Diedre Brown

INFO 640/Fall 2019 Data Analysis

Predictive Data Analysis

22 October 2019

Using Predictive Data Analysis to Determine Population Correlation in Three Species of Freshwater Shrimp in Quebrada Prieta

Defined as the myriad of interactions that have made Earth habitable for billions of years (Carrington, 2018), understanding biodiversity is central to developing and evaluating sustainable means of ecological resilience. In "What is biodiversity and why does it matter to us?", Damien Carrington describes biodiversity as being "comprised of several levels, starting with genes, then individual species, then communities of creatures and finally entire ecosystems, such as forests or coral reefs, where life interplays with the physical environment" (Carrington, 2018). Eager to gain knowledge of biodiversity and understand how data analysis can play a role in defining 'what biodiversity means to us', I selected to analyze a dataset on shrimp populations in Quebrada Prieta in Puerto Rico. Though I know nothing about freshwater shrimp, as this definition of biodiversity implies that population growth of various species is correlated within an ecosystem and time, I was curious to see if this logic could be inferred based on found population data.

Data and Exploratory Analysis

To find data related to this topic, I searched the *Nature Research* journals.[1] Though *Nature* does not host data ("Recommended Data Repositories | Scientific Data," n.d.), the data

[1] As a leader in scientific publication, *Nature* has a strict policy regarding data the accuracy and validity of data in submitted publications. Seeing a need to "improve the infrastructure supporting the reuse of scholarly data," since 2016 *Nature* has endorsed the **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR) Data Principles, and has developed a publication and data management system called *Scientific Data* (Wilkinson et al., 2016).

collected for this analysis comes from one of its recommended data repositories the Environmental Data Initiative Data Portal (EDI). While data collection varies from individual to individual, the one assurance I had in using data from EDI was that contextually rich and came with verifiable metadata.

This study comprises three datasets (each with six variables) on the freshwater shrimp populations of *Atya Ianipes, Xiphocaris elongate*, and *Macrobrachium spp.*, which was collected six times yearly from 1988-2016 at in four pools in the Quebrada Prieta Valley within the Luquillo Forest.[2,3] Each dataset is specific to each pool, and includes three species-count specific variables (ATYACPUE, XIPHCPUE, and MACCPUE, respectively), as well as, a variable for year collected and month collected. The three datasets were joined into two separate vectors: one long dataset of four variables and 1,023 observations, and one wide dataset of five variables and 341 observations. Though the notes on the datasets did state that counts of shrimps were equal to:

> The total number of freshwater species shrimps captured divided by the number of traps from the corresponding pool in the Quebrada Prieta and then released. Number of traps in each pool can vary but usually are 34, 3, and 2 for Pools 0, 8, and 15 respectively. Record is missing when data is missing (Crowl, 2017)

upon reviewing the unified dataset, I found that there were inconsistencies in the date of collection among pools (i.e. in some pools there were six readings collected for the year in others five or seven), and see that there were 120 observations taken at Pool 0, 109 observations taken

[2] Since 1988, the collection site is part of the National Science Foundation (NSF) Long-Term Ecological Research (LTER) Program, which was established in 1980 to understand the dynamics of ecosystem processes (“ABOUT US | Luquillo LTER,” n.d.).

[3] While the site maintains data on four pools (0, 8, 9, and 15, respectively), only three were used for this study, as the data available on Pool 9 only covered weekly sampling since 1993.

at Pool 8, and 112 observations taken at Pool 15. Moreover, this coupled with the realization that the counts were averages of counts species found per pool and not the actual counts meant that information could be misinterpreted if I further averaged the counts by year[4] without full population data. Instead, I elected to start by exploring the data graphically as is, then through predictive analysis determine if this data could provide the information I was looking for: a correlation between population size and time, and a correlation between species population sizes.

A preliminary scatterplot of the unified data (Figure 1), suggests a relationship between the counts of the *Atya lanipes* and *Xiphocaris elongate* shrimp in all pools, however, a relationship between any species and the *Macrobrachium spp*. could not be readily determined from this plot. I plotted all values of each species on a base ten logarithmic scale (Figure 2) to see if by expanding the coordinate scales, more graphical information could be revealed. And, to isolated the data on *Macrobrachium spp*, I created a separate plot (Figure 3). These plots illustrated a consistency changes in population size of all species in the period of 1995-2015. After calculating the mean of all species, respectively (Table 1), I assumed that *Atya lanipes* was the most independent species per pool and *Macrobrachium spp* was the most dependent species per pool. Using this assumption, I plotted the species in relation to each other per pool (Figure 4, Figure 5, Figure 6).

Using exploratory correlograms (Figure 7 and Figure 8),R calculated a 0.135 correlation coefficient between collection date and counts of all species, and values of 0.005, 0.528, -0.255, as each species (ATYACPUE, XIPHCPUE, and MACCPUE), relates to time respectively.

[4] In other words, I could not just take the average for twelve months of data when I only had information for five months.

Though overall these values are small, they are enough to suggest further study of these variables.

