetted or otherwise validated. Within this vetted dataset, a set of habitat variables is selected and justified in step five. For steps six and seven, the data are “wrangled”, or put into the correct format to use in software or other programs, and then explored, respectively. This includes finding descriptions, units, and uses for the dataset. For step eight, a specific subset of statistical questions is formulated from the initial concerns and research objectives to the chosen set of data identified in step 1. Step nine is broken into six sub-steps (Figure 3.2) that help guide the data analysis phase. This phase includes using ArcMap to visualize where each fish species is present and absent. Then, histograms and box plots of each habitat condition sampled are created to look at the distribution and range within the sampled data. Next, for parsimony, similar variables are compared using box plots to decide which specific variable has the most versatility. Multiple logistic regression models for each fish species are then created.

Figure 3.2 Step 9, Statistical analysis subloop (personal communication, Dr. Martha Mather)



The tenth and final step is to make predictions from the statistical analyses completed in step nine, while also considering the literature in step three. The ten steps represent a prototype developed to connect fish habitat variables to fish presence and absence data to understand distribution, and provide a systematic framework for a comprehensive, responsible, and a useful approach to monitoring data. In the following sections, this framework will be applied to gain insight into observed fish species distributions in Kansas from local habitat monitoring data.

3.3. Application of framework to Kansas fish and habitat monitoring data

3.3.1. General questions

The general questions used in this study for Kansas fish and habitat data are as follows:

(1) Where are specific taxa present? Where are they absent?

(2) Can a model be built with presence and absence data?

(3) Can the model performance be evaluated?

(4) Can the model help to identify which habitat and impact variables are influential?

(5) Is local habitat data useful? If so, what types of local habitat data are useful?

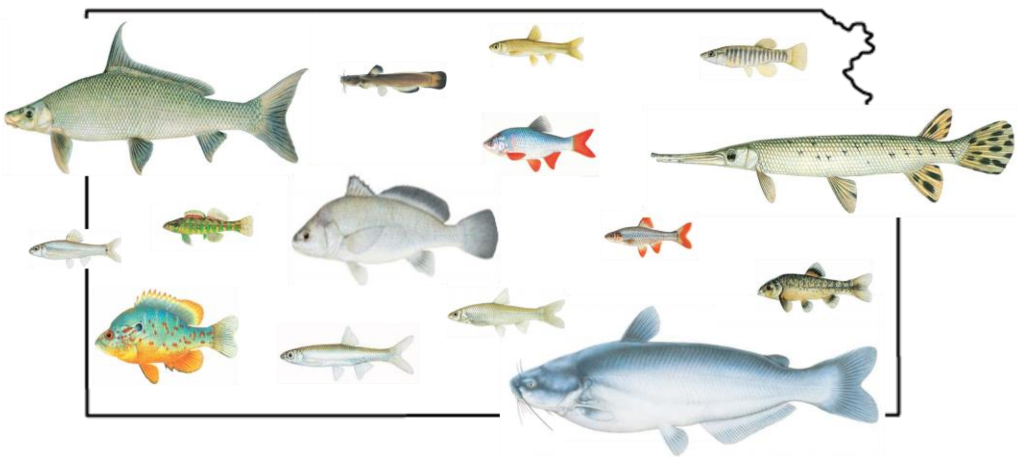
(6) Can the models predict where to sample next?

(7) Can a series of steps that provide a systematic framework for interpretation of multiple data analyses obtained through an ensemble analysis be established?

3.3.2. Taxa and scale

Fifteen fish species were chosen for analysis, including: Topeka Shiner, Neosho Madtom, Bullhead Minnow, Emerald Shiner, Plains Minnow, Blue Sucker, Banded Darter, Blue Catfish, Longnose Gar, Freshwater Drum, Central Stoneroller, Orange-spotted Sunfish, Red Shiner, Northern Plains Killifish, and the Brassy Minnow (Table 3.1;Figure 3.3). These species were chosen to represent a range in reproductive guilds, habitat preferences, diet, and conservation need. The inclusion of species of different needs was important to ensure flexibility of the framework and models. A summary of the fish and associated conservation needs is shown below in Table 3.1.

Figure 3.3 Fish Species Chosen for Analysis (personal communication, Olivia Rode)



A summary of the fish and associated conservation needs is shown below in Table 3.1.

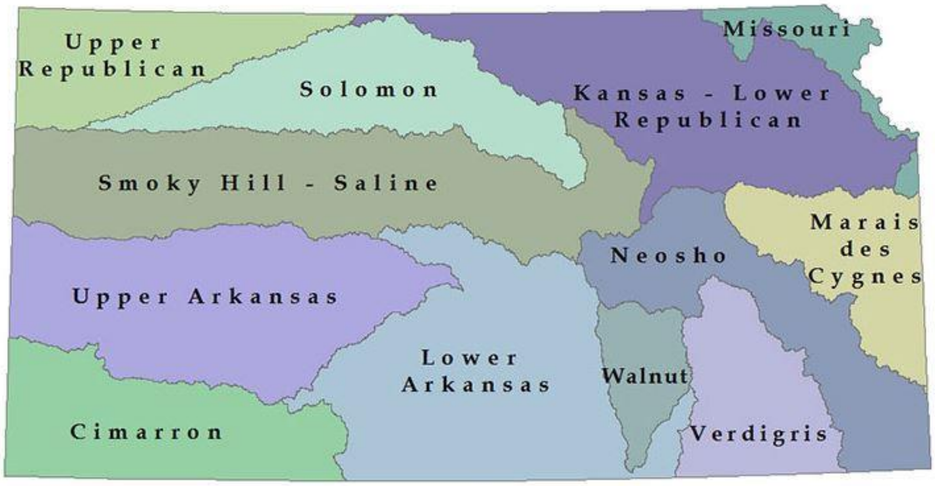
Table 3.1 Summary of Fish Characteristics

Common Name	**Scientific Name**	Priority
Topeka Shiner	Notropis topeka	Threatened & Endangered

Blue Sucker	Cycleptus elongatus	Species in need of Conservation (SINC)
Blue Catfish	Ictalurus furcatus	Common
Freshwater Drum	Aplodinotus grunniens	Common
Emerald Shiner	Notropis atherinoides	Common
Longnose Gar	Lepisosteus osseus	Common
Central Stoneroller	Campostoma anomalum	Common
Bullhead Minnow	Pimephales vigilax	Common
Orangespotted Sunfish	Lepomis humilis	Common
Red Shiner	Cyprinella lutrensis	Common
Plains Minnow	Hybognathus placitus	SINC
Brassy Minnow	Hybognathus hankinsoni	SINC
Neosho Madtom	Noturus placidus	Threatened & Endangered
Banded Darter	Etheostoma zonale	SINC
Northern Plains Killifish	Fundulus zebrinus	SINC

The entire state of Kansas was used to evaluate species presence and absence, as well as associations between each species and habitat variables. The state of Kansas can be divided into twelve major basins that could be used for a more in-depth analysis in the future (Figure 3.4). As for the temporal scale of analysis, the monitoring dataset included habitat observations collected between the 1970s and 2019, while corresponding fish data were collected between 1994 and 2019.

Figure 3.4 Study Area (from ksoutdoors.com)



3.3.3. Literature

The fifteen selected species differed by reproductive guilds, habitat preferences, diet, and conservation status in the state of Kansas (Eberle, et al.; Table 3.2).

Table 3.2 Fish Characteristics

Fish Species	Associated Habitat	Food Type	Spawning Condition
Topeka Shiner	Small streams, adjacent riparian area, clean gravel, cobble or coarse sand bottoms, clay hardpan, silt, quiet, open pools, cool temperatures, low fish diversity[1] Tolerates: range of temperatures	Flies, mayflies, midges, zooplankton, algae, plant material[1]	On Sunfish nests, gravel, rubble substrate, pools[1]
Blue Sucker	Swift water in large rivers, turbulent riffles, deep channels, large	Aquatic insects, clams, amphipods, copepods, algae[1]	Cobble or bedrock substrate, deep, swift riffles[1]

	woody debris, gravel, sand, hard clay[1]		
Blue Catfish	Large rivers, swift, deep channels, sand or coarser substrates[1]	Fish, zooplankton, insects, zebra mussels, crayfish[1]	Cavity nest[1]
Freshwater Drum	Large rivers and lakes, large woody debris, clean sand or gravel, deep water[1]	Insects, fish, crustaceans, clams, snails, midge larvae, crayfish[1]	May to July, deep pools, broadcast spawner[1]
Emerald Shiner	Big rivers, all depths, low gradient, moderate current, sandy[1] Tolerates: low oxygen, high turbidity, variety of substrates[1]	Fish eggs, insects, zooplankton, algae, midges, worms, amphipods[1]	Pelagic spawner, gravel, rubble or sand, late Spring-Summer[1]
Longnose Gar	Large rivers, deep water, slower pools[1]	Ambush predator, mainly fish, insects, small crustaceans[1]	Small streams, structure, or weedy shallow areas in lakes[1]
Central Stoneroller	Small streams, riffles and pools, gravel substrate[1] Somewhat tolerant of turbidity and siltation.[1] Not tolerant where sediment loads are high.[1]	Algae, diatoms, insects[1]	Nests on gravel riffles, early Spring[1]
Bullhead Minnow	Pools, sand, silt, shallow water, debris[1]	Detritus in stream bed, insects, invertebrates[1]	Under flat stones or debris (cover needed), shallow pools[1]

Orangespotted Sunfish	Pools, small or large streams, sand, gravel, mud[1]	Insects, crustaceans[1]	Gravel or sand nest, shallow water[1]
Red Shiner	Low gradient, shallow water, pools[1]	Invertebrates, fish eggs, algae[1]	Clean gravel, sand, or riffles, submerged rocks or plants[1]
Plains Minnow	Shallow streams with sand, diverse flow habitat, backwaters, pools[1] Tolerates: turbidity, low dissolved oxygen, temps up to 104 °F[1] Rarely occurs where there's no sand[1]	Algae, diatoms, ventral mouth[1]	Produce non-adhesive, semi-buoyant eggs[1]
Brassy Minnow	Small streams, low velocity pools, back waters, cover, sand, silt[1] Tolerates: high temperatures, low dissolved oxygen	Algae, detritus, bacteria[1]	Shallow, vegetated backwater[1]
Neosho Madtom	Riffles with gravel or pebble substrates, moderate depths[1]	Insects[1]	Gravel nests[1]
Banded Darter	Moderate streams with perennial flow, deep, turbulent riffles, large, course substrates, algae and plants[1] Does not occur in ponds or lakes.[1]	Midges, black flies, mayflies, small crustaceans[1]	Mats of algae, deep water[1]

Northern Plains Killifish	Shallow water, sand, pools, bank[1]	Insects, plant material[1]	Gravel, shallow water, Spring-Summer[1]

1. Eberle et al., 2014

3.3.4. Vet fish data

Local empirical habitat variables--collected from monitoring field work done by the state of Kansas-- were obtained from a datasheet supplied by the KDWP. KDWP has a stream assessment and monitoring program aimed at recording the range and distribution of stream species. Agency staff collected habitat data at more than 1675 sites between the 1970s and present day. The data at each site is collected across 10 to 11 transects, depending on the stream width. The transects are labeled A through K, with the A-transect located at the downstream end of the site, and the K-transect at the upstream end, putting the F-transect in the middle of the stream. Transects are spaced 10 m to 30 m apart such that the total length of surveyed sites is between a minimum of 150 meters and a maximum of 300 meters. A separate datasheet obtained from the KDWP included fish presence data. Prior to collecting habitat data at the site, fish are sampled by electrofishing. Usually, only one pass is made through the stream reach for fish sampling (KDWP, n.d.). Each species of fish found is recorded in the respective reach sampled. Each species not found at each sampling location was later added to the datasheet as an absence.

As part of the data vetting process, habitat data were examined to identify potential outliers or inconsistencies in entry of categorical habitat data. Ensuring consistent and correct data entry is crucial to subsequent data analysis steps. Due to the nature of data collection, the data are not always recorded the same way, depending on the collector. For example, in the habitat data, a simple “yes” or “no” is recorded to document sediment. It is recorded in many different formats though, including, “Y or N”, “yes or no”, “Yes or No”, etc. Similarly, categorical data such as substrate type or channel unit included entries that did not match the habitat datasheet key supplied by KDWP. When inconsistencies were detected, the habitat monitoring coordinator at KDWP was consulted to ensure the correct categorical data descriptors were assigned. Along with this, the channel unit codes provided.

3.3.5. Choose habitat-impact data

From the KDWP dataset, a subset of variables was chosen to further analyze the correlation between habitat and fish population as informed by the literature review of habitat needs (Table 3.2). The variables thought to be of major importance, and the empirical habitat data available for those categories, are shown in Table 3.3. Further justification for selecting these variables follows.

Table 3.3 Empirical Habitat Variables

Variable Category	Empirical Habitat Data
Size	Mean width, median width, mean depth, median depth
Gradient	Slope
Water flow	Discharge
Sinuosity	Compass bearing
Substrate	Most common, diversity index
Riparian Land Use	Sum riparian area, average riparian area
Channel Units	Most common, diversity index
Woody Debris	Total
Canopy Cover	Total
Cover (instream)	Filamentous algae, macrophytes, small brush, overhanging limbs, undercut banks, boulders, hard structures, total of all instream cover

Embeddedness	Percent of cover over substrates

3.3.5.1. Size

The size of the stream was characterized using both the width and depth measurements of the stream. The required width and depth vary based on fish species and geographic areas. Some fish species need colder water to survive, meaning they will need more depth. More depth also allows for more oxygen storage in the water (Kasco, 2023). In the KDWP habitat dataset, the size of a stream is determined by taking a wetted width measurement at each transect and dividing this measurement by four, creating five equally spaced points across the stream. At each point, a measurement is recorded for stream depth in centimeters and width in meters (KDWP, n.d.). The width and depth measures selected for subsequent analyses were total wetted width and median depth.

3.3.5.2. Gradient

The gradient of the stream is essentially the longitudinal slope from upstream to downstream at the sample site measured as the drop in elevation per unit distance. Slope can be related to elevation, substrate, and current flow. A higher slope can prevent a body of water from filling with sediment and organic matter, but a steeper slope can also lead to quicker erosion. If the substrate is a hard solid such as rock, the slope can be steeper because the rock won't erode as quickly as it would if there were a softer substrate in its place (Kasco, 2023). In the KDWP habitat monitoring dataset, the gradient is measured using a clinometer (a tool that measures angle of elevation or angle from the ground using a right-angled triangle at eye-level while standing at the edge of the stream (KDWP, n.d.). The resulting clinometer measurement, which has units of meter/meter, was utilized in subsequent data analyses.

3.3.5.3. Water flow

Water flow is important for fish and is assessed using current velocity and discharge. Here, water flow was measured as discharge (total volume of water moving down an entire stream or river channel cross section per unit time). Different species may have different preferences for faster versus slower-moving water. Discharge may affect the concentration of nutrients in the water, with higher concentrations tending to occur during low-flow periods. In the KDWP dataset, discharge is measured using a portable flowmeter at 30 equal distances (points) across the stream width to record velocity in 30 stream segments. The velocity and area, which is measured as the width and depth of the segment in which the velocity measurement was taken, are multiplied to obtain discharge in cubic feet per second (cfs). The 30 discharge measurements are added together for a total discharge of the stream at the monitoring site (KDWP, n.d.).

3.3.5.4. Sinuosity

The sinuosity of the stream, which is an indication of how much a stream meanders, or bends, is assessed by evaluating the compass bearing. The sinuosity of a stream can reflect habitat complexity with higher sinuosity at a meander, meaning there will be more water in that location and deeper pools (Net Map Tools, 2014). In the KDWP dataset, the sinuosity is recorded using an Azimuth compass in the middle of each transect of the stream (KDWP, n.d.). This provides a single indicator of sinuosity for each site that could range from 0 (very low sinuosity/straight channel) to 360 (highly sinuous) degrees.

3.3.5.5. Substrates

The substrates in the stream are classified by size and type. Substrate is an important habitat variable. Individual fish taxa rely on specific substrate types for spawning, predation, and shelter. Changes in substrate can influence growth and survival of fish in aquatic environments (Lynch 2019). The substrate in a stream is measured at the same set of 5 equally spaced points across each transect and at the same time the width and depth measurements are recorded (KDWP, n.d.). The dominant substrate types found at each observation point are classified as shown in Figure 3.5.

Figure 3.5 Substrate Size Class Code (KDWP Physical Habitat Form, communication Ryan Waters)

SUBSTRATE SIZE CLASS CODES
RS = BEDROCK (SMOOTH) *(LARGER THAN A CAR) RR = BEDROCK (ROUGH) *(LARGER THAN A CAR) BL = BOULDER (250 TO 4000 MM) *(BASKETBALL TO CAR) CB = COBBLE (64 TO 250 MM) *(TENNIS BALL TO BASKETBALL) GC = COARSE GRAVEL (16 TO 64 MM) *(MARBLE TO TENNIS BALL) GF = FINE GRAVEL (2 TO 16 MM) *(LADYBUG TO MARBLE) SA = SAND (0.06 TO 2 MM) *(GRITTY - UP TO LADYBUG SIZE) FN = SILT/CLAY, MUCK (NOT GRITTY) WD = WOOD (ANY SIZE) OT = OTHER (COMMENT - FLAG)

The substrate monitoring procedure results in 50 or more substrate type observations within each reach. This set of measurements was reduced in this analysis to two variables representing the most common substrate type encountered in the reach and the diversity of substrate types encountered in the reach. The most common substrate type was determined by identifying the substrate class with the highest number of occurrences at each site. In this analysis, we omitted rare substrates (i.e., boulders). The heterogeneity of substrates was characterized by the Shannon diversity index (Shannon, 1948). This diversity index provided a measure of the number of different types of substrates as well as the evenness of substrate distributions (e.g., Morris, et al., 2014). The Shannon diversity index can be found using the following formula (Equation 2):

Equation 2 Diversity Index

$$H=-\Sigma p_i*\ln(p_i)$$

where ln is the natural log and p_i is the proportion of species *i* relative to the entire community (Morris, et al., 2014).

Also measured with substrates is embeddedness (the extent to which hard substrates, i.e. gravel, cobble are covered or surrounded by fine substrate, i.e., silt, sand). It is measured with the same protocol as substrates, but it is taken as a percentage from 0-100%.

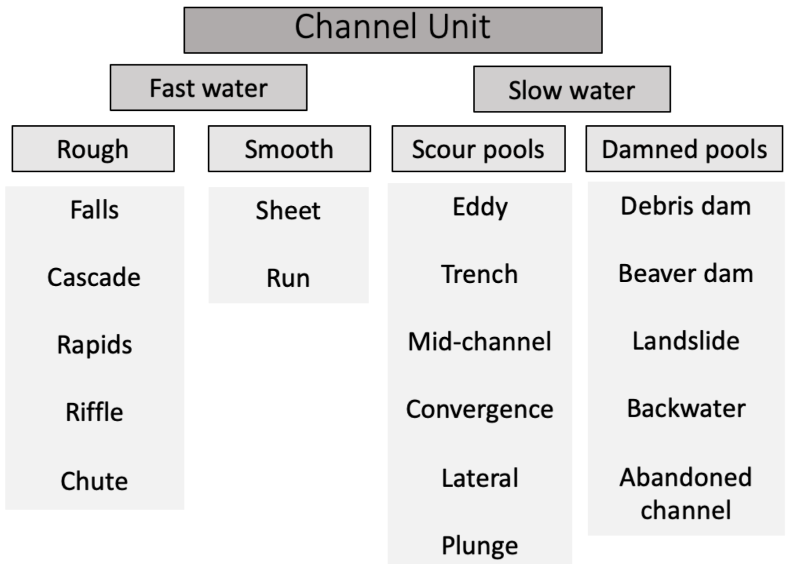
3.3.5.7. Riparian Land Use

The riparian land use variable assessed is riparian width. Riparian areas work to filter sediment, nutrients, and other matter before it reaches a stream. The presence of a contiguous riparian buffer with native vegetation is generally associated with higher in-stream habitat quality and biodiversity. For example, Poulton and Tao (2019) documented higher aquatic macroinvertebrate species richness in urban watersheds with intact riparian buffers relative to urban watersheds where native riparian cover had been disturbed. At each of the transects, boundaries are applied to the left and right banks, and the riparian width perpendicular to the stream is visually estimated with a maximum measurement of 50 meters (KDWP, n.d.). For this analysis, the average and sum of riparian width were considered.

3.3.5.8. Channel Units

Channel units are created by interactions between flow and roughness elements of the stream (Bisson, et al., 2017). The resulting bedforms range from those that form in faster moving water along steeper gradients (e.g., cascades and riffles) to slow moving areas with very low water surface slopes (e.g., pools). The more variety in channel units, the more diverse the habitat for fish will be. KDWP records observations of channel unit types within each habitat monitoring site at the thalweg (that is, at the deepest point in the stream cross section) at each transect (KDWP, n.d.). Channel units are classified by Bisson et al. (2017) as diagramed in Figure 3.6. The classification scheme used by KDWP is a simplification of Bisson et al. (2017) and includes pools, glides (and runs), riffle, rapid, cascade, falls, and dry channel. As with categorical substrate data, for this analysis, the multiple channel unit observations at each habitat monitoring site were aggregated to two variables to indicate the most commonly occurring unit (that is, the mode) and heterogeneity of channel units within the site (that is, by computing the Shannon Diversity Index given in Equation 2). In this analysis, we only used channel units that were common in Kansas rivers and streams (pools, glides, riffle, dry channel).

Figure 3.6 Channel Unit classification scheme (adapted from Bisson et al., 2017)



3.3.5.9. Woody Debris

Woody debris provides habitat for certain species of fish and so can be an important indicator of habitat quality. KDWP documents the relative number of woody debris at each stream monitoring site by taking a visual estimate at each transect. Woody debris presence is recorded on a scale of 0-4, with zero being absent and four being very heavy cover (KDWP, n.d.). To analyze the woody debris in a stream, the sum of all wood debris visual estimates were tabulated and used in subsequent analyses.

3.3.5.10. Instream Cover and Canopy Cover

The total cover of non-wood instream materials is also assessed as part of KDWP's stream habitat monitoring program. This includes algae, macrophytes, overhanging vegetation and undercut banks, and others. The cover is measured by taking a visual estimate. Similar to woody debris, other instream cover types are all recorded on a scale of 0-4 with zero being absent and four being very heavy cover (KDWP, n.d.). Cover can be important for feeding and avoiding predation by structure-oriented fish. However, cover on the surface of a stream can prevent sunlight from reaching plants, which will limit the oxygen being produced, and in-water structure can impede schooling and open-water fish. For this analysis, all important instream cover variables were summarized as sum, mean, and standard deviation of all instream cover variables.

Canopy cover, or extent to which the stream is shaded by vegetation, was also assessed. Canopy cover is measured using a densiometer. This is an important variable as it can provide shade and hence, cooler water. The variable set selected for further analysis consisted of the sum, mean, and standard deviation of the canopy cover variable.

3.3.6. Wrangle data

In "wrangling" the data, an important discrepancy that had to be corrected before analysis was matching the habitat data to latitude and longitude coordinates by their specific site identification, and then converting latitude and longitude data from degrees/minutes/seconds to decimal degrees. In addition, some habitat variables needed to be "collapsed" from multiple measurements taken across multiple transects at each site to a single value representative of the spatial scale (the site) represented by the fish data. For example, depth is measured at 5 locations along 10 to 11 transects at each site, resulting in at least 50 depth measurements. This collection of depth measurements was aggregated to the site scale by taking the median depth recorded at the site. Likewise, categorical variables used to describe channel units and substrate were recorded at multiple points within each monitoring site. In the case of such categorical variables, the mode was used to characterize the central tendency of the habitat variable. For example, in the case of channel units, KDWP records the type of channel unit at each of the points where depth is measured, resulting in 50 or more channel unit observations in a single reach. The most commonly occurring channel unit was determined from these multiple observations to produce a single value for the channel unit variable at the site scale.

3.4. Data analysis

3.4.1. Explore data

Locations with fish presence/absence and habitat monitoring data for the target fish species were displayed geographically using latitude-longitude data obtained from KDWP's stream monitoring program in ArcMap (ESRI). ArcMap is a software used to display geographic information on a map using layers that contain latitude and longitude information. The locations sampled for the chosen fish species in Kansas is shown in Figure 3.7, there were 1,675 locations in which both habitat and fish presence/absence data were collected. The fish monitoring data highly favors the eastern part of Kansas, as there are not as many bodies of water to sample in the western portion of the state due to its dry climate and historic dewatering of streams for groundwater mining.

Figure 3.7 Fish Species and Habitat Sampling Locations in Kansas

Presence and Absence Habitat Data

The next important step in exploring the data is understanding how much data is available and how the data is distributed. For the species chosen, the number of times they were present or absent can be seen in Table 3.4. For many of the fish species in this dataset, the number of times they were absent is more than the number of times they were present. This will be an important consideration when interpreting the data results.

Table 3.4 Number of Presences and Absences of Species

Species	Number of Presences	Number of Absences
Banded Darter	8	1667
Topeka Shiner	39	1636
Red Shiner	1473	204
Plains Minnow	60	1615
Orangespotted Sunfish	734	942
Northern Plains Killifish	435	1242
Neosho Madtom	2	1673

Longnose Gar	278	1397
Blue Catfish	4	1671
Blue Sucker	7	1668
Freshwater Drum	333	1343
Brassy Minnow	23	1652
Bullhead Minnow	409	1266
Emerald Shiner	144	1532
Central Stoneroller	1236	441

To further explore the data, the distribution of the habitat variables was assessed. For each habitat variable, the mean, median, and standard deviation were computed as summarized in Table 3.5.

Table 3.5 Descriptive Statistics of Habitat Variables

Habitat Variable	Mean	Median	Standard Deviation
H' Substrate	1.02	1.08	0.50
H' Channel Units	0.76	0.79	0.39
Width (m)	11.91	7.8	13.32
Depth (cm)	25.71	20.5	20.04
Slope (m/m)	0.20	0.1	0.30
Bearing (degrees)	135.39	125.82	62.45
Discharge (cfs)	53.46	3.25	109.89
Riparian Width (m)	33.85	46.36	20.41
Canopy Cover (count)	49.56	51	29.69
Instream Cover (count)	3.36	3.09	1.74
Woody Debris (count)	0.28	0.18	0.34
Embeddedness (%)	82.76	91.36	20.74

Boxplots and histograms of each habitat variable for this dataset were also assessed to understand the distribution of each variable. The boxplots and histograms can be found in Appendix A.

3.4.2. Specific questions

As introduced previously, the following are the specific questions to be answered through following this framework process.

(1) Where are the fish present? Where are the fish absent?

(2) What range of habitat conditions were sampled?

(3) What metric and scale options exist for habitat variables?

(4) How does habitat differ when fish data is present and absent?

(5) Can model(s) be created? Can the model(s) quality be evaluated?

(6) Can the importance of different habitat impact variables for a taxa be evaluated?

(7) Are data representative? Are there gaps? Are some areas oversampled?

(8) Can questions 1-7 be used to make predictions that guide future sampling?

3.4.3. Statistics

The statistics framework follows a series of eight steps based off the specific questions to be answered as listed above. The 15 fish species chosen for analyses were assessed according to these steps. The following section focuses on step six of the statistics subloop (which variables are associated with fish presence/absence) and will detail the data analysis and models of the habitat variables and their correlation to the fifteen species of fish chosen.

To create models for each species and the variables that affect them the most, multiple logistic regressions were ran using R Studio (Posit, 2023). Like other regression analyses, multiple logistic regression provides a predictive analysis, in this case regarding the likelihood that a given species will be present at a given site as a function of the site's habitat characteristics. The multiple logistic regression is a predictive analysis used to identify the relationship between one dependent binary variable (fish taxa presence or absence at a site) and many independent variables (habitat data in the form of nominal, ordinal, interval, or ratios) (Statistics Solutions, 2023). Multiple logistic regression is well-suited to this dataset as the predicted variable, presence, or absence, is binary (0 or 1) rather than continuous. Three types of multiple logistic regressions were run for each fish taxa. In all the models, random absences were removed to balance presence and absence. First, in the ecological model approach, the multiple logistic regression was run using one habitat variable from each of the ecological categories described in Section 3.3.5.

In the second quantitative approach, noncorrelated habitat variables were included as regressors. In this approach, multiple variables or no variables might be included from a single ecological category. This approach only considered the statistical role of variables.

In the third "reduced model" approach, the quantitative regression was re-run with a smaller set of variables retaining only regressors that showed significance ($p \leq 0.05$). The most reduced version of the models is shown below.

For significant regressors, probability plots were constructed to show the relationship between an individual regressor and the probability of a fish being present. Lastly, the model quality was assessed using a cross-validated confusion matrix.

For five of the 15 species, a multiple logistic regression model could not be run because of an inadequate number of sites at which the species was presence. These species include the Brassy Minnow, Neosho Madtom, Blue Catfish, Blue Sucker, and Banded Darter.

Lastly, the model quality was assessed using a cross-validated confusion matrix and probability plots (Appendix B) were made to indicate how the direction of a habitat variable affects species presence and absence.

3.4.3.1. Bullhead Minnow

The final model created by repeating logistic regressions until only the significant habitat variables were left for Bullhead Minnow is summarized in Table 3.6.

Table 3.6 Bullhead Minnow Model

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.26596	0.348494	-3.633	0.000281	***
Total Canopy	-0.00643	0.003318	-1.938	0.052663	.
Discharge	-0.0097	0.002272	-4.269	1.96E-05	***
Overhanging	-0.62901	0.155787	-4.038	5.4E-05	***
Riparian Width Sum	0.010573	0.002174	4.863	1.16E-06	***
Median Width	0.113778	0.01321	8.613	< 0.0000000000000002	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

From the model results above and probability plots (Appendix B), an increase in canopy cover, discharge, and overhanging vegetation decrease the likelihood of a Bullhead Minnow being present. An increase in width and riparian area increases the likelihood of a Bullhead Minnow occurrence. The confusion matrix showed this model being 73% accurate at predicting presences and absences.

From the literature, it is known that the Bullhead Minnow is associated with pools with a low discharge and sand or silt substrates (Eberle et al., 2014). A lack of data could be the reason some of these habitat variables didn’t show up as significant in the model. Out of 1,675 total samples, Bullhead Minnow’s were present in 409.

3.4.3.2. Brassy Minnow

The final model created by repeating logistic regressions until only the significant habitat variables were left for Brassy Minnow is summarized in Table 3.7.

Table 3.7 Brassy Minnow Model

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.8361	0.9065	-2.026	0.0428	*
H' Substrate	0.5017	0.5905	0.85	0.3955	
Total Cover	0.3618	0.1721	2.103	0.0355	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the model results above and probability plots (Appendix B), an increase in substrate diversity and total cover increase the likelihood of a Brassy Minnow occurrence. The confusion matrix showed this model being 61% accurate at predicting presences and absences. The Brassy Minnow species is listed as in need of conservation in Kansas and the lack of available data on the species could contribute to a lower accuracy at predicting a model. Out of 1,675 total samples, the Brassy Minnow was only present in 23.

From the literature, it is known that the Brassy Minnow is associated with deep pools, with sand and silt substrates, which is supported above in the model with the diversity index of substrates. They also need woody debris or vegetation to spawn, which is supported above in the model by total cover (Eberle et al., 2014). Since the literature and model were not conclusive, this was the cutoff point for creating models with low presences and shows that a lack of data can heavily influence biased model results. Models were not created for any fish species with less than 23 presences.

3.4.3.3. Central Stoneroller

The final model created by repeating logistic regressions until only the significant habitat variables were left for Central Stoneroller is summarized in Table 3.8.

Table 3.8 Central Stoneroller Model

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.729853	0.18463	-3.953	7.72E-05	***
Median Depth	-0.029207	0.004312	-6.774	1.25E-11	***
H' Substrate	1.88679	0.176689	10.679	< 0.0000000000000002	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

From the model results above and probability plots (Appendix B), an increase in depth decreases the likeliness of a Central Stoneroller being present. An increase in diversity index of substrates increases the likelihood of a Central Stoneroller occurrence. The confusion matrix showed this model being 68% accurate at predicting presences and absences. Out of 1,675 total samples, the Central Stoneroller was present in 1,236.

From the literature, it is known that the Central Stoneroller is associated with small streams, deep pools with gravel, and mainly feed off algae and spawn on gravel riffles (Eberle et al., 2014). When a model shows the diversity index of substrates being important, that indicates that more than the common substrates are important to this species. Kansas streams primarily have sand substrate, so the diversity index being important means other substrates such as gravel are important for a species, and this is certainly the case for the Central Stoneroller. The model also predicted that a decrease in depth will result in an increase in Central Stoneroller presence, which matches the literature.

3.4.3.4. Emerald Shiner

The final model created by repeating logistic regressions until only the significant habitat variables were left for Emerald Shiner is summarized in Table 3.9.

Table 3.9 Emerald Shiner Model

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.72245	1.408415	-2.643	0.008217	**
Total Canopy	-0.01903	0.006218	-3.06	0.002212	**
Bearing	0.007337	0.002874	2.553	0.010692	*
Embeddedness	0.039946	0.01263	3.163	0.001562	**
Total Cover	-0.43938	0.125119	-3.512	0.000445	***
Median Width	0.073776	0.015627	4.721	2.35E-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the model results above and probability plots (Appendix B), an increase in total cover and total canopy decreases the likelihood of an Emerald Shiner occurrence. An increase in width, bearing and embeddedness increases the likelihood of an Emerald Shiner occurrence. The confusion matrix showed this model being 76% accurate at predicting presences and absences. Out of the 1,675 total samples, only 144 had Emerald Shiners present, which could skew the model results.

From the literature, it is known that the Emerald Shiner is associated with large, sandy rivers, but is tolerant of many substrates. They have a terminal mouth, so they are not dependent on the bottom of streams (Eberle et al., 2014). The Emerald Shiner population in Kansas is not listed as threatened and endangered or a species in need of conservation, like many other pelagic spawners are.

3.4.3.5. Topeka Shiner

The final model created by repeating logistic regressions until only the significant habitat variables were left for Topeka Shiner is summarized in Table 3.10.

Table 3.10 Topeka Shiner Model

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.533826	1.907718	-2.377	0.017474	*
H' substrate	4.387315	1.260105	3.482	0.000498	***
Macrophytes	2.108569	0.714228	2.952	0.003155	**
Riparian Area Sum	-0.009298	0.010189	-0.913	0.361435	
Slope	-4.086611	2.376769	-1.719	0.085542	.
Woody Debris	-6.41421	1.957644	-3.276	0.001051	**
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

From the model results above and probability plots (Appendix B), an increase in woody debris shows a decrease in the likelihood of a Topeka Shiner occurrence. An increase in the diversity index of substrates and macrophytes show an increase in the likelihood of a Topeka Shiner occurrence. The confusion matrix showed this model being 65% accurate at predicting presences and absences. Out of 1,675 total samples, the Topeka Shiner was only present in 39.

From the literature, it is known that the Topeka Shiner is associated with small, cool temperature, and low discharge streams. They do have diverse substrates they're found in, including gravel, cobble, and clay, which matches the model. The Topeka Shiner spawns on Sunfish nests and primarily eat flies, algae, and plant material (Eberle et al., 2014), which indicates why macrophytes were found to be of importance for Topeka Shiners.

3.4.3.6. Red Shiner

The final model created by repeating logistic regressions until only the significant habitat variables were left for Red Shiner is summarized in Table 3.11.

Table 3.11 Red Shiner Model

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.469631	0.980074	-0.479	0.631811	
Total Canopy	-0.018168	0.005153	-3.526	0.000422	***
Embeddedness	0.023942	0.007425	3.225	0.001261	**
H' Substrate	0.892008	0.339921	2.624	0.008686	**
Total Cover	-0.588345	0.089217	-6.595	4.27E-11	***
Median Width	0.082096	0.02328	3.526	0.000421	***
Woody Debris	1.623977	0.499011	3.254	0.001136	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the model results above and probability plots (Appendix B), an increase in total cover and total canopy decrease the likelihood of a Red Shiner occurrence. An increase in the diversity index of substrates, width, embeddedness, and woody debris increase the likeliness of a Red Shiner occurrence. The confusion matrix showed this model being 73% accurate at predicting presences and absences. Out of 1,675 total samples, the Red Shiner was present in 1,473 samples. The Red Shiner is a common fish in Kansas and is increasing in numbers.

From the literature, it is known that the Red Shiner is associated with shallow pools, primarily feed on fish eggs and invertebrates, and spawn on a variety of substrates, including gravel, sand, and rock substrates, and plant material (Eberle et al., 2014). The Red Shiner is extremely tolerable of harsh conditions and will do good almost anywhere. The Red Shiner is an open water fish, so the model predicting that a decrease in total cover and canopy is expected. Woody debris showing importance is relevant as they serve as a host for invertebrates, which is a source of food for Red Shiners. The diversity index of substrates increasing the likeliness of a Red Shiner is also probably due to the fact that the fish are associated with gravel.

3.4.3.7. Longnose Gar

The final model created by repeating logistic regressions until only the significant habitat variables were left for Longnose Gar is summarized in Table 3.12.

Table 3.12 Longnose Gar Model

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.30333	0.547149	-6.037	1.57E-09	***
Total Canopy	-0.01488	0.005153	-2.887	0.003893	**
Median Depth	0.027786	0.006676	4.162	3.16E-05	***
H’ Substrate	1.381074	0.301092	4.587	4.5E-06	***
H’ Channel Unit	1.188341	0.334381	3.554	0.00038	***
Overhanging	-0.89424	0.230283	-3.883	0.000103	***
Median Width	0.068123	0.010667	6.386	1.7E-10	***
Woody Debris	1.036939	0.347487	2.984	0.002844	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

From the model results above and probability plots (Appendix B), an increase in the diversity index of substrates, the diversity index of channel units, the specified width, depth, and woody debris increase the likelihood of a Longnose Gar occurrence. An increase in overhanging vegetation and total canopy decreases the likelihood of a Longnose Gar occurrence. The confusion matrix showed this model being 79% accurate at predicting presences and absences. Out of 1,675 total samples, the Longnose Gar were present in 278 samples. The Longnose Gar species is stable, but the lack of samples is probably due to where sampling occurs.

From the literature it is known that the Longnose Gar is associated with large rivers but reside in the shallow areas or pools. The Longnose Gar is an ambush predator and need structure to hide behind. They feed on other small fish and need slow moving, weedy areas to spawn (Eberle et al., 2014). The positive associations with depth, width, and woody debris matches the literature.

3.4.3.8. Freshwater Drum

The final model created by repeating logistic regressions until only the significant habitat variables were left for Freshwater Drum is summarized in Table 3.13.

Table 3.13 Freshwater Drum Model

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.5337193	0.29641263	-8.548	< 0.0000000000000002	***
Median Depth	0.01480156	0.00542452	2.729	0.00636	**
Discharge	-7.238E-05	0.00001743	-4.152	0.000033	***
Macrophytes	-0.6213543	0.18915776	-3.285	0.00102	**
Riparian Area Sum	0.00949704	0.00248045	3.829	0.000129	***
Median Width	0.10236006	0.0117589	8.705	< 0.0000000000000002	***
Woody Debris	0.76525125	0.28181163	2.715	0.006618	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

From the model results above and probability plots (Appendix B), an increase in width, depth, woody debris, and riparian area increase the likelihood of a Freshwater Drum. An increase in macrophytes and discharge decrease the likelihood of a Freshwater Drum. The confusion matrix showed this model being 74% accurate at predicting presences and absences. Out of 1,675 total samples, the Freshwater Drum were present in 333 samples.

From the literature it is known that Freshwater Drum is associated with deep pools, tolerate a variety of substrates and large woody debris. They primarily feed on fish and fish eggs and are pelagic spawners (Eberle et al., 2014). The model agrees with literature for width, depth, woody debris, and an inverse relationship with discharge.

3.4.3.9. Orangespotted Sunfish

The final model created by repeating logistic regressions until only the significant habitat variables were left for Orangespotted Sunfish is summarized in Table 3.14.

Table 3.14 Orangespotted Sunfish Model

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.17427	0.206438	-10.532	< 0.000000 00000000 02	***
Median Depth	0.025751	0.003743	6.879	6.02E-12	***
Discharge	-0.00674	0.001711	-3.939	8.17E-05	***
H' Substrate	0.721443	0.123139	5.859	4.66E-09	***
Riparian Width Sum	0.004352	0.001434	3.035	0.00241	**
Median Width	0.028632	0.006624	4.322	1.54E-05	***
Woody Debris	0.421153	0.18118	2.324	0.0201	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

From the model results above and probability plots (Appendix B), an increase in discharge decreases the likelihood of an Orangespotted Sunfish occurrence. An increase in the diversity index of substrates, and the specified depth, width, woody debris, and riparian area increase the likelihood of an Orangespotted Sunfish occurrence. The confusion matrix showed this model having a 66% accuracy at predicting presence and absence. Out of 1,675 total samples, the Orangespotted Sunfish were in 734 samples. They are not of conservation concern.

From the literature it is known that the Orangespotted Sunfish is associated with small or large streams, pools, sand, gravel, and mud. They spawn on gravel or sand (Eberle et al., 2014).

3.4.3.10. Northern Plains Killifish

The final model created by repeating logistic regressions until only the significant habitat variables were left for Northern Plains Killifish is summarized in Table 3.15.

Table 3.15 Northern Plains Killifish Model

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.51257	0.889294	-0.576	0.56436	
Total Canopy	-0.04481	0.004473	-10.018	< 0.000000 00000000 02	***
Median Depth	-0.05734	0.008772	-6.536	6.3E-11	***
Embeddedness	0.041104	0.008213	5.005	5.6E-07	***
H' Substrate	-2.07148	0.321802	-6.437	1.22E-10	***
Riparian Width Sum	0.020606	0.002971	6.936	4.03E-12	***
Overhanging	0.673116	0.293372	2.294	0.02177	*
Macrophyte	-0.67675	0.231108	-2.928	0.00341	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

From the model results above and probability plots (Appendix B), an increase in the diversity index of substrates, macrophytes, specified depth, and total canopy decrease the likeliness of a Northern Plains Killifish occurrence. An increase in embeddedness, overhanging vegetation and riparian area increases the likeliness of a Northern Plains Killifish occurrence. The confusion matrix showed this model being 87% accurate at predicting presences and absences. Out of 1,675 total samples, the Northern Plains Killifish were present in 435 samples.

From the literature it is known that the Northern Plains Killifish is associated with shallow pools with sand bottoms, which is explained by the inverse relationship with depth above. They feed on insects and plant material and spawn on gravel (Eberle et al., 2014), which is explained by the diversity index of substrates in the model. Riparian area is commonly associated with an increase in fish presence for most species.

3.4.3.11. Plains Minnow

The final model created by repeating logistic regressions until only the significant habitat variables were left for Plains Minnow is summarized in Table 3.16.

Table 3.16 Plains Minnow Model

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.061707	1.172667	3.464	0.000533	***
Bearing	-0.01182	0.005299	-2.23	0.025729	*
H' Substrate	-1.75791	0.535014	-3.286	0.001017	**
Riparian Width Sum	0.014127	0.005735	2.463	0.013763	*
Total Cover	-1.14032	0.297654	-3.831	0.000128	***
Median Width	0.053307	0.017587	3.031	0.002437	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

From the model results above and probability plots (Appendix B), an increase in the diversity index of substrates, specified total cover and bearing decrease the likeliness of a Plains Minnow occurrence. An increase in width and riparian area increases the likeliness of a Plains Minnow occurrence. The confusion matrix showed this model being 77% accurate at predicting presences and absences. Out of 1,675 samples, the Plains Minnow was present in 60 samples.

From the literature, it is known that the Plains Minnow is associated with shallow pools with sand and primarily feed on algae, diatoms, and are pelagic spawners (Eberle et al., 2014).

3.4.4. Predictions

As the models continue to be developed and new habitat variables tested, the strongest correlating variables will be found. Even though some fish species didn't show strong correlations to certain habitat variables that literature argues they should, it is most likely due to the dis-proportional number of absences compared to presences when sampling for a species of fish or the lack of suitable habitat for specific fish species in Kansas streams. The models for most of the species do not have enough presence to inform a justifiable model for empirical habitat data. The empirical habitat data is primarily collected in small, wadeable streams, that might favor some species compared to others. For example, from literature, it is known that Blue Catfish prefer large, deep bodies of water, so they are not as likely to be found in the streams being surveyed as a fish such as the Central Stoneroller that prefer smaller, shallower streams. Similarly, species such as the Freshwater Drum and the Orangespotted Sunfish prefer gravel substrate, but that isn't available in many Kansas streams. The data can be biased in this way. The models not always matching the literature is still informative, as it shows areas where monitoring data is sparse. This analysis took a state-wide focus, but narrowing down to a specific basin might be more indicative of the results. When narrowing down, there will be an even bigger loss of data though. The current proposed framework itself will also continue to evolve and be modified.

3.5. Conclusion

Monitoring data is essential for conservation, and it is usually collected with justifiable methodology. However, there are challenges to understanding and analyzing monitoring data. One of the main challenges, which was echoed through this data analysis is that one dataset cannot answer all questions. There is also a struggle to identify significance with such a broad set of data and so few data points. This dataset focuses on in-stream habitat data, with awareness that fish presence and absence rely on much more than that. While this framework does work to identify the presence of fish species, it will continue to be improved with each model created and evaluated.

Chapter 4 - General Conclusions

Establishing a consistent and repeatable way to analyze state-wide monitoring data is important for conservation. The goal of these data analyses for both the blue-green algae bloom analysis and the fish habitat analysis was not to come to a single significant conclusion, but more so an approach to process and review variables. While it is known that monitoring data is essential for maintaining the ecosystems of Kansas lakes, reservoirs, and streams, there are some challenges to understanding the data. Monitoring data is usually collected with justifiable methodology, but sometimes there is a struggle to identify significance with such a broad scope of data being collected. One of the main challenges, which was echoed through this data analysis is that one dataset cannot answer all questions. These datasets focused on a few variables, with the awareness that cyanobacteria blooms and fish distribution can be influenced by many other variables not assessed. The models presented here will continue to be improved with each development. With the development of such models, a systematic framework is established for a comprehensive, responsible, and a useful approach to further understand and analyze monitoring data.

References

Bisson, P., Montgomery, D., & Buffington, J. (2017). *Valley segments, stream reaches, and channel units*. Methods in Stream Ecology, Volume 1 (Third Edition). Retrieved from https://www.sciencedirect.com/science/article/pii/B9780124165588000020

Bouma-Gregson, K., & Hsu, K. (2013). *Cyanobacteria Factsheet*. Eel River Recovery. Retrieved from https://www.krisweb.com/ERRP/cyanobacteria%20factsheet_FINAL.pdf

Carlson, R. (1977). *A trophic state index for lakes.* Retrieved from https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.4319/lo.1977.22.2.0361

Carney, E. (2012). *The HAB Experience in Kansas*. Retrieved from https://www.nalms.org/product/lakeline-32-3-habs/

Cooper A.R., Infante D.M., Daniel W.M., Wehrly K.E., Wang L., Brenden T.O. (2017). *Assessment of dam effects on streams and fish assemblages of the Conterminous USA*. The Science of the total environment. Retrieved from https://pubmed.ncbi.nlm.nih.gov/28233615/

Dzialowski, A.R., S.H. Wang, N.C. Lim, W.W. Spotts & D.G. Huggins. (2005). *Nutrient Limitation of Phytoplankton Growth in Central Plains Reservoirs, USA.* (pp. 587-595). Journal of Plankton Research.

Dodds, W., Carney, E., & Angelo, R. (2006). *Determining ecoregional reference conditions for nutrients, secchi* ...Retrieved from https://www.researchgate.net/publication/240310874_Determining_Ecoregional_Reference_Conditions_for_Nutrients_Secchi_Depth_and_Chlorophyll_a_in_Kansas_Lakes_and_Reservoirs

Eberle, M., Edds, D., & Tomelleri, J. (2014). Kansas Fishes.

Environmental Protection Agency (EPA). (2022). *Learn about Cyanobacteria and Cyanotoxins*. EPA. Retrieved from https://www.epa.gov/cyanohabs/learn-about-cyanobacteria-and-cyanotoxins

Garcia-Pichel, F. (2009). Cyanobacteria. *Encyclopedia of Microbiology* (pp. 107-124). Elsevier Inc.. https://doi.org/10.1016/B978-012373944-5.00250-9

Gerken, J. & Paukert, C. (2013). Fish assemblage and habitat factors associated with the distribution of Topeka shiner (*Notropis topeka*) in Kansas streams. *Journal of Freshwater Ecology*, 28:4,503-516, DOI: 10.1080/02705060.2013.792754

JMP®, Version 16. SAS Institute Inc., Cary, NC, 1989–2021.

Kansas Department of Wildlife and Parks (KDWP). (2022). *Kansas State Wildlife Action Plan.* Retrieved from https://ksoutdoors.com/Services/Kansas-SWAP

Kansas Department of Wildlife and Parks (KDWP). (n.d.). *Regions Map.* Retrieved from https://ksoutdoors.com/KDWP-Info/Locations/Hunting-Fishing-Atlas/Regional/Regions-Map

Kansas Department of Wildlife and Parks. (n.d.). *Stream Assessment and Monitoring Program.* Retrieved from https://ksoutdoors.com/Services/Stream-Assessment-and-Monitoring-Program

Kasco. (2023). *Proper Depth for Fish Health.* Retrieved from https://kascomarine.com/blog/using-maintaining-proper-depth-fish-health/

Kansas Department of Health and Environment (KDHE). (2022). *KDHES Habs Response Program*. Retrieved from https://www.kdhe.ks.gov/DocumentCenter/View/6266/2022-HAB-Response-Plan-Complete-PDF?bidId=

Kansas Department of Health and Environment. (n.d.). *Kansas-Lower Republican Basin Total Maximum Daily Load*. Retrieved from https://www.kdhe.ks.gov/DocumentCenter/View/14056/Milford-Lake-EU-PDF

Kansas Department of Health and Environment. (2022). *Lake & Wetland Monitoring Program*. Retrieved from https://www.kdhe.ks.gov/1287/Lake-Wetland-Monitoring-Program

Kansas Department of Health and Environment Bureau of Water. (2011). *Water Quality Standards White Paper.* Retrieved from https://www.kdhe.ks.gov/DocumentCenter/View/9396/Chlorophyll---A-Criteria-for-Public-Water-Supply-Lakes-or-Reservoirs-PDF

Likens, G. E. (2009). *Encyclopedia of Inland Waters*. Elsevier.

Love, S., Phelps, Q., Tripp, S., & Herzog, D. (2016). The importance of shallow-low velocity habitats to juvenile fish in the Middle Mississippi River. Retrieved from https://www.researchgate.net/publication/307933911_The_Importance_of_Shallow-Low_Velocity_Habitats_to_Juvenile_Fish_in_the_Middle_Mississippi_River_The_Importance_of_Shallow-Low_Velocity_Habitats_to_Juvenile_Fish

Lynch, A. (2019). *Substrate.* Fishionary. Retrieved from https://fishionary.fisheries.org/substrate/

Morris, K. E., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., Meiners, T., Müller, C., Obermaier, E., Prati, D., Socher, S. A., Sonnemann, I., Wäschke, N., Wubet, T., Wurst, S., & Rillig, M. C. (2014). Choosing and using diversity indices: Insights for ecological applications from the German biodiversity exploratories. *Ecology and Evolution*, *4*(18), 3514–3524. https://doi.org/10.1002/ece3.1155

Net Map Tools. (2014). *Channel Sinuosity.* Retrieved from https://www.netmaptools.org/Pages/NetMapHelp/channel_sinuosity.htm

National Oceanic and Atmospheric Administration (NOAA). (2019). *What is eutrophication?* Retrieved from https://oceanservice.noaa.gov/facts/eutrophication.html

Percival, S. L., Yates, M. V., Williams, D., Chalmers, R. M., & Gray, N. I. F. (2014). *Microbiology of waterborne diseases: Microbiological aspects and risks*. Elsevier.

Poulton, B, & Tao, J. (2019). *Evaluation of EPT macroinvertebrate metrics in small streams located within the non-connected stormwater management region of Kansas City, Missouri, USA.* USGS. Retrieved from https://pubs.er.usgs.gov/publication/70212585

R Studio, Version 0-386. Posit Software, Boston, MA, 2023.

Savic R, Stajic M, Blagojević B, Bezdan A, Vranesevic M, Nikolić Jokanović V, Baumgertel A, Bubalo Kovačić M, Horvatinec J, & Ondrasek G. (2022). *Nitrogen and Phosphorus Concentrations and Their Ratios as Indicators of Water Quality and Eutrophication of the Hydro-System Danube–Tisza–Danube.* Agriculture. Retrieved from https://doi.org/10.3390/agriculture12070935

Shannon, C. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27:379–423, 623–656.

Statistics Solutions. (2023). What is Logistic Regression. Retrieved from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/

United States Geological Survey (USGS). (2016). *Cyanobacterial (blue-green algal) blooms: Tastes, odors, and ...* Retrieved from https://www.usgs.gov/centers/kansas-water-science-center/science/cyanobacterial-blue-green-algal-blooms-tastes-odors-and

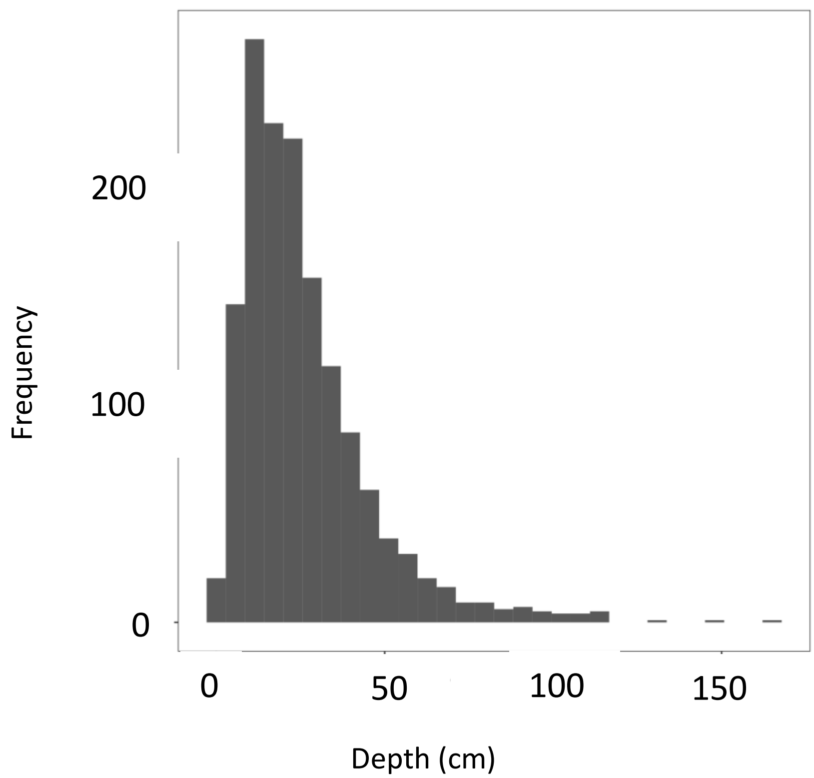
Wall, S., Berry, C., Blausey, C., Jenks, J., & Kopplin, C. (2004). *Fish-habitat modeling for gap analysis to conserve the endangered Topeka Shiner (Notropis Topeka)*. Retreived from https://www.usgs.gov/publications/fish-habitat-modeling-gap-analysis-conserve-endangered-topeka-shiner-notropis-topeka

Washington State Department of Health. (n.d.). *Blue-green algae*. Retrieved from https://doh.wa.gov/community-and-environment/contaminants/blue-green-algae

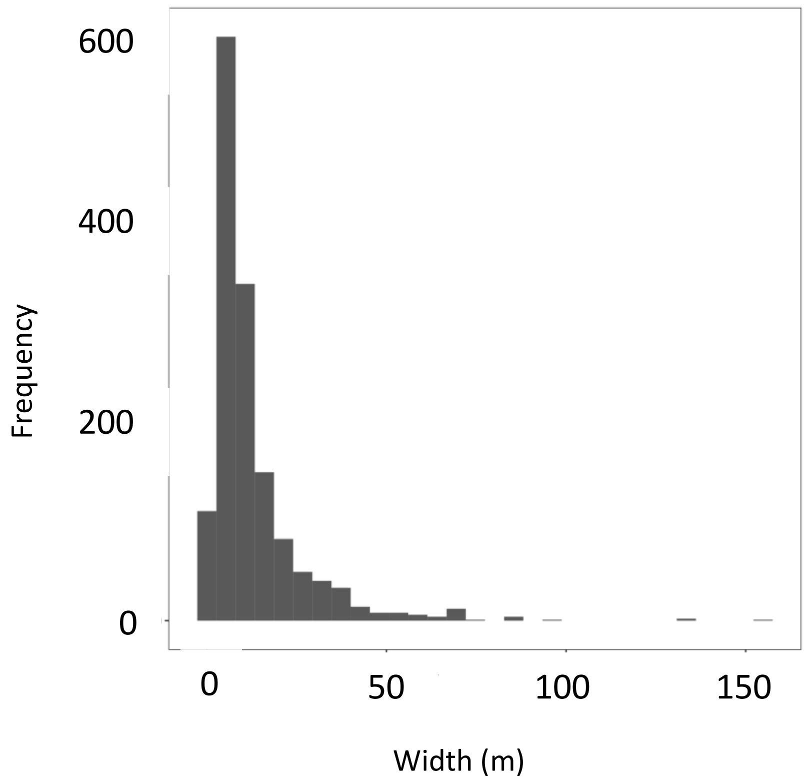
Appendix A - Histograms and Boxplots for Kansas Fish Dataset

- These histograms and boxplots are for the entire habitat dataset, and not for individual fish species.

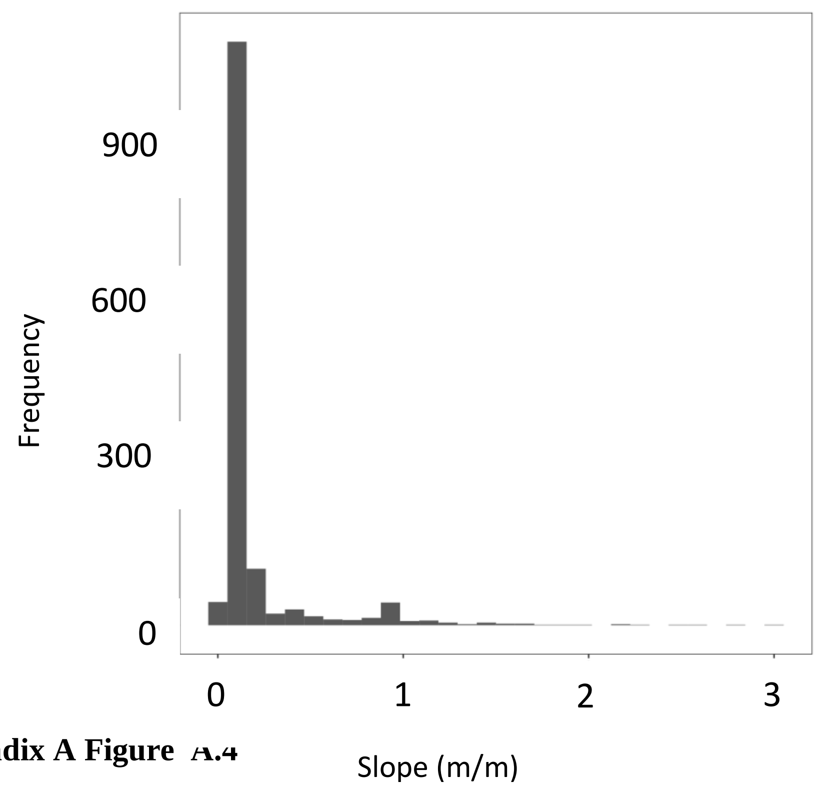
Appendix A Figure A.1



Appendix A Figure A.2

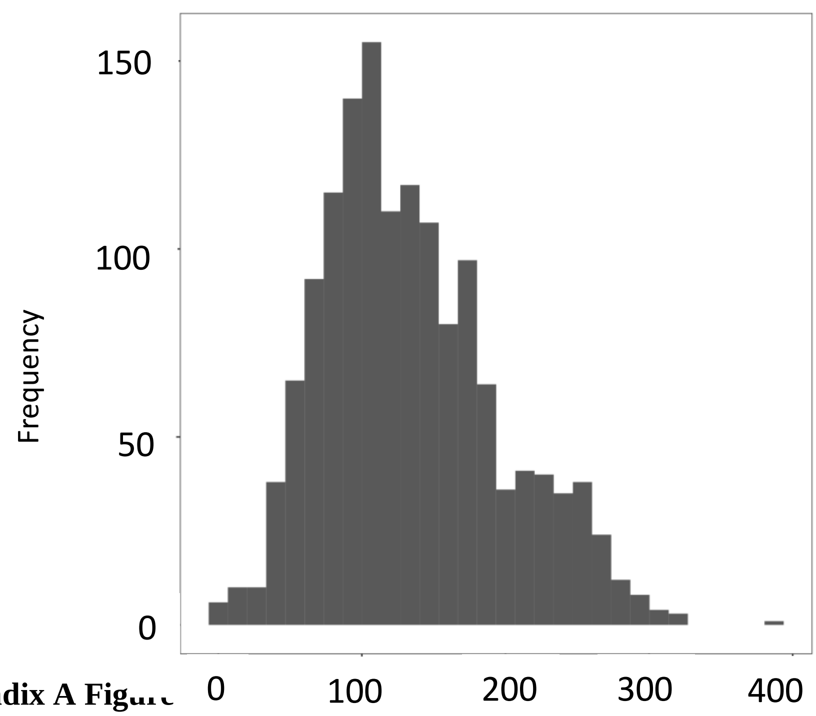


Appendix A Figure A.3

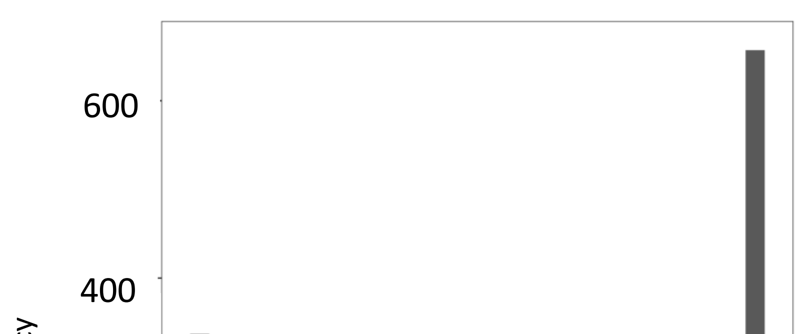


Appendix A Figure A.4

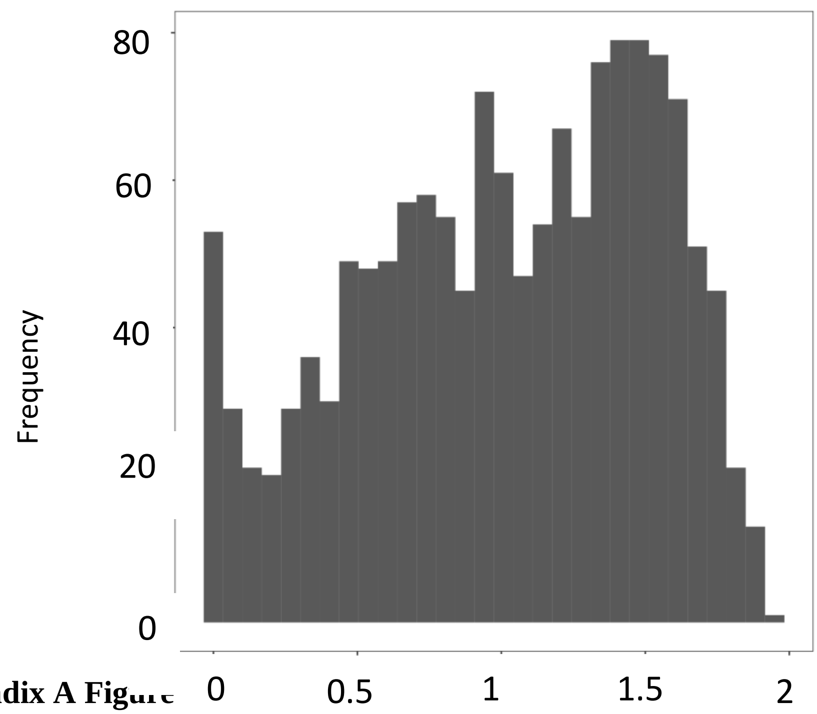
Appendix A Figure A.5



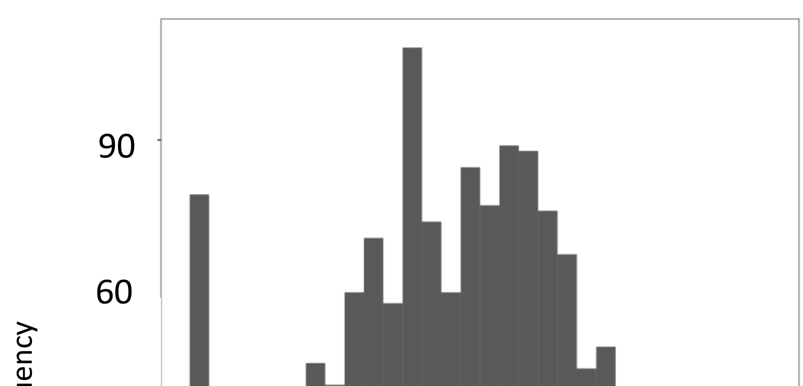
Appendix A Figure



Appendix A Figure A.7

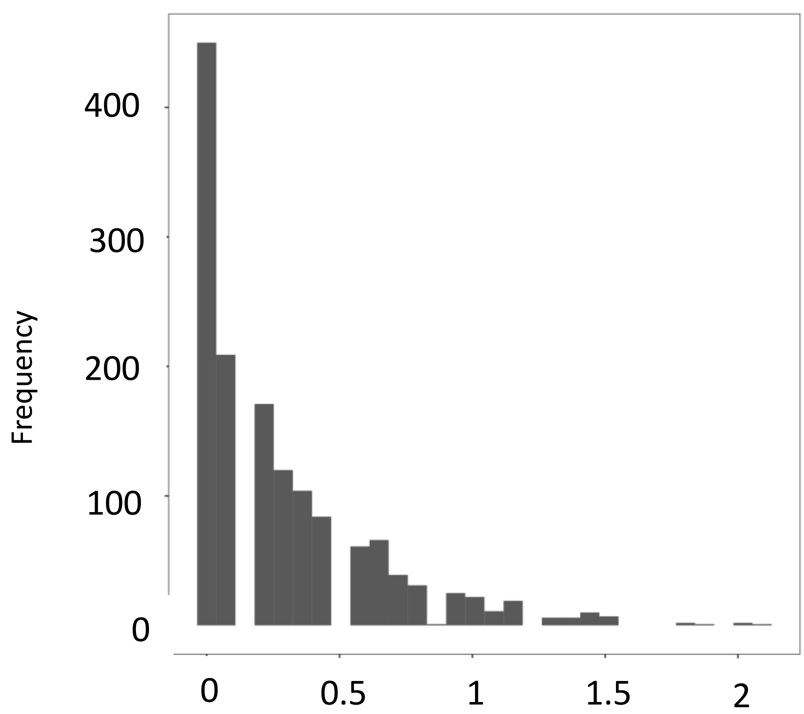


Appendix A Figure

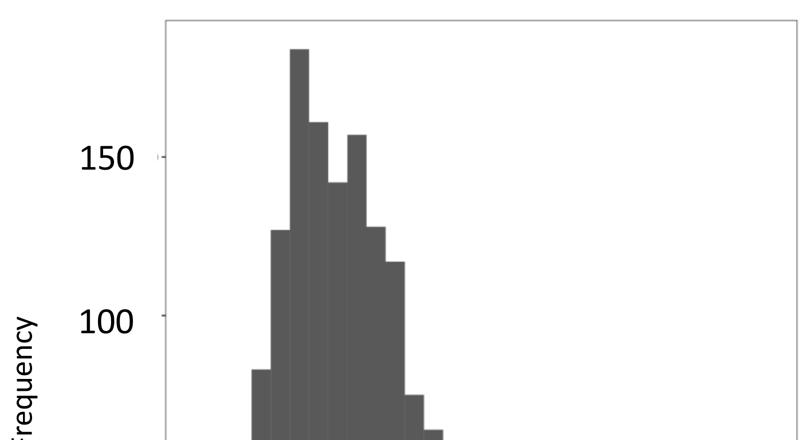


Channel Unit Diversity Index

Appendix A Figure A.9

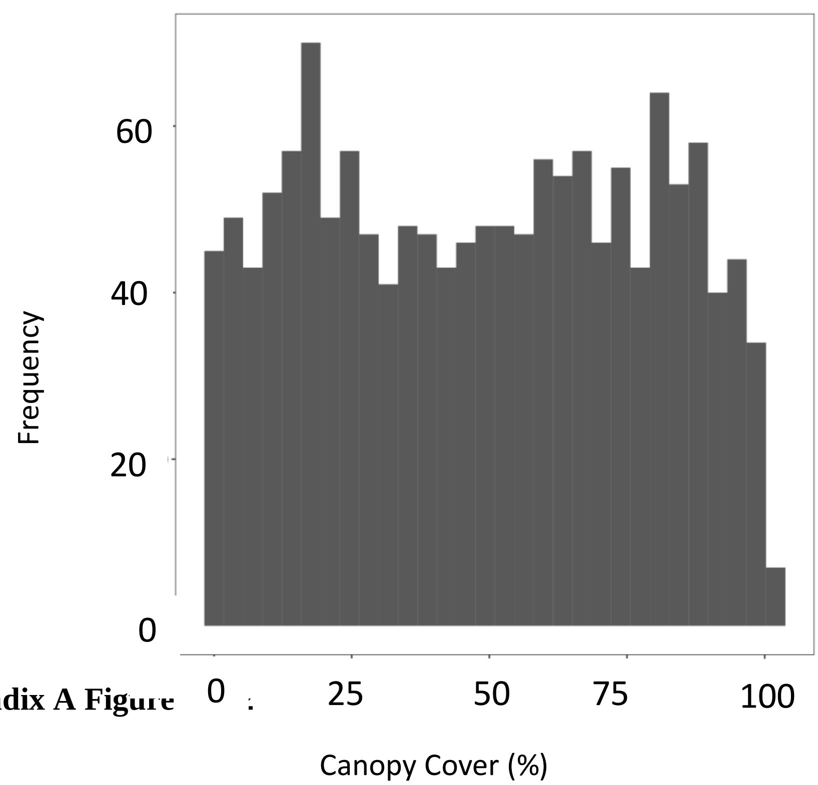


Appendix A Figure A.10

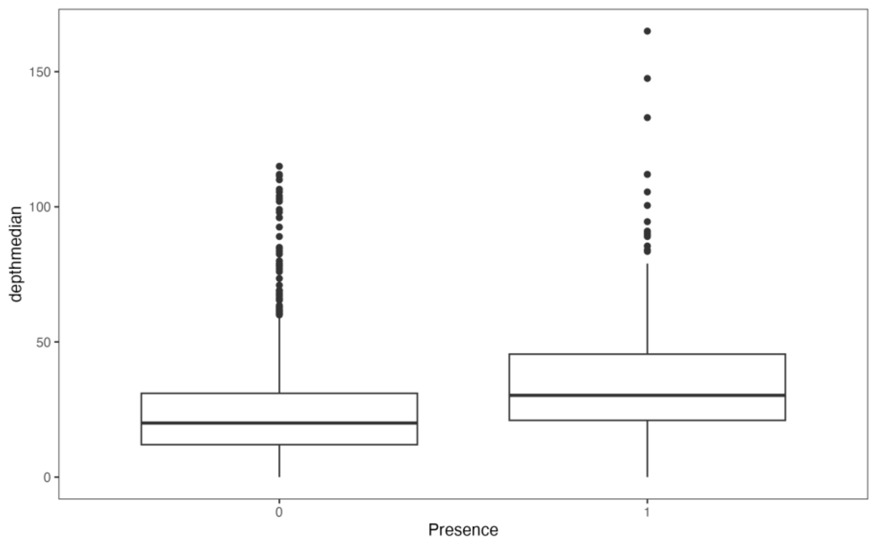


Total Instream Cover (count)

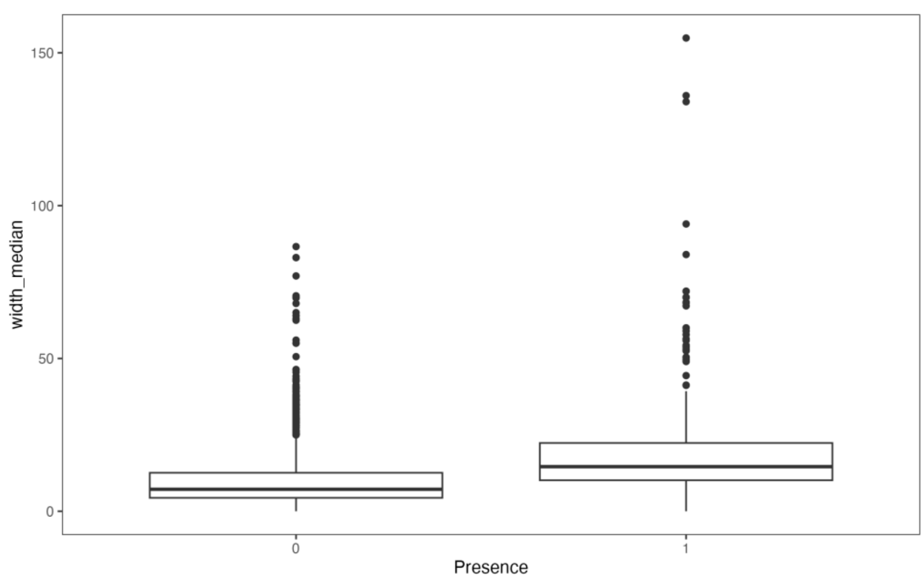
Appendix A Figure A.11



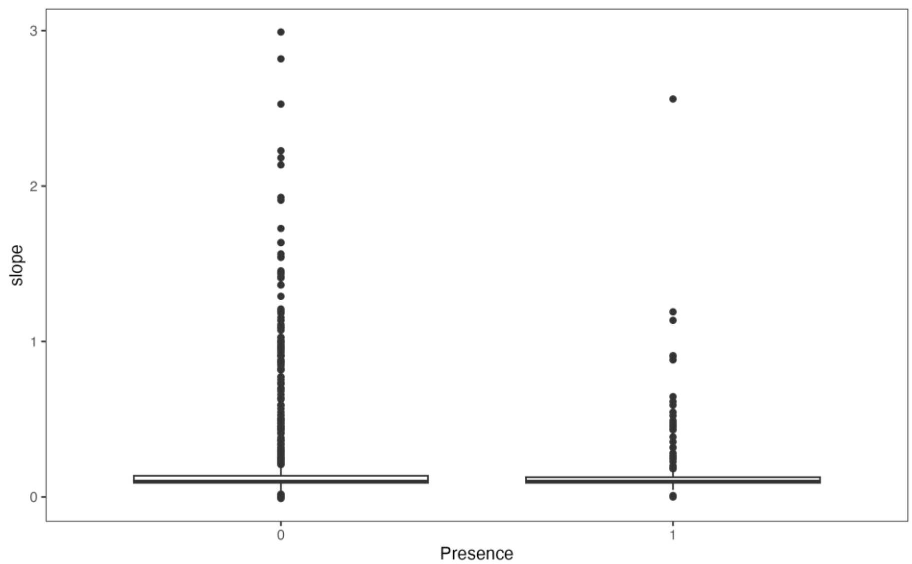
Appendix A Figure



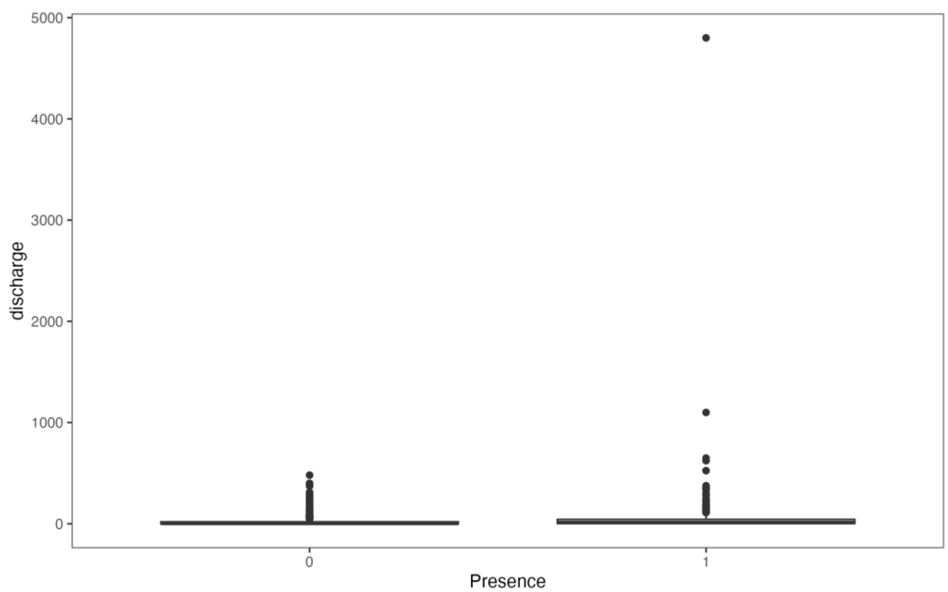
Appendix A Figure A.13



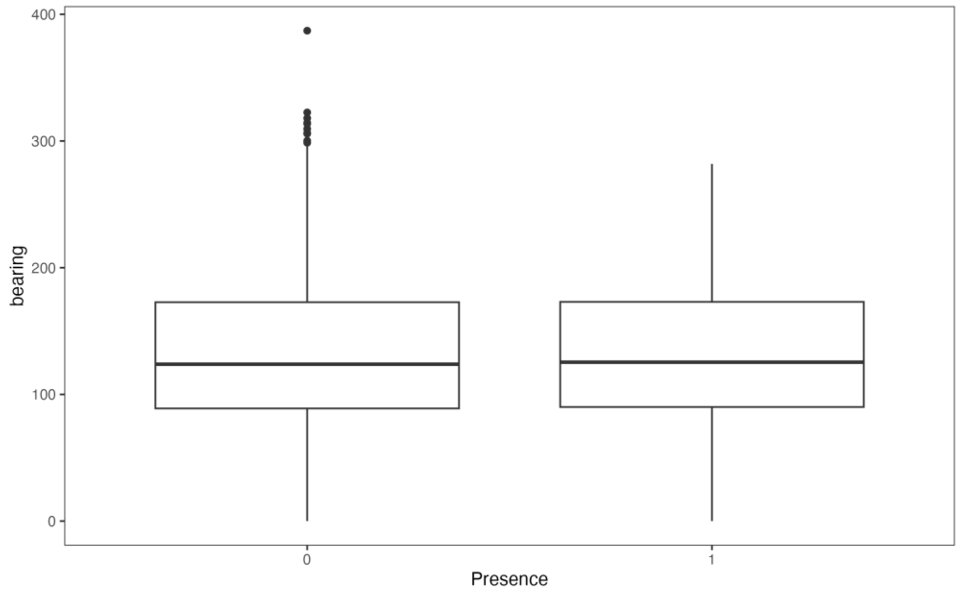
Appendix A Figure A.14



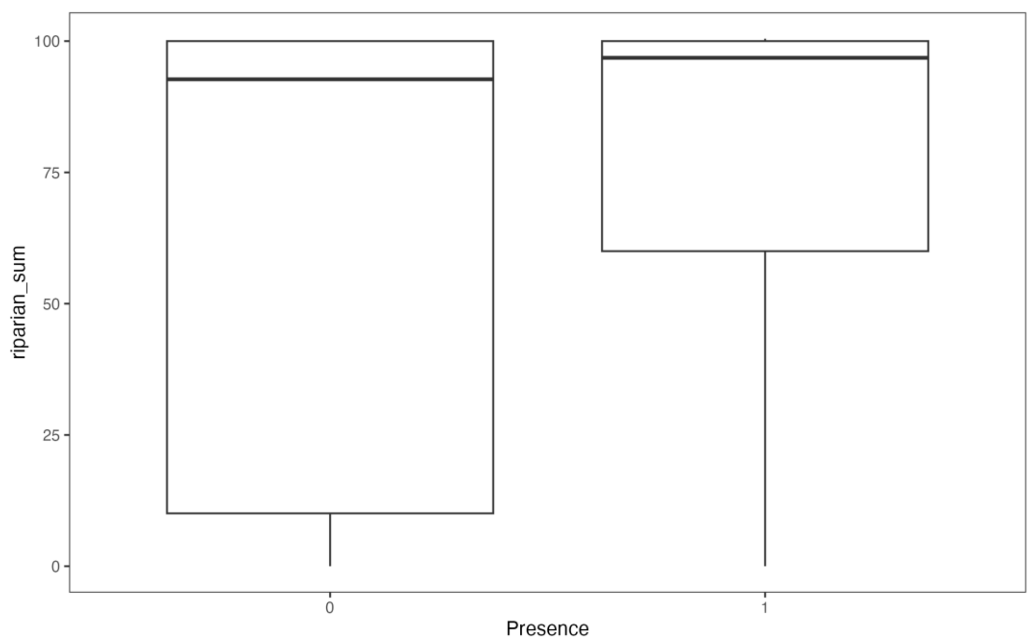
Appendix A Figure A. 15



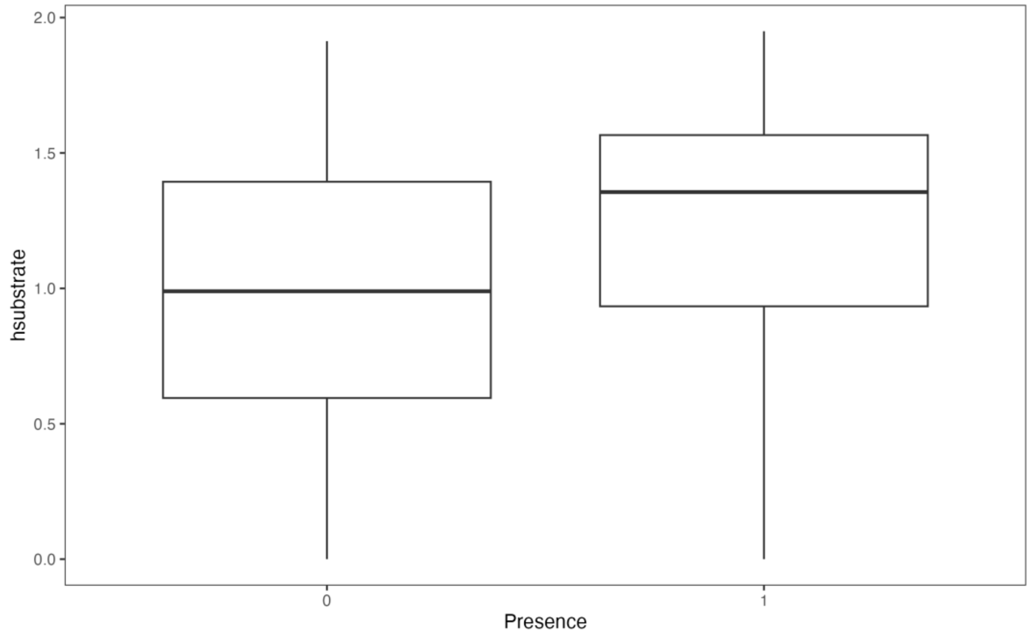
Appendix A Figure A.16



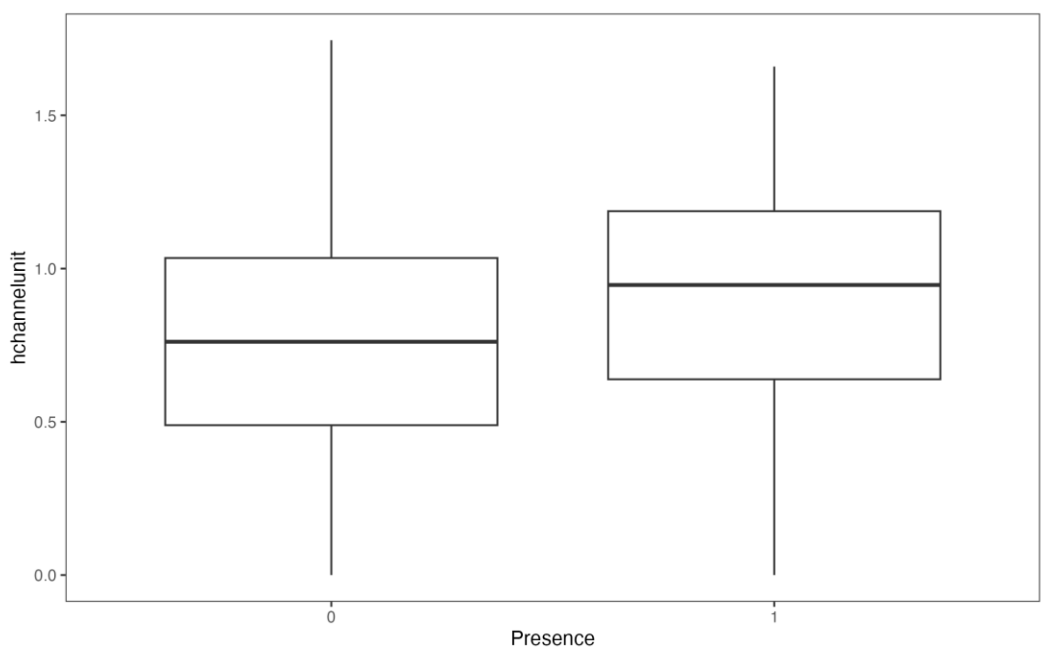
Appendix A Figure A.17



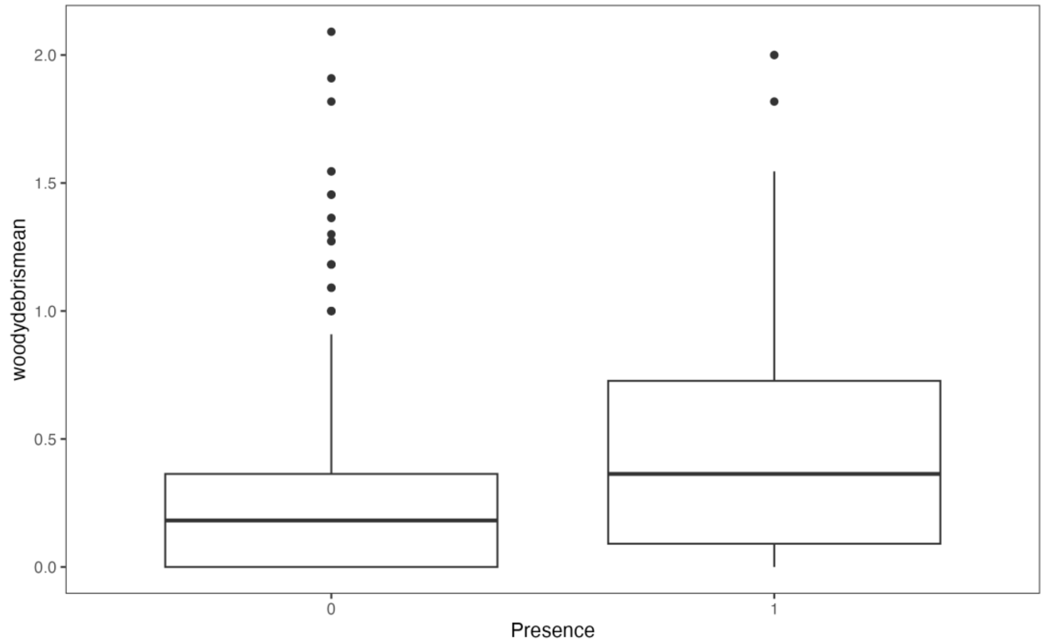
Appendix A Figure A.18



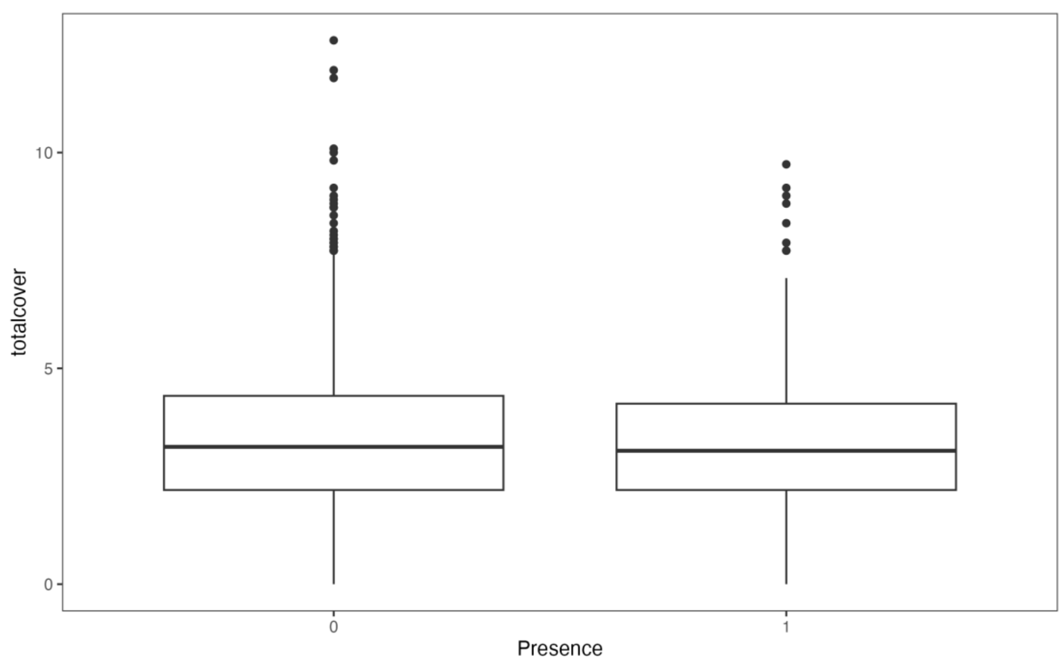
Appendix A Figure A.19



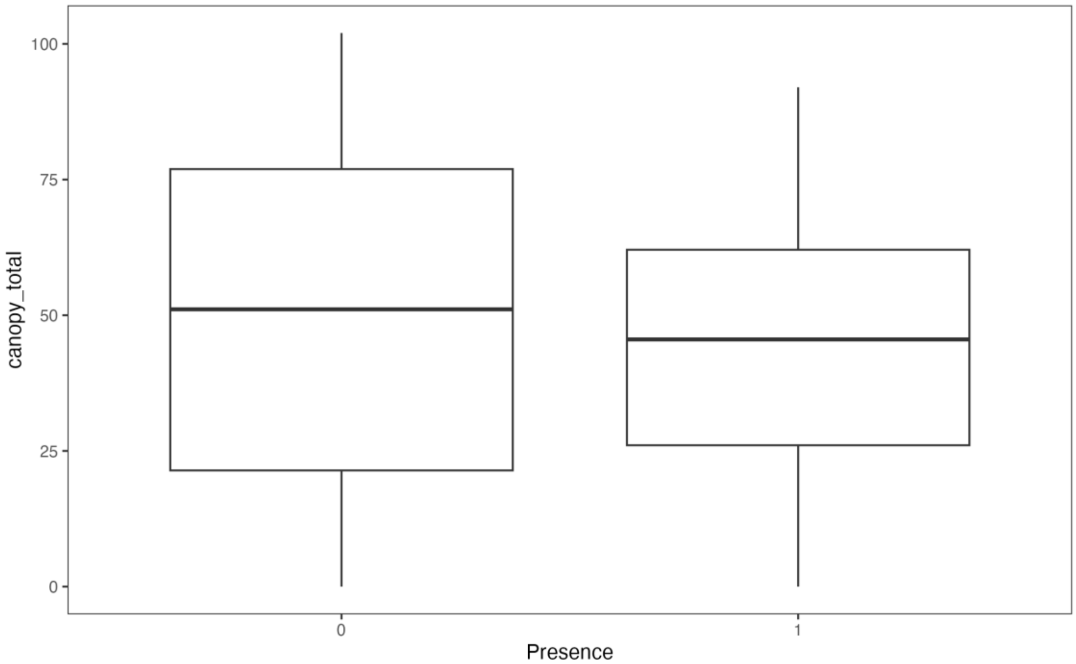
Appendix A Figure A.20



Appendix A Figure A.21

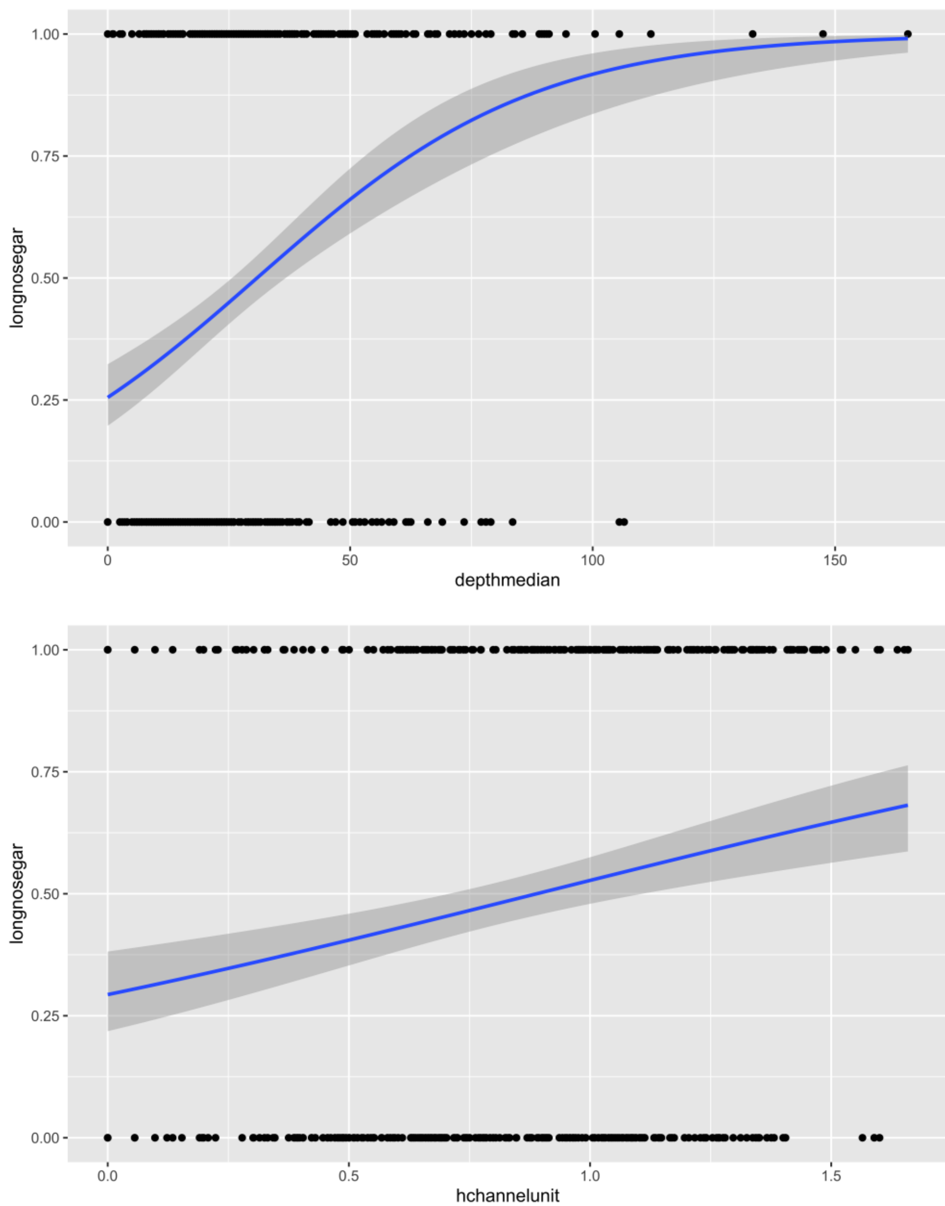


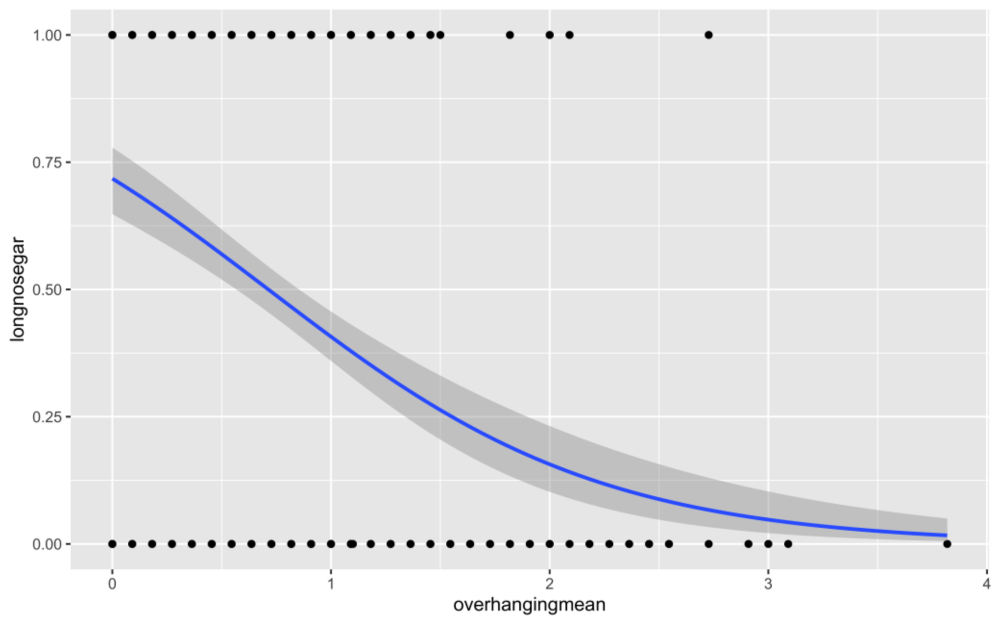
Appendix A Figure A.22


100
75
50
25
0
canopy_total
0
1
Presence

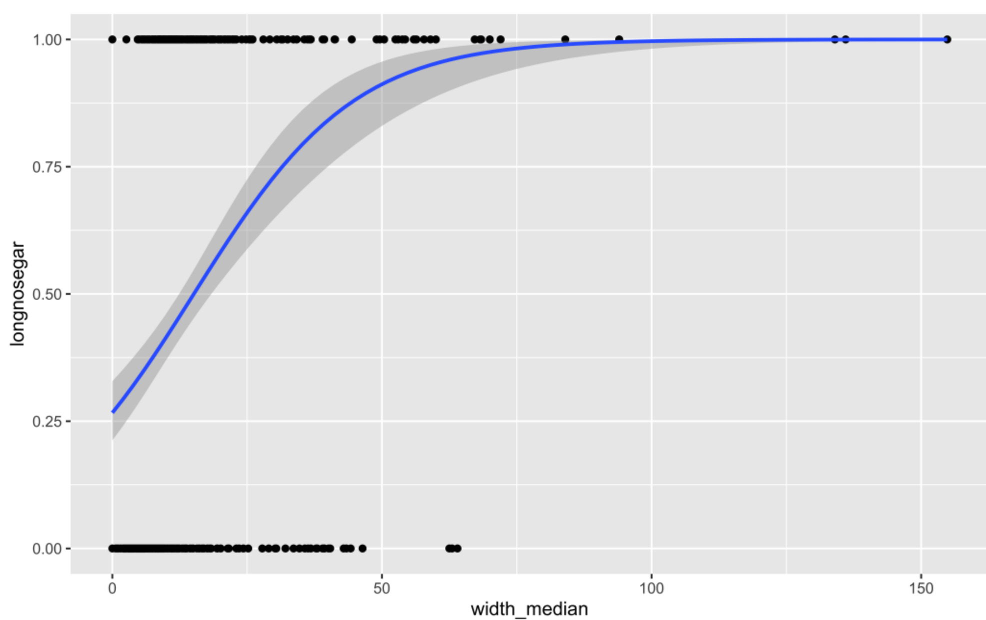
Appendix B - Probability Curves for Reduced Models

Appendix B Figure B.1. Longnose Gar Probability Curves

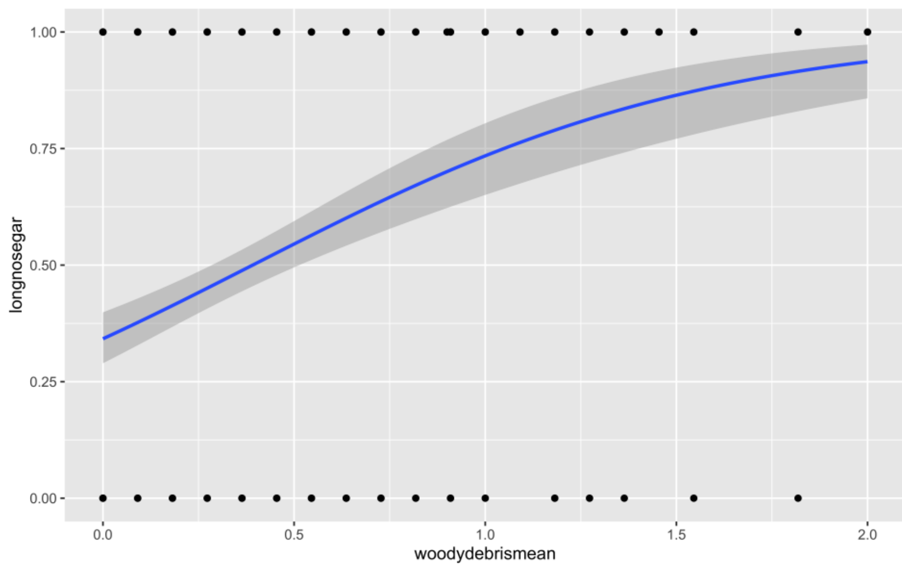




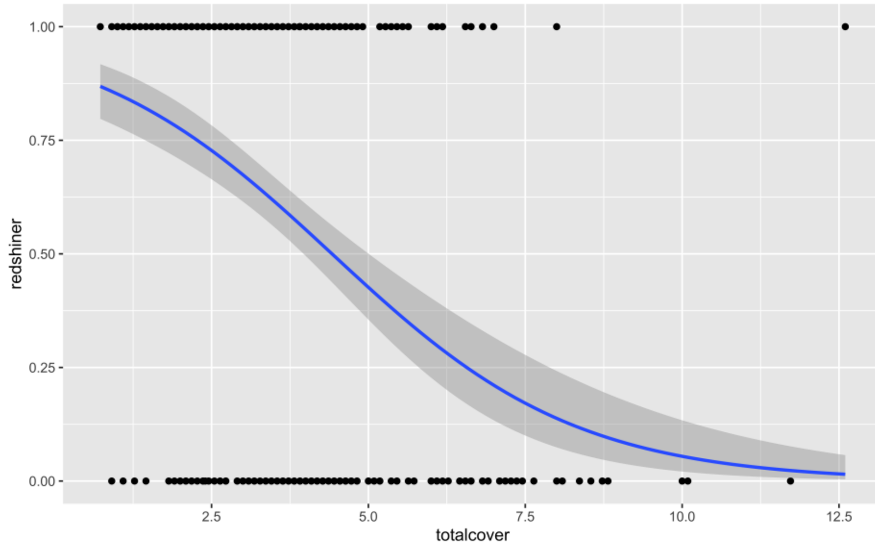
1.00
0.75
0.50
0.25
0.00
longnosegar
0
1
2
3
4
overhangingmean



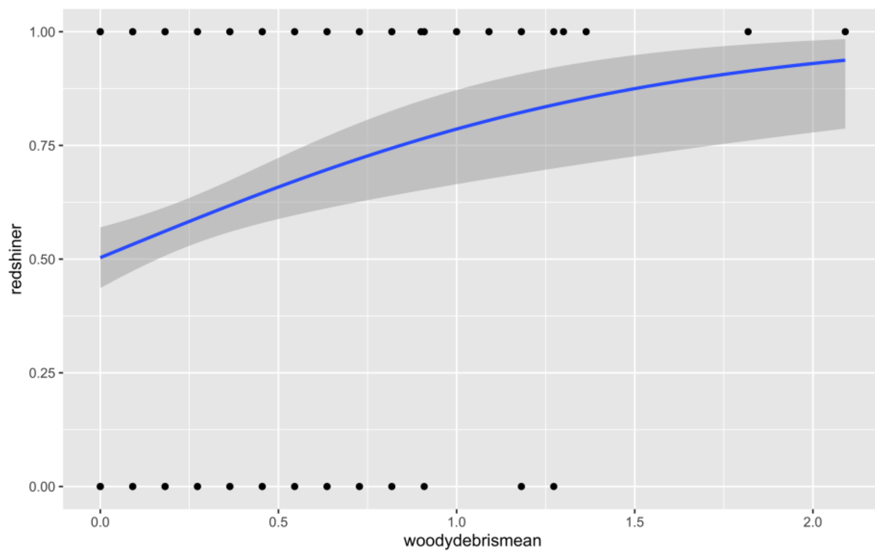
1.00
0.75
0.50
0.25
0.00
longnosegar
0
50
100
150
width_median



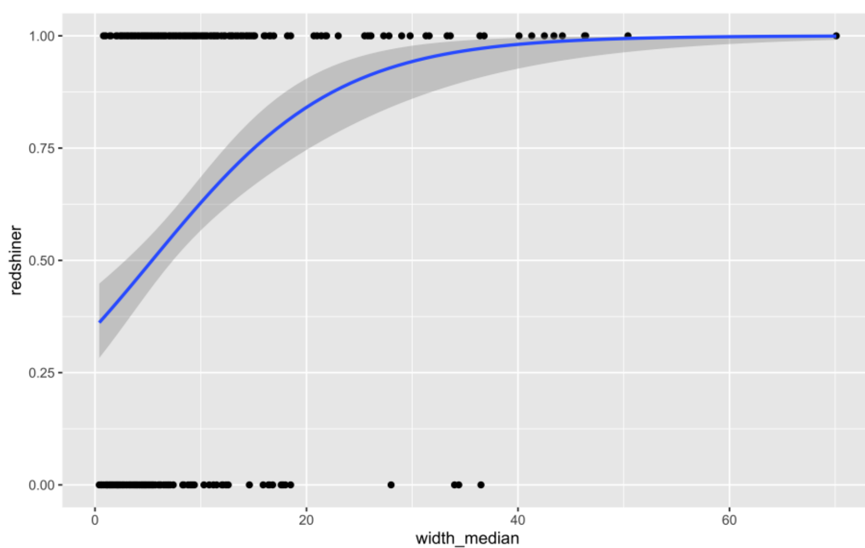
Appendix B Figure B.2. Red Shiner Probability Curves



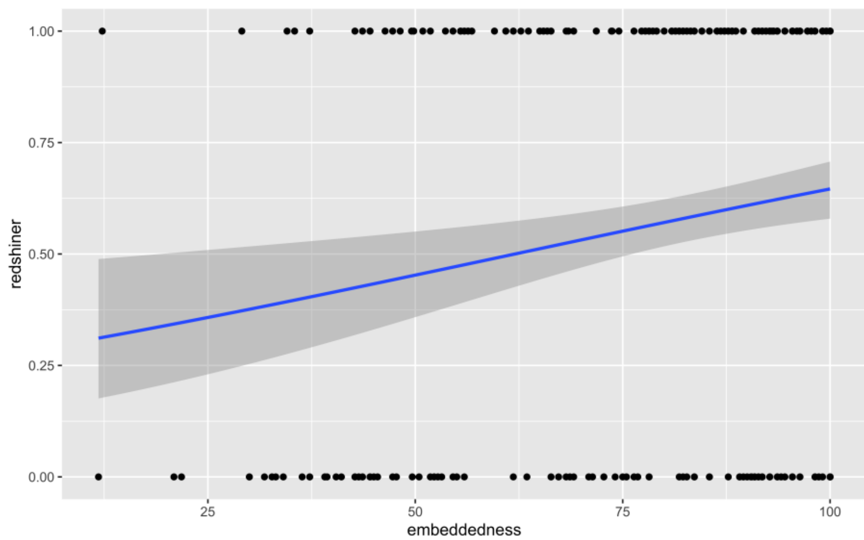
1.00
0.75
0.50
0.25
0.00
redshiner
2.5
5.0
7.5
10.0
12.5
totalcover



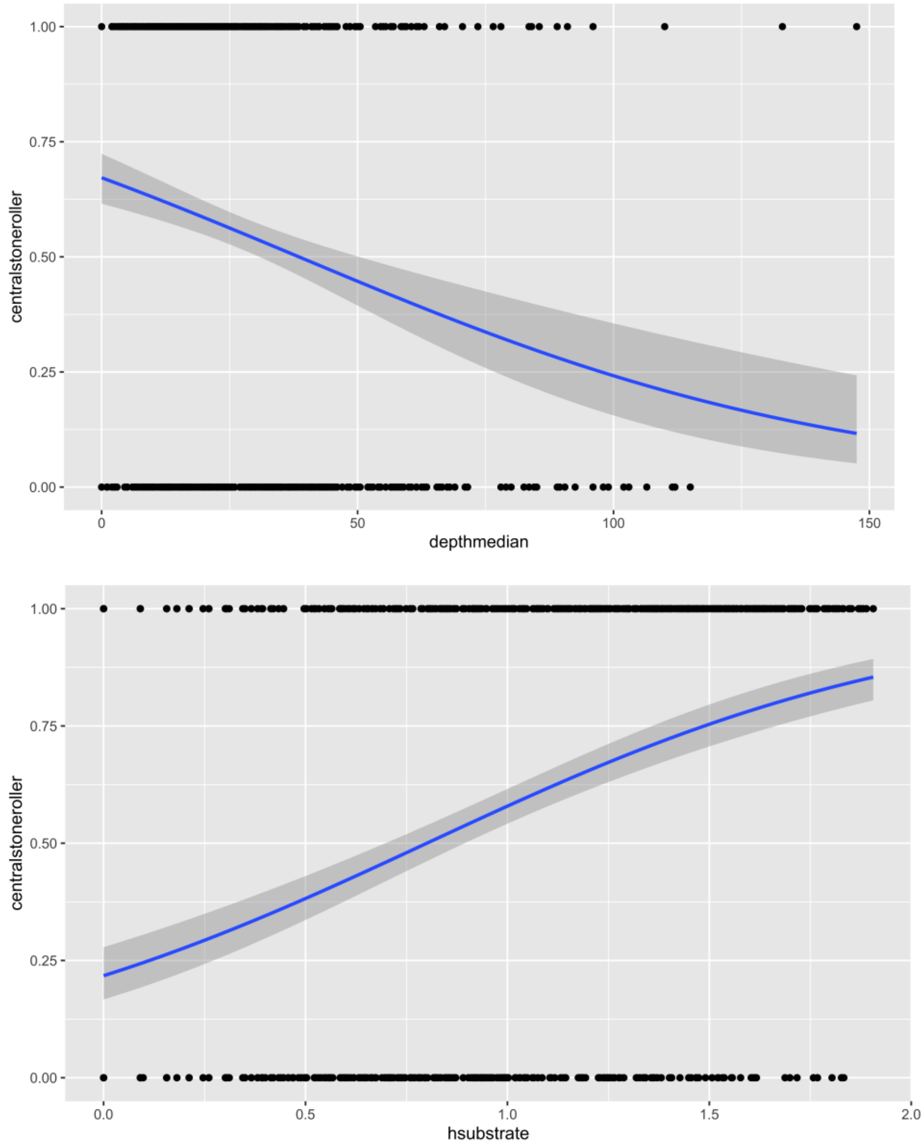
1.00
0.75
0.50
0.25
0.00
redshiner
0.0
0.5
1.0
1.5
2.0
woodydebrismean



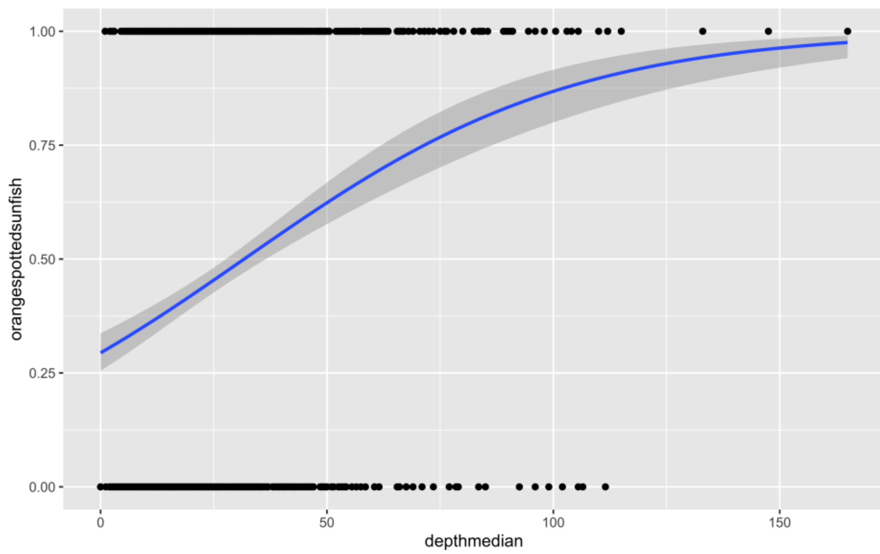
1.00
0.75
0.50
0.25
0.00
redshiner
0
20
40
60
width_median



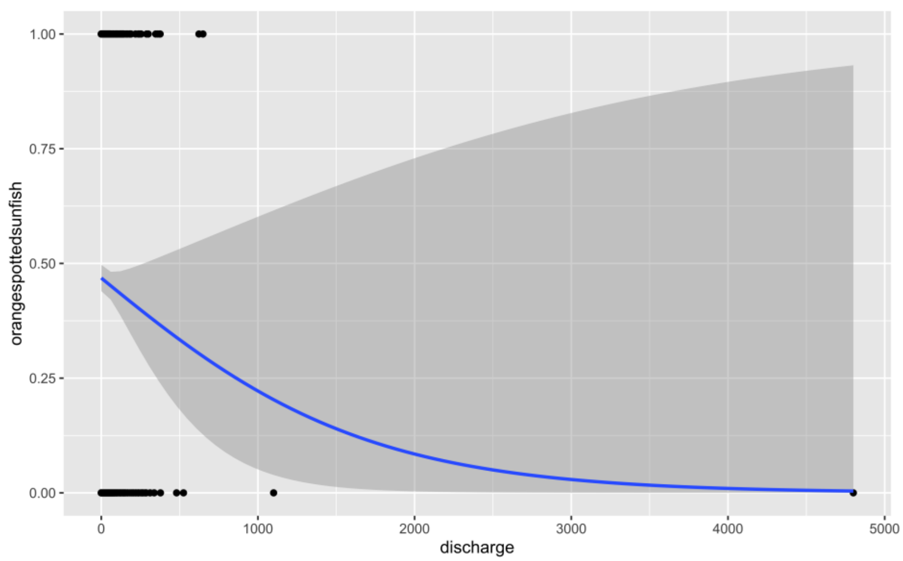
Appendix B Figure B.3. Central Stoneroller Probability Curves



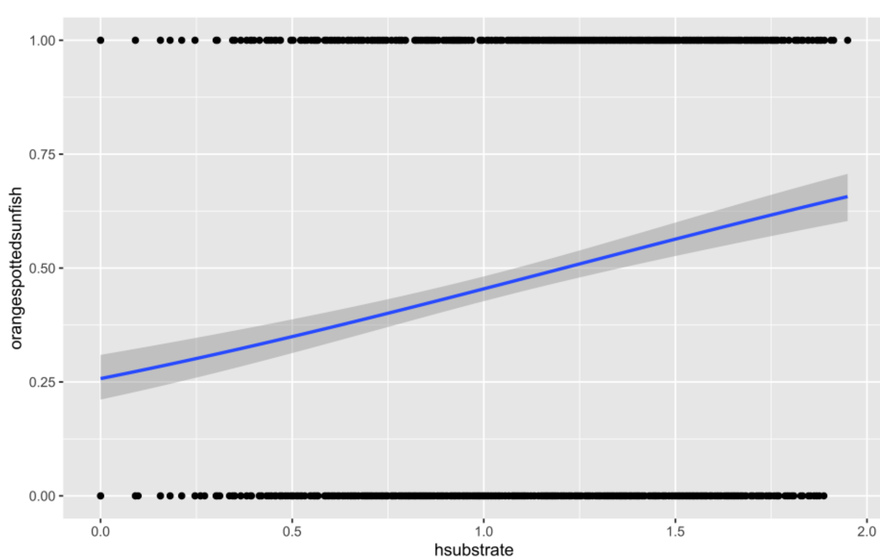
Appendix B Figure B.4. Orangespotted Sunfish Probability Curves



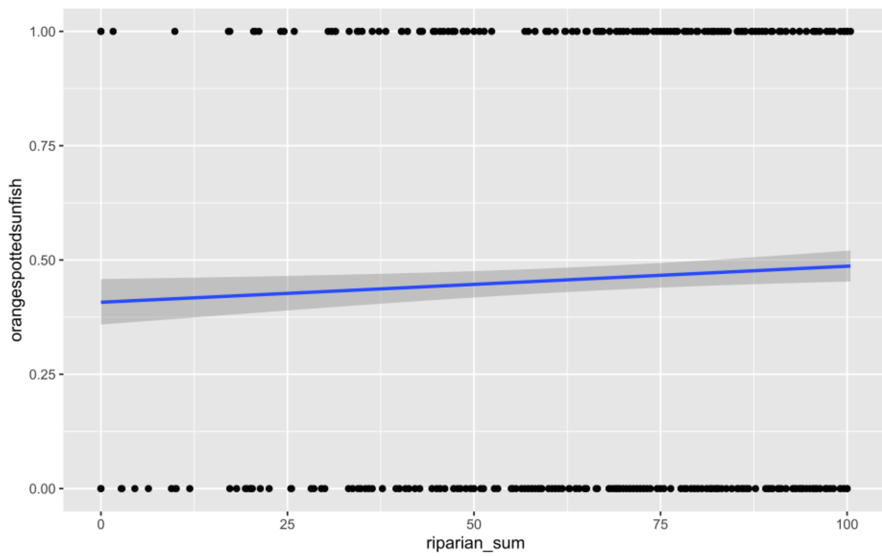
orangespottedsunfish
1.00
0.75
0.50
0.25
0.00
0
50
100
150
depthmedian



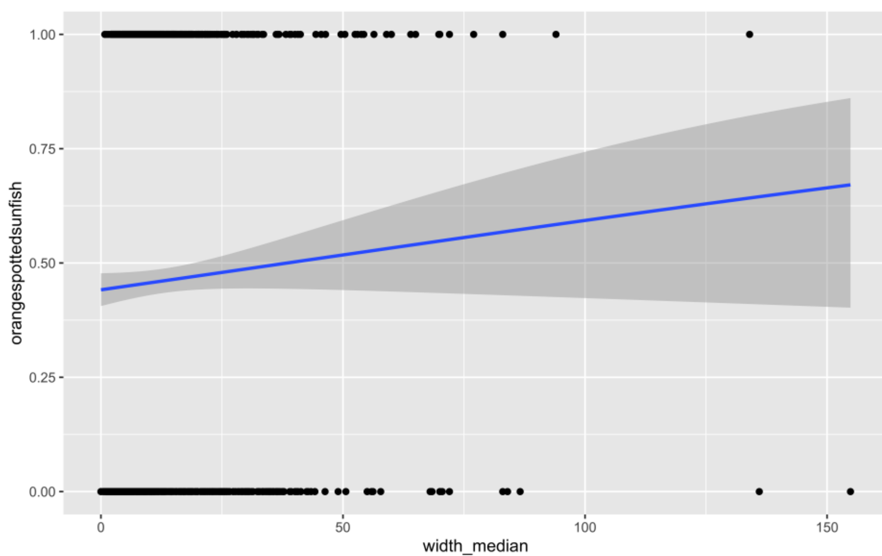
orangespottedsunfish
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
discharge



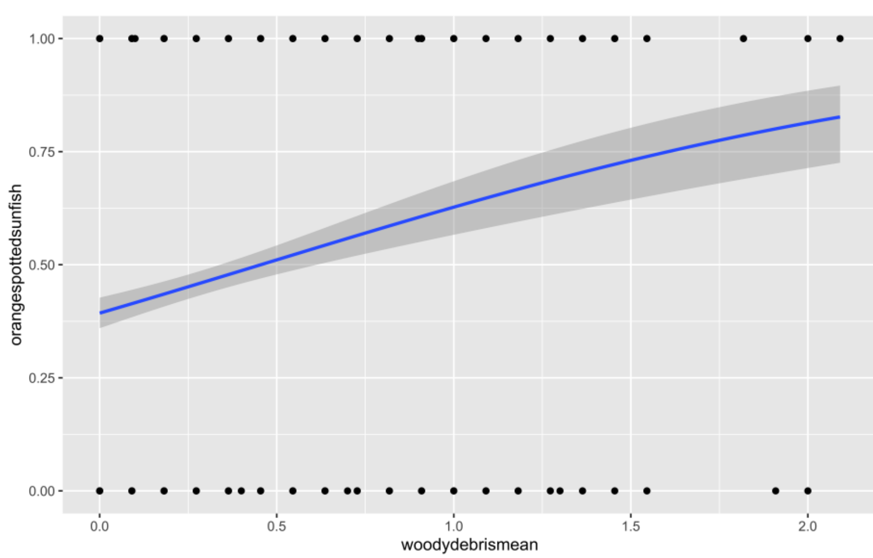
orangespottedsunfish
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
hsubstrate



1.00
0.75
0.50
0.25
0.00
orangespottedsunfish
0
25
50
75
100
riparian_sum

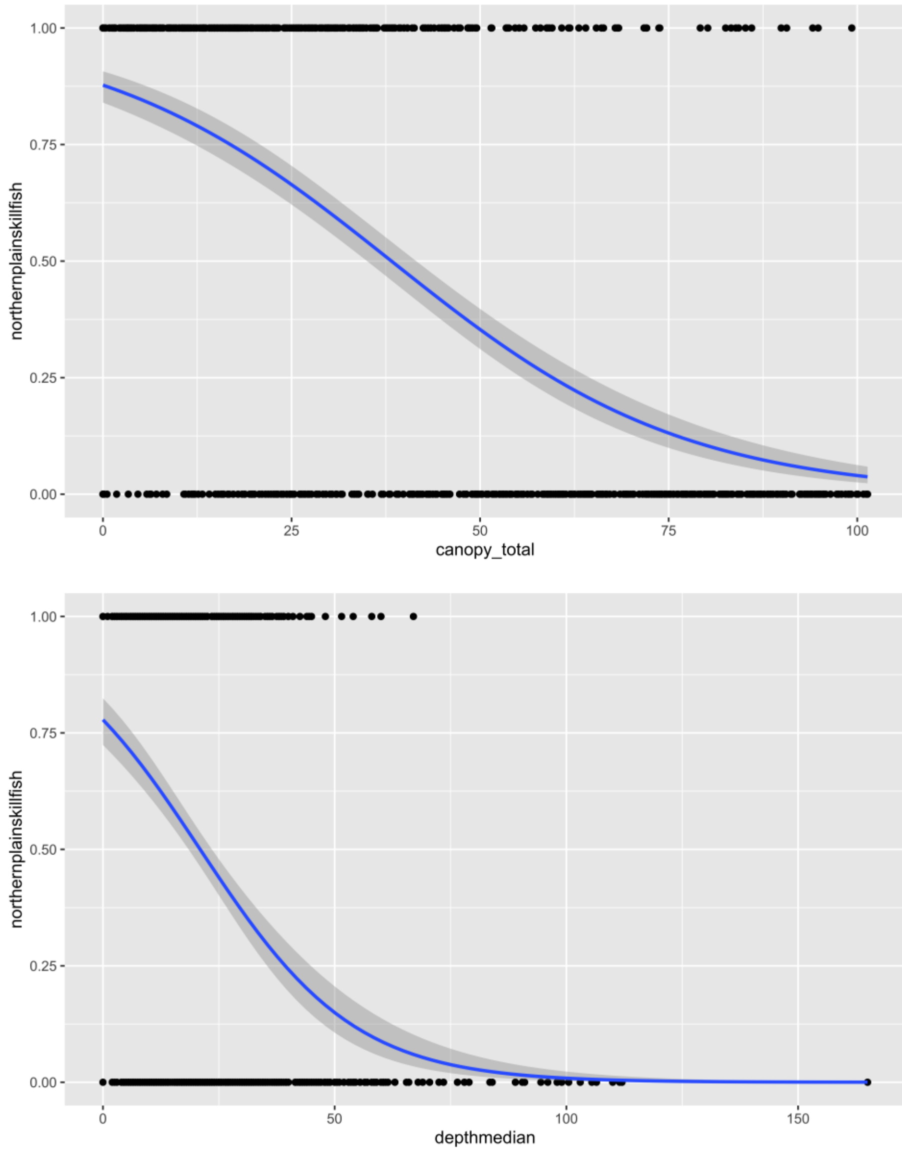


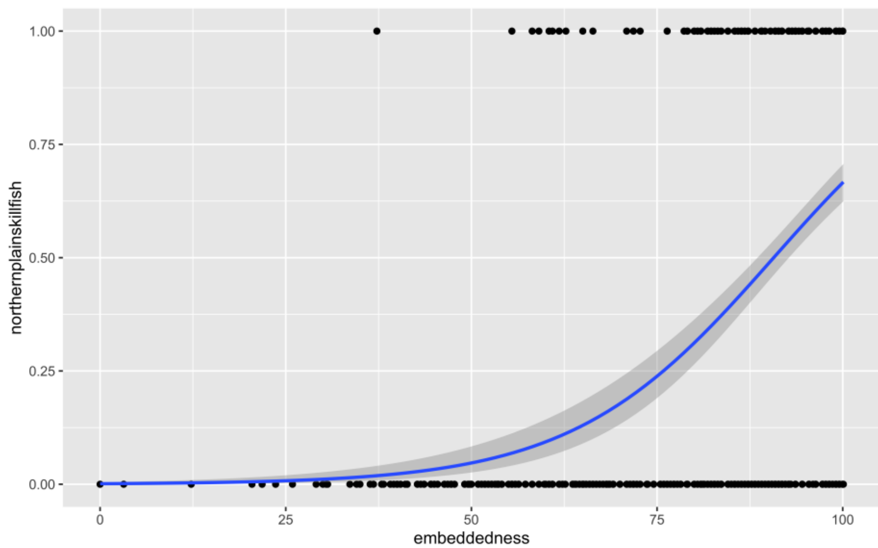
1.00
0.75
0.50
0.25
0.00
orangespottedsunfish
0
50
100
150
width_median



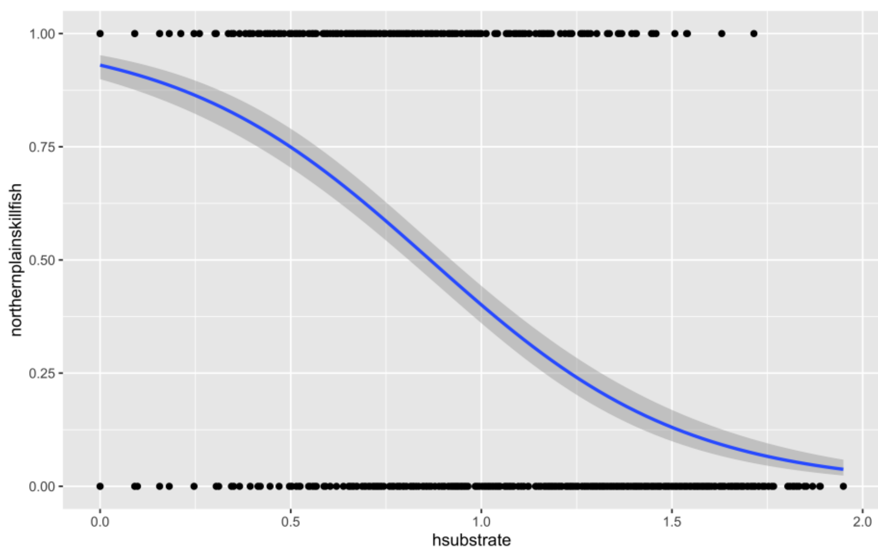
1.00
0.75
0.50
0.25
0.00
orangespottedsunfish
0.0
0.5
1.0
1.5
2.0
woodydebrismean

Appendix B Figure B.5. Northern Plains Killifish Probability Curves

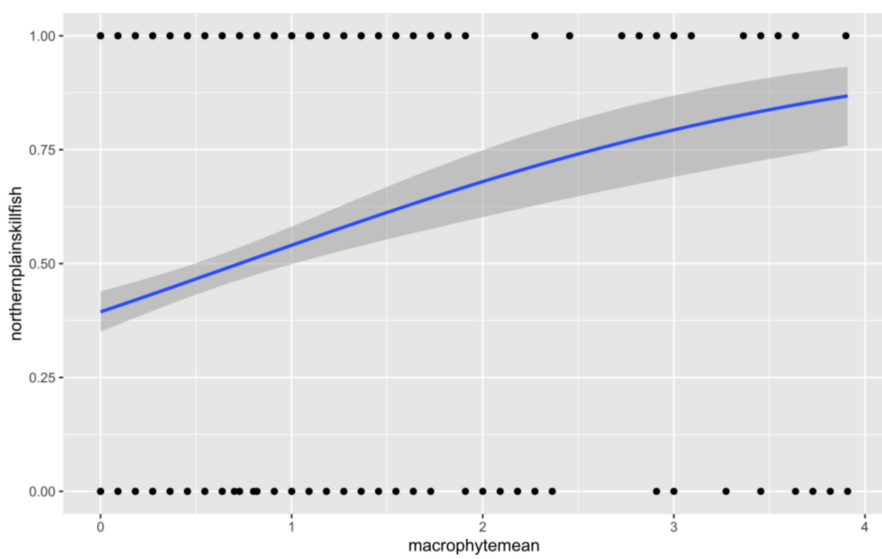




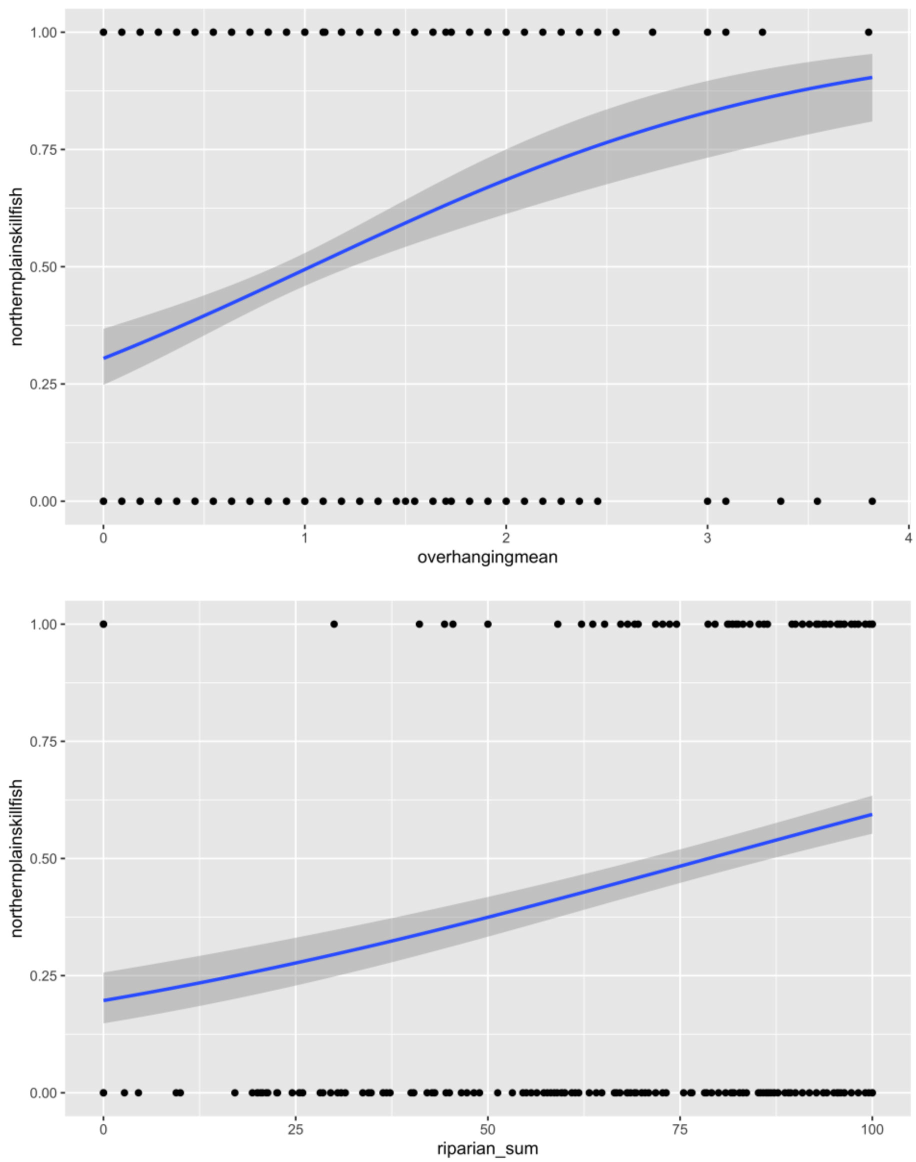
1.00
0.75
0.50
0.25
0.00
northernplainskillfish
0
25
50
75
100
embeddedness



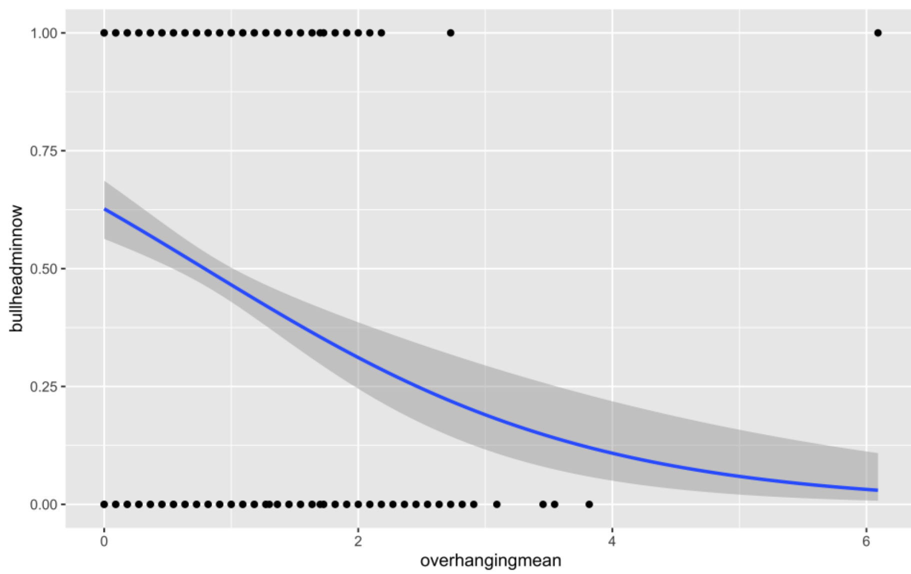
1.00
0.75
0.50
0.25
0.00
northernplainskillfish
0.0
0.5
1.0
1.5
2.0
hsubstrate

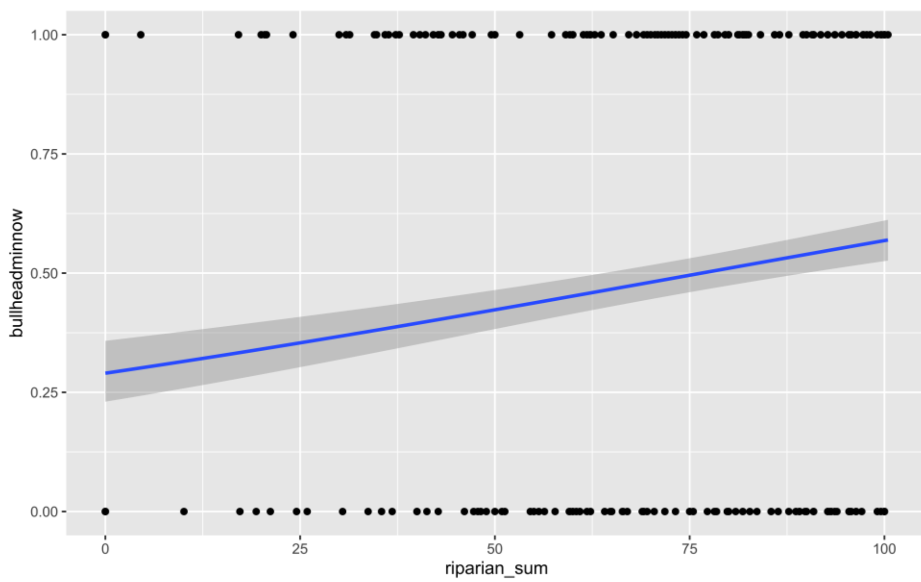


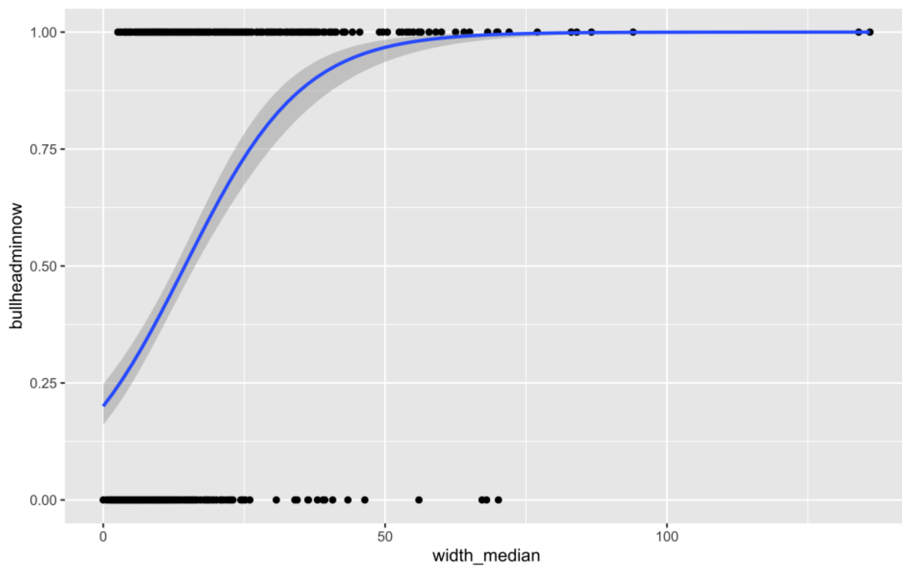
1.00
0.75
0.50
0.25
0.00
northernplainskillfish
0
1
2
3
4
macrophytemean

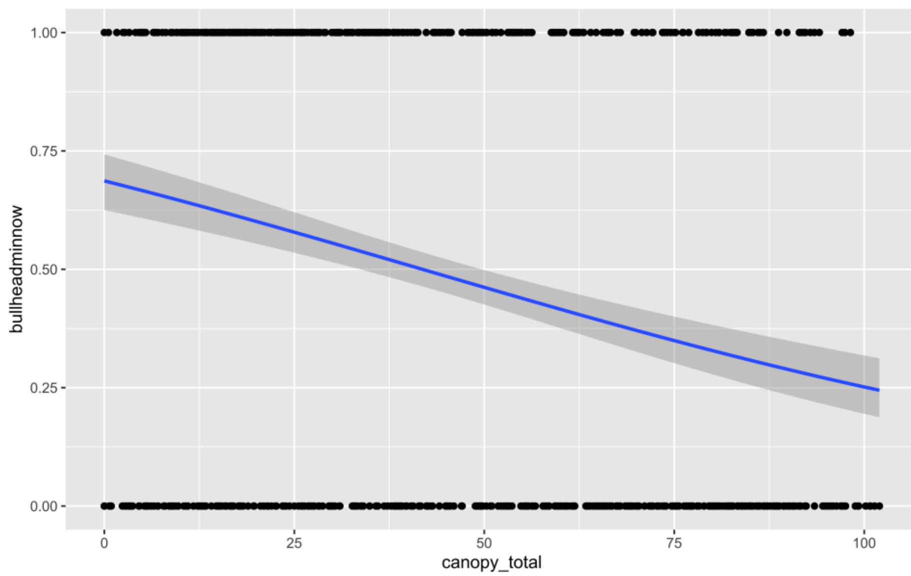


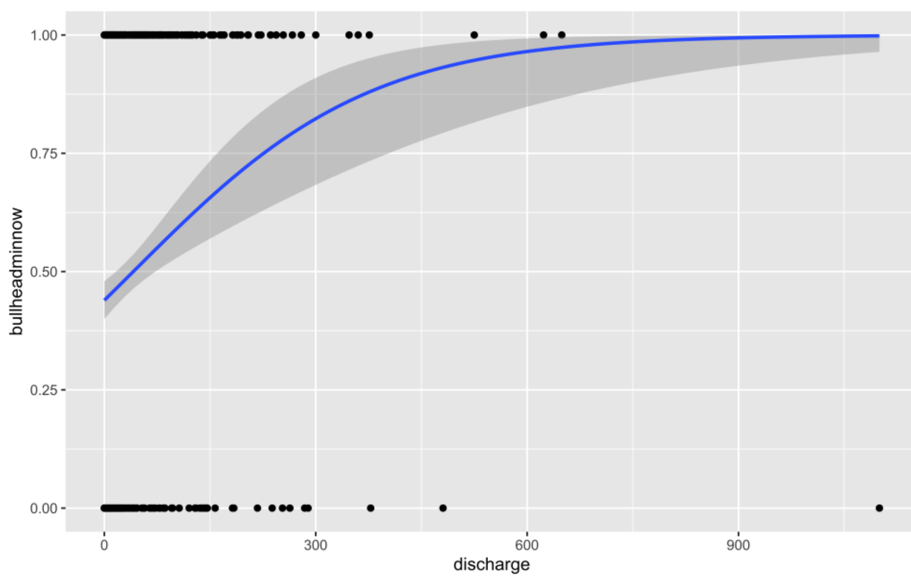
Appendix B Figure B.6. Bullhead Minnow Probability Curves


1.00
0.75
0.50
0.25
0.00
bullheadminnow
0
2
4
6
overhangingmean

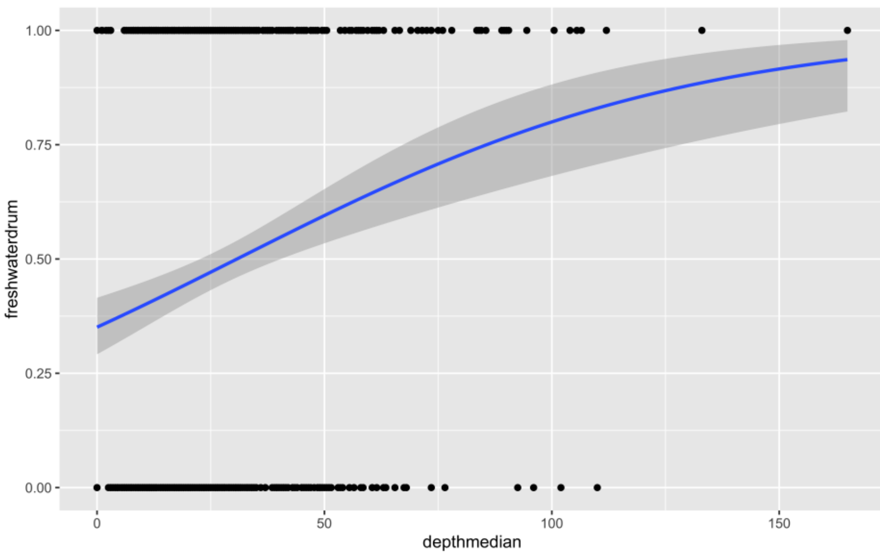

1.00
0.75
0.50
0.25
0.00
bullheadminnow
0
25
50
75
100
riparian_sum

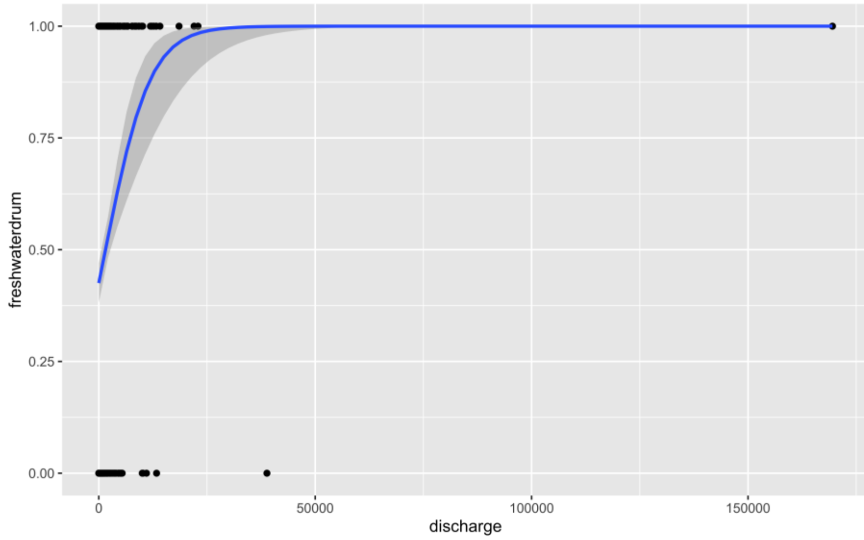

1.00
0.75
0.50
0.25
0.00
bullheadminnow
0
50
100
width_median

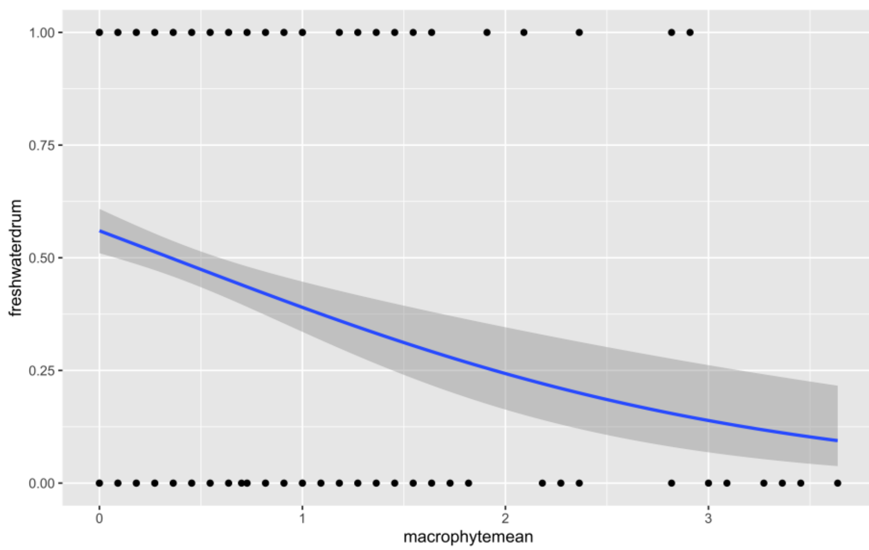

1.00
0.75
0.50
0.25
0.00
bullheadminnow
0
25
50
75
100
canopy_total

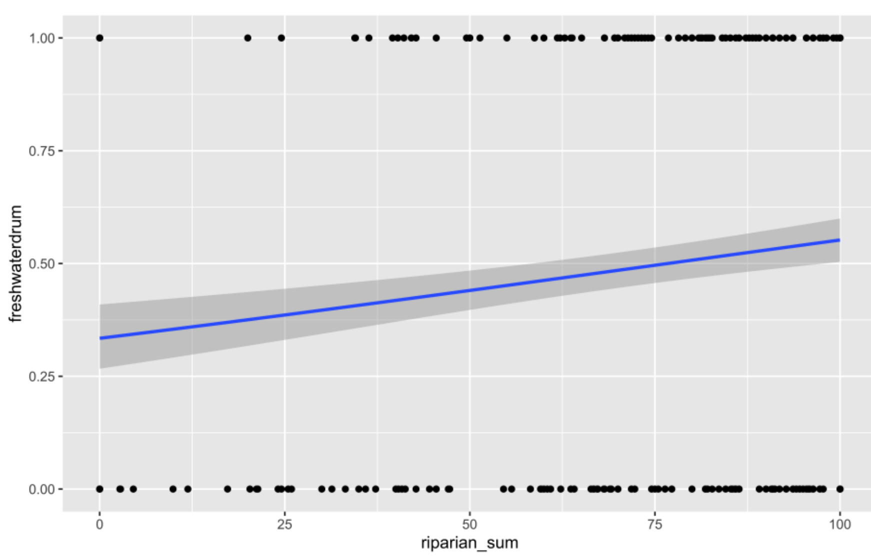

1.00
0.75
0.50
0.25
0.00
bullheadminnow
0
300
600
900
discharge

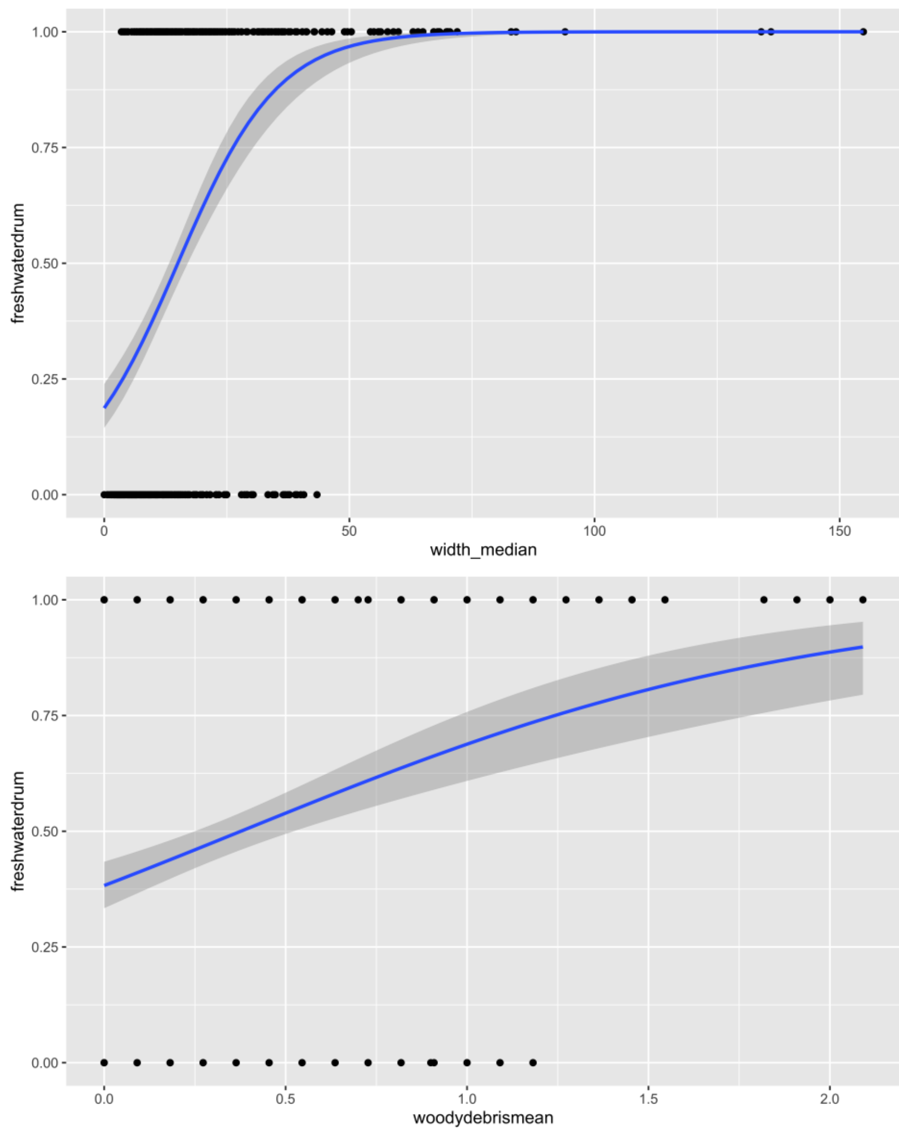
Appendix B Figure B.7. Freshwater Drum Probability Curves



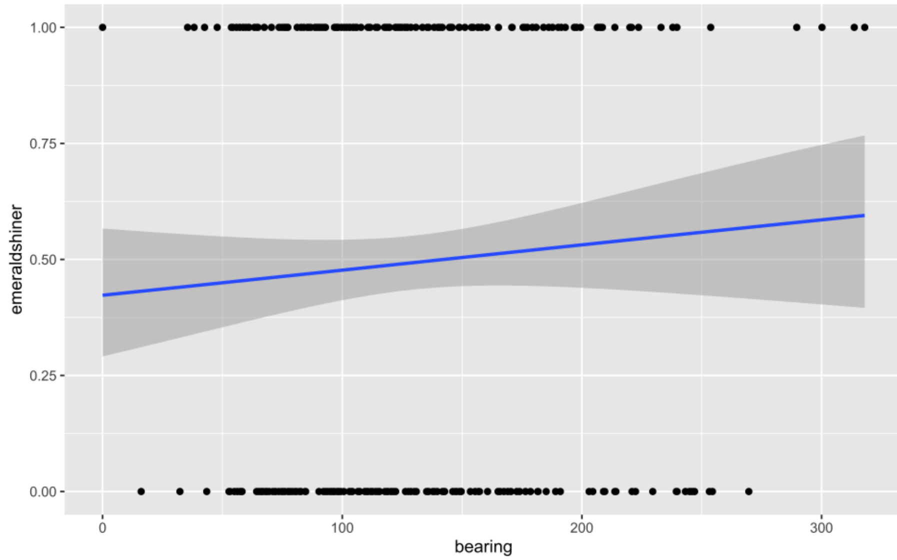

freshwaterdrum
1.00
0.75
0.50
0.25
0.00
0
50000
100000
150000
discharge

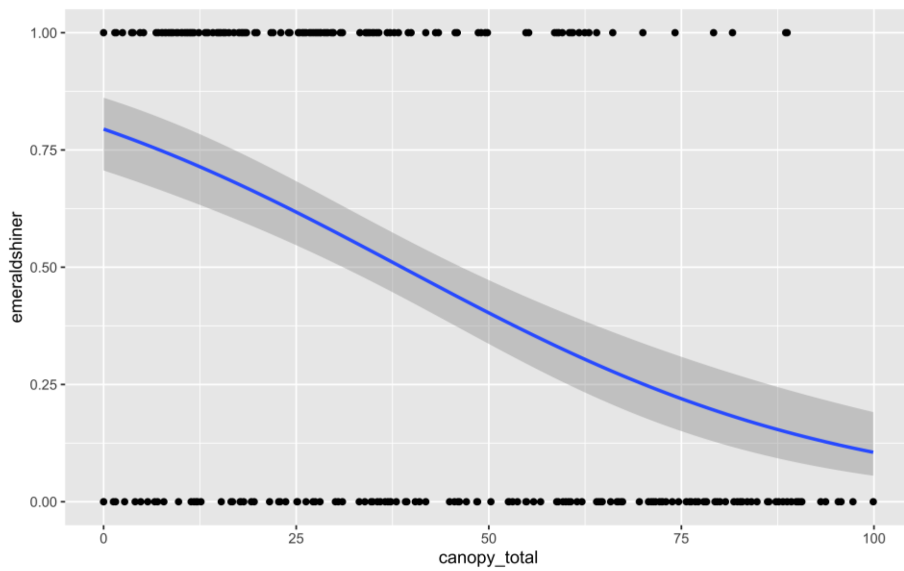

freshwaterdrum
1.00
0.75
0.50
0.25
0.00
0
1
2
3
macrophytemean

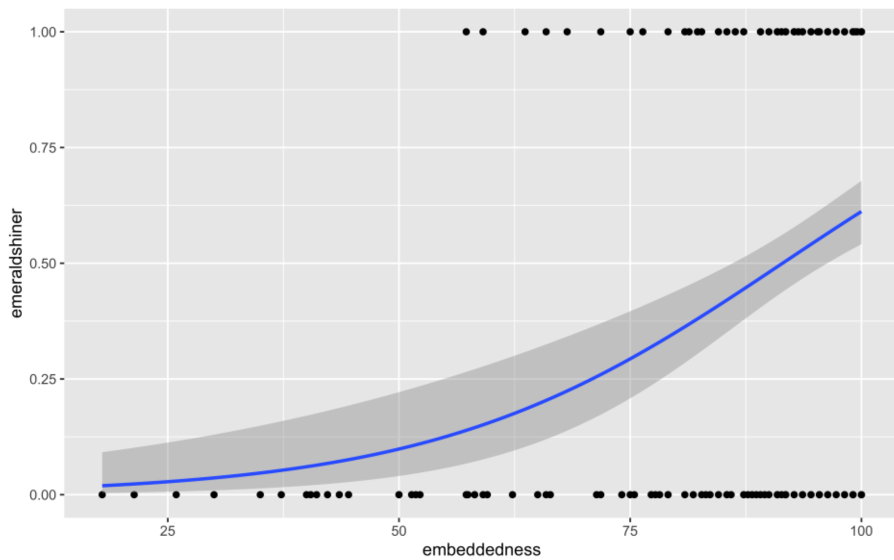

freshwaterdrum
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
riparian_sum



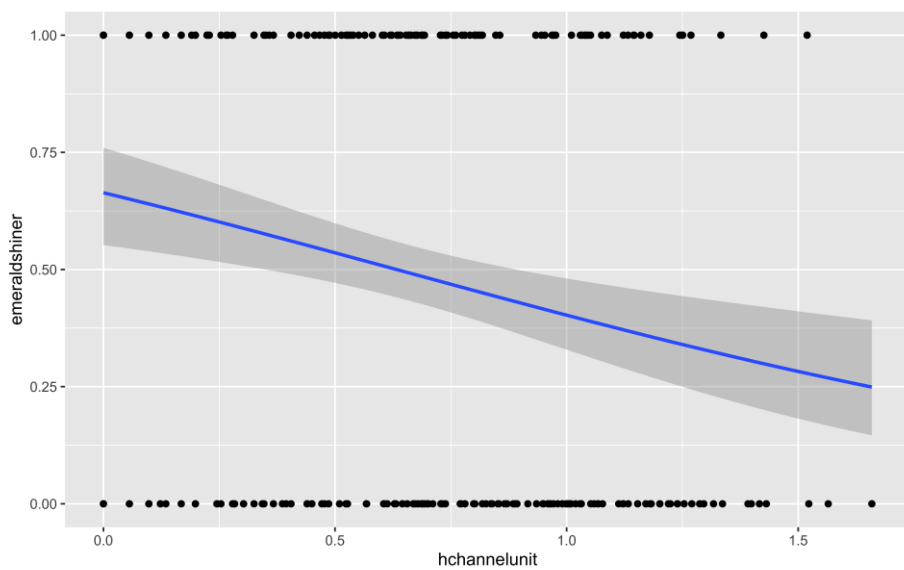
Appendix B Figure B.8. Emerald Shiner Probability Curves


1.00
0.75
0.50
0.25
0.00
emeraldshiner
0
100
200
300
bearing

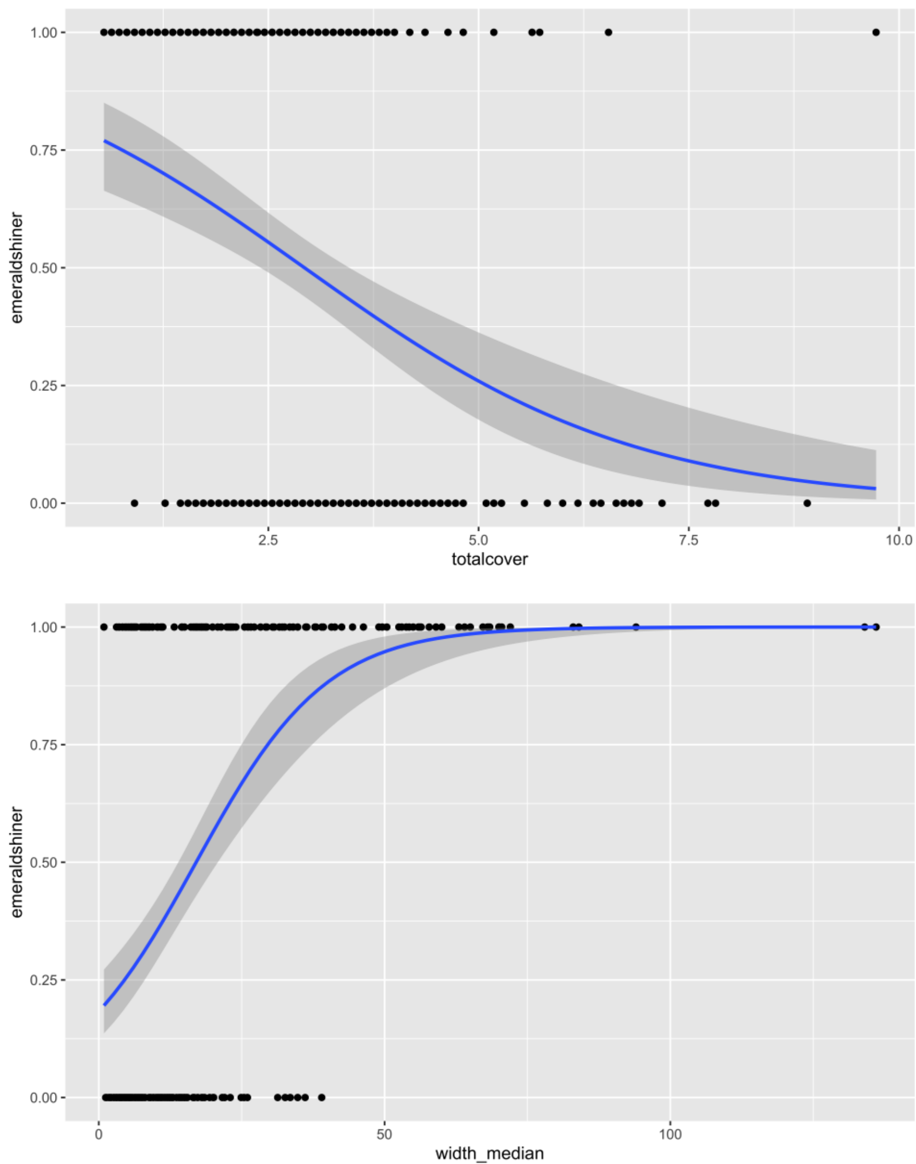

1.00
0.75
0.50
0.25
0.00
emeraldshiner
0
25
50
75
100
canopy_total



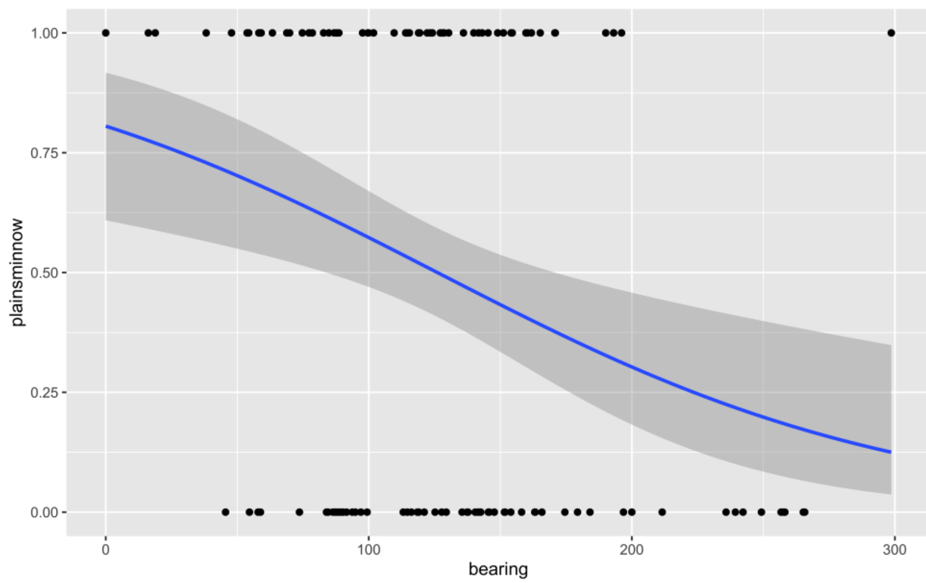
emeraldshiner
1.00
0.75
0.50
0.25
0.00
25
50
75
100
embeddedness

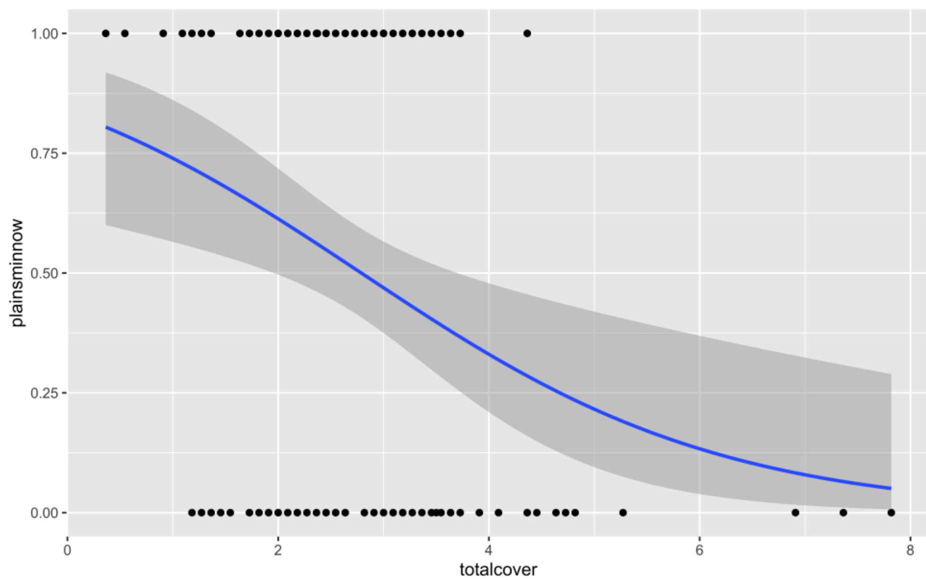


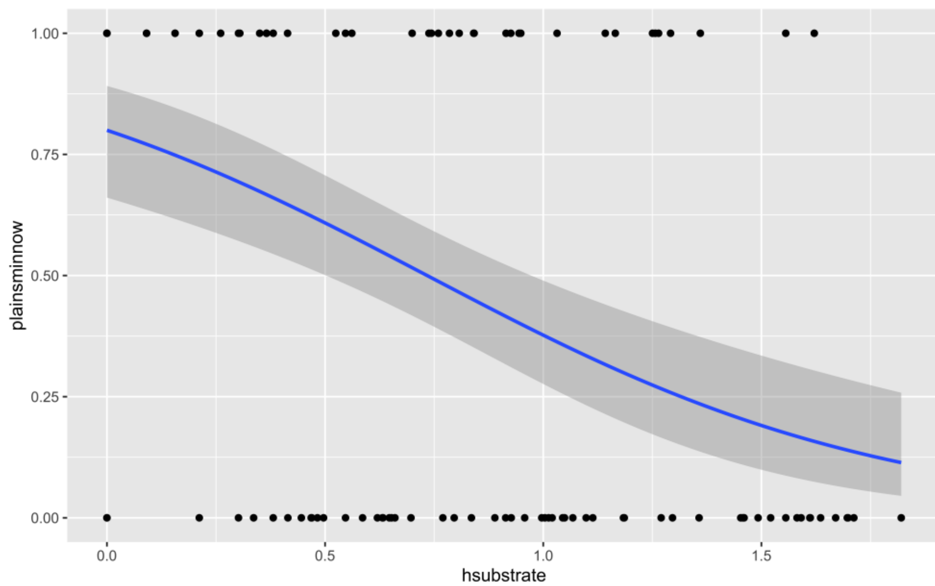
emeraldshiner
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
hchannelunit

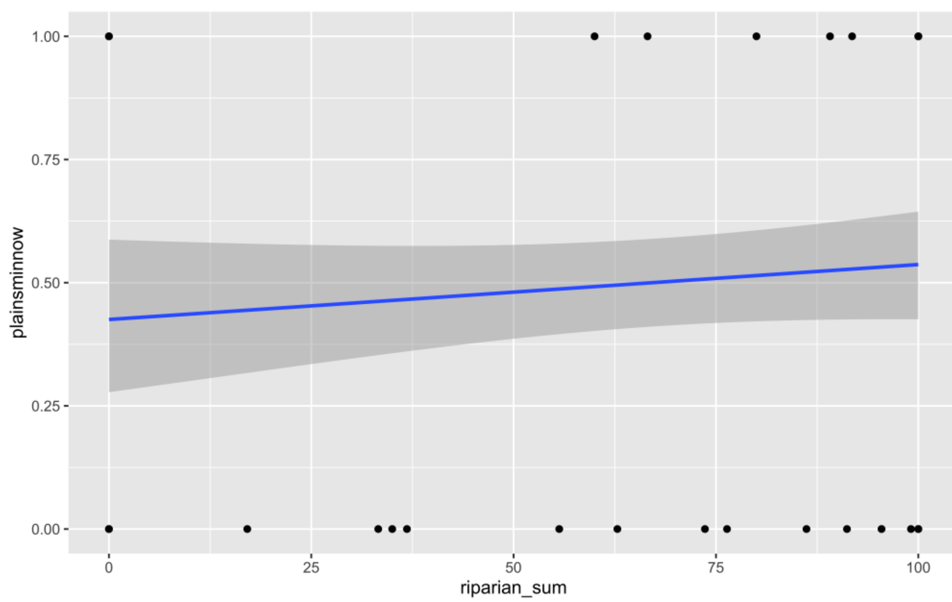


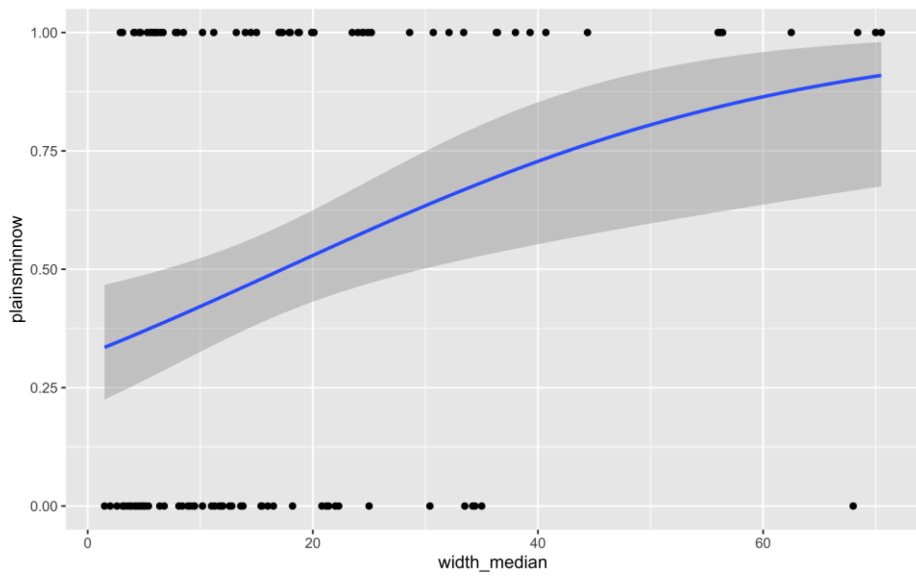
Appendix B Figure B.9. Plains Minnow Probability Curves


plainsminnow
1.00
0.75
0.50
0.25
0.00
0
100
200
300
bearing

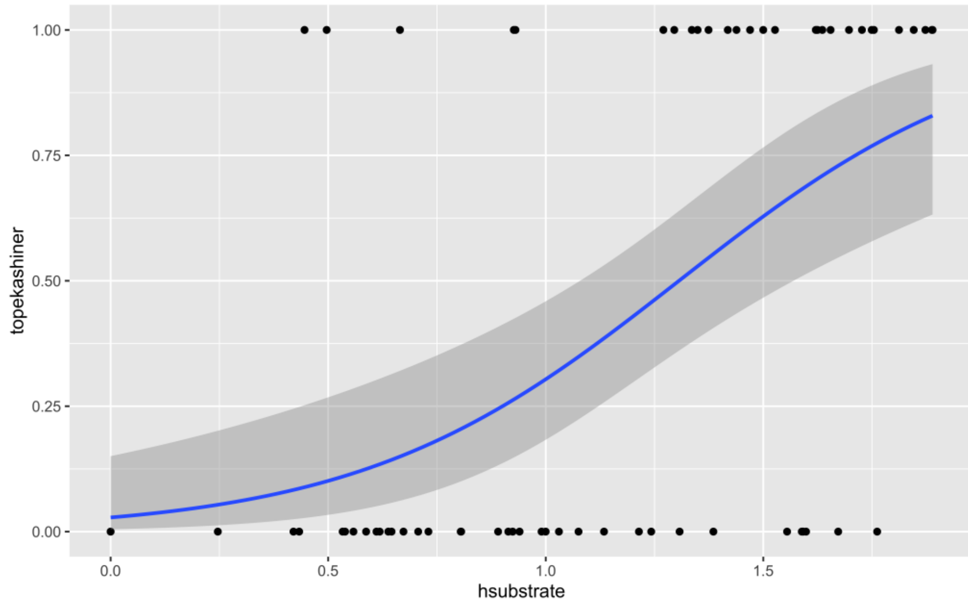

plainsminnow
1.00
0.75
0.50
0.25
0.00
0
2
4
6
8
totalcover

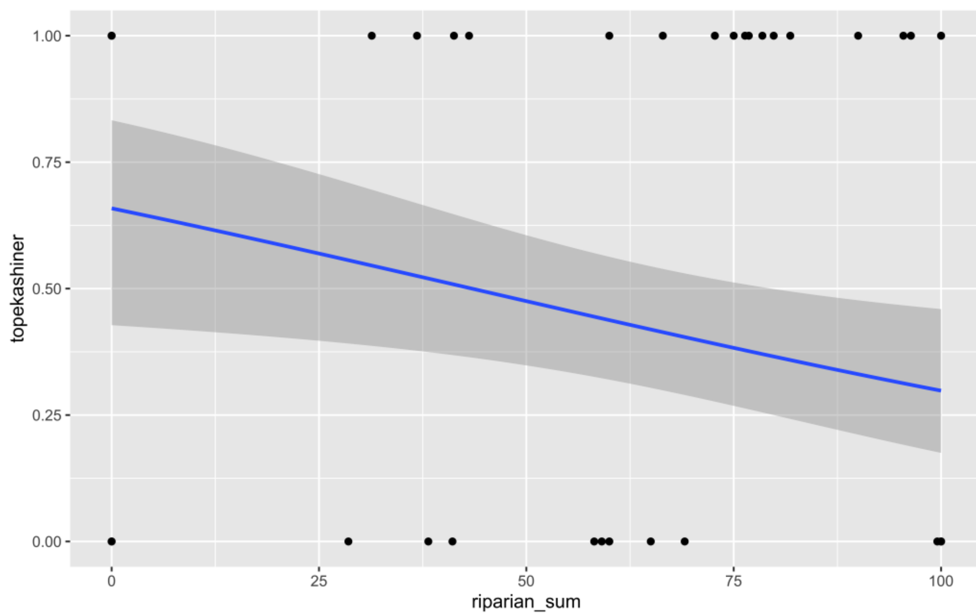

1.00
0.75
0.50
0.25
0.00
plainsminnow
0.0
0.5
1.0
1.5
hsubstrate

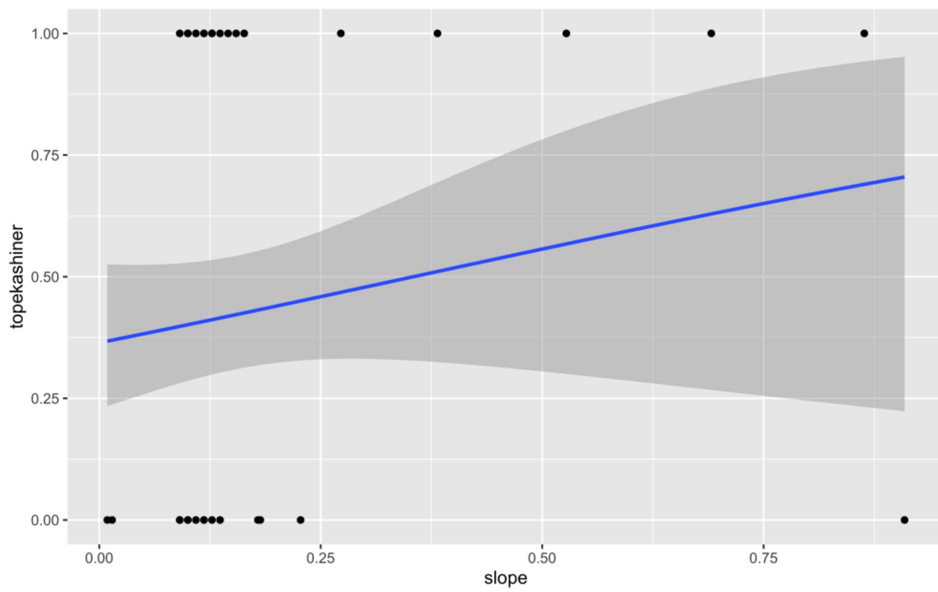

1.00
0.75
0.50
0.25
0.00
plainsminnow
0
25
50
75
100
riparian_sum

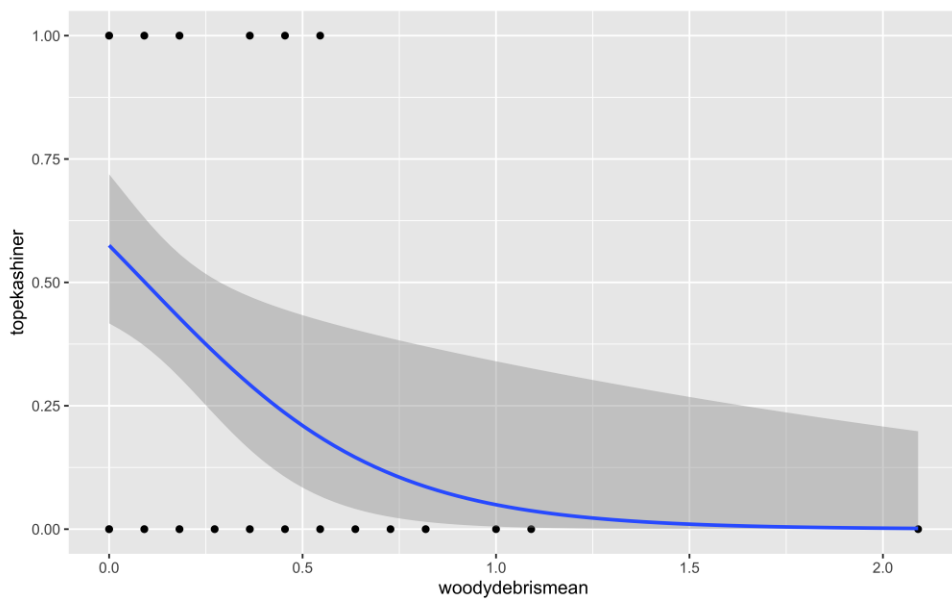


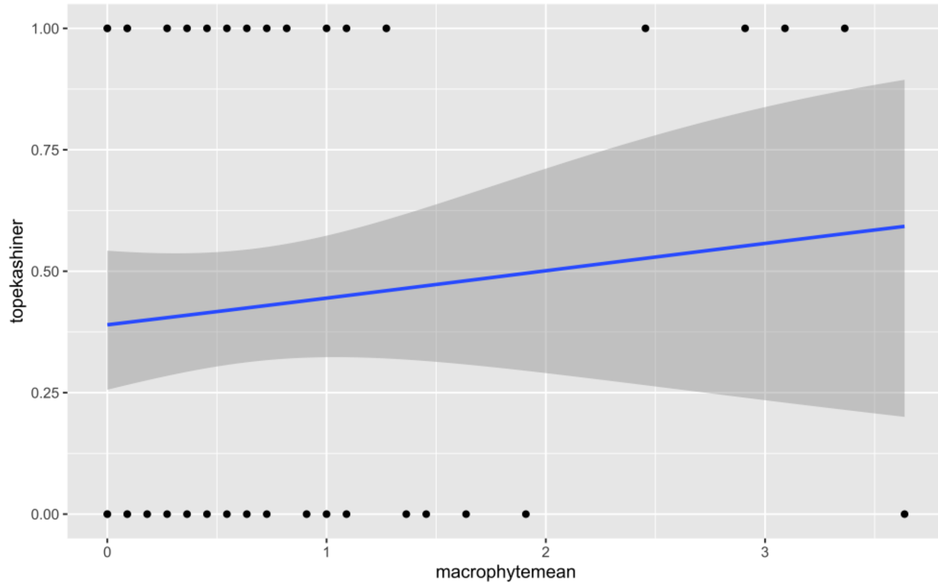
Appendix B Figure B.20. Topeka Shiner Probability Curves

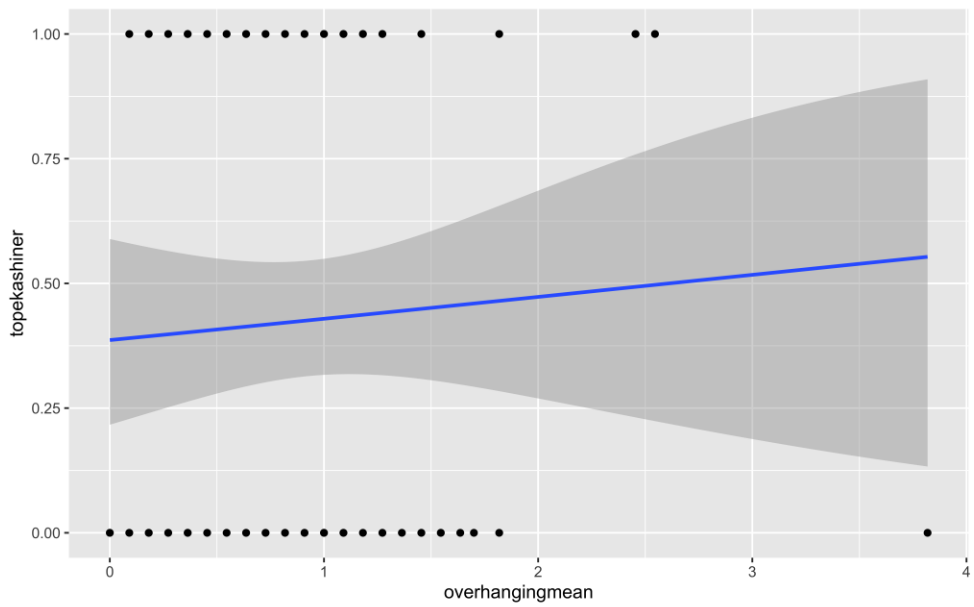

1.00
0.75
0.50
0.25
0.00
topekashiner
0.0
0.5
1.0
1.5
hsubstrate


1.00
0.75
0.50
0.25
0.00
topekashiner
0
25
50
75
100
riparian_sum


1.00
0.75
0.50
0.25
0.00
topekashiner
0.00
0.25
0.50
0.75
slope


1.00
0.75
0.50
0.25
0.00
topekashiner
0.0
0.5
1.0
1.5
2.0
woodydebrismean


1.00
0.75
0.50
0.25
0.00
topekashiner
0
1
2
3
macrophytemean


1.00
0.75
0.50
0.25
0.00
topekashiner
0
1
2
3
4
overhangingmean