Conclusions from Predictive Analysis (Table 2)

Though most of the bivariate data tested cannot be said to be mathematically correlated, as the found coefficients of determination showed an almost absence of variability for all tests. However, as the margin of error was most narrow for the Species ~Time Study, I believe that the relationship between population of the shrimp species and time is likely to have a correlation not accounted for by these variables alone. Similarly, as the *Xiphocaris elongate* to *Atya lanipes* model showed a 10% variability, I believe that the relationship between these species is a better representation of unseen data. What this dataset did illustrate is that there is a lot of unseen data—uneven count samplings, possibility to test with other variables. While I used lines of regression to help with this issue, without knowing the cause to why these counts are missing (natural disaster, population not present, competition between species, lack of resources, etc.), a fair and true assessment of the data cannot be attained.

Aside from the need for more variables (such as temperature, time of day, bacterial count, mineral content, debris, weather condition, etc.), due to the extreme outliers, each model did pose questions for further study during specific periods:

1. Species ~Time Study: June 1996. There was a high average count of the *Atya lanipes* species (Count 152) in Pool 15. As the average counts for the *Xiphocaris* and *Macrobrachium* shrimp were only 13.5 and 0, respectively, what conditions caused *Atya* to proliferate so?

2. *Xiphocaris ~Atya lanipes*: In November 2015, the average count of the *Xiphocaris* (Count 113) population exceeded that of the *Atya lanipes* (Count 60) population in Pool 8. What conditions caused this dynamic shift?
3. *Macrobrachium ~ Atya lanipes*: In both December 1989 and June 1996, the *Atya lanipes* shrimp showed low average counts (6 and 25, respectively) for its species compared to *Macrobrachium in Pool 8*. What caused this change in Pool 8?
4. *Macrobrachium ~ Xiphocaris*: In Pool 0 in January 1988 both the *Atya lanipes* and *Xiphocaris* were 8 and 8.25, respectively. What caused these shrimp populations to be nearly equal? And in Pool 8 in June 1996, a similar evenness in the average counts of *Atya lanipes* and *Xiphocaris* shrimp.
5. As June 1996 showed up in a number of the outliers, what environmental conditions caused the extremes in the shrimp populations?

With 1,023 observations, the list of questions for further study can go on. As answers to many of these would require additional data, the only *true* correlation I can make from the data is that an ecosystem's biodiversity is dependent on a multitude of variables.

Appendix A: Figures and Tables

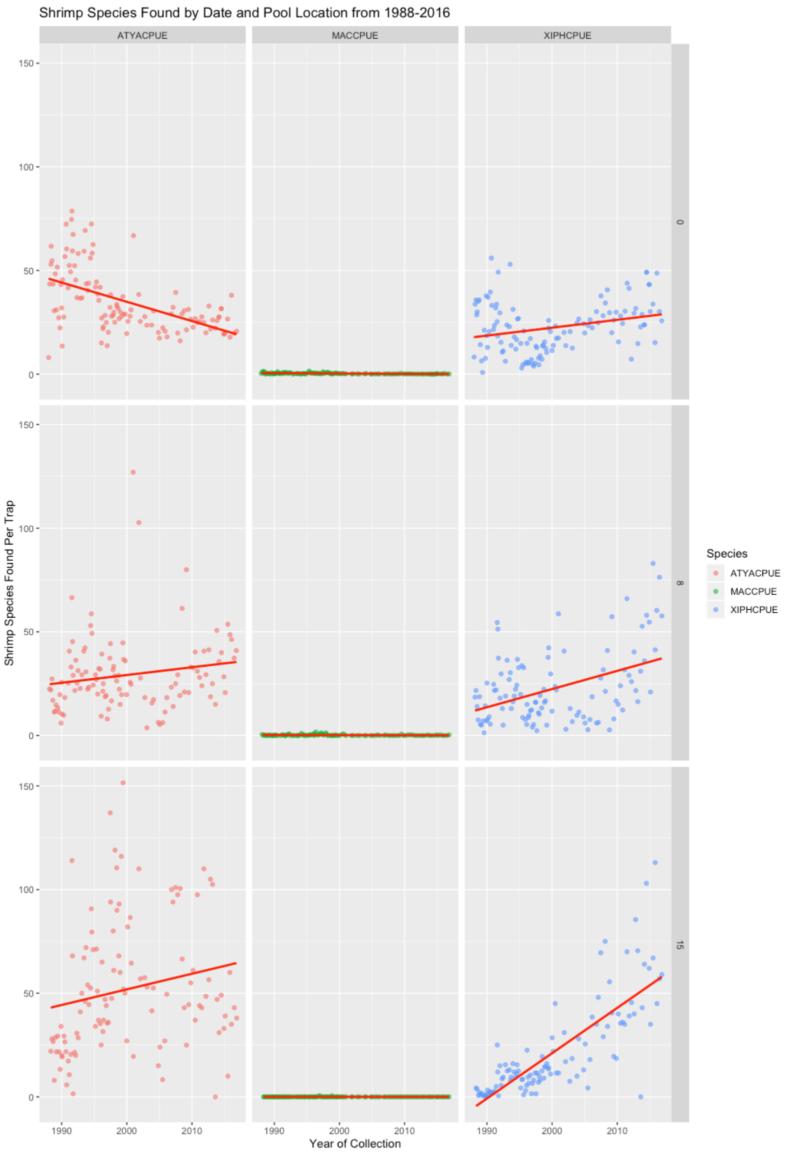

Figure 1. Preliminary Analysis of Shrimp Species Found by Data and Pool Location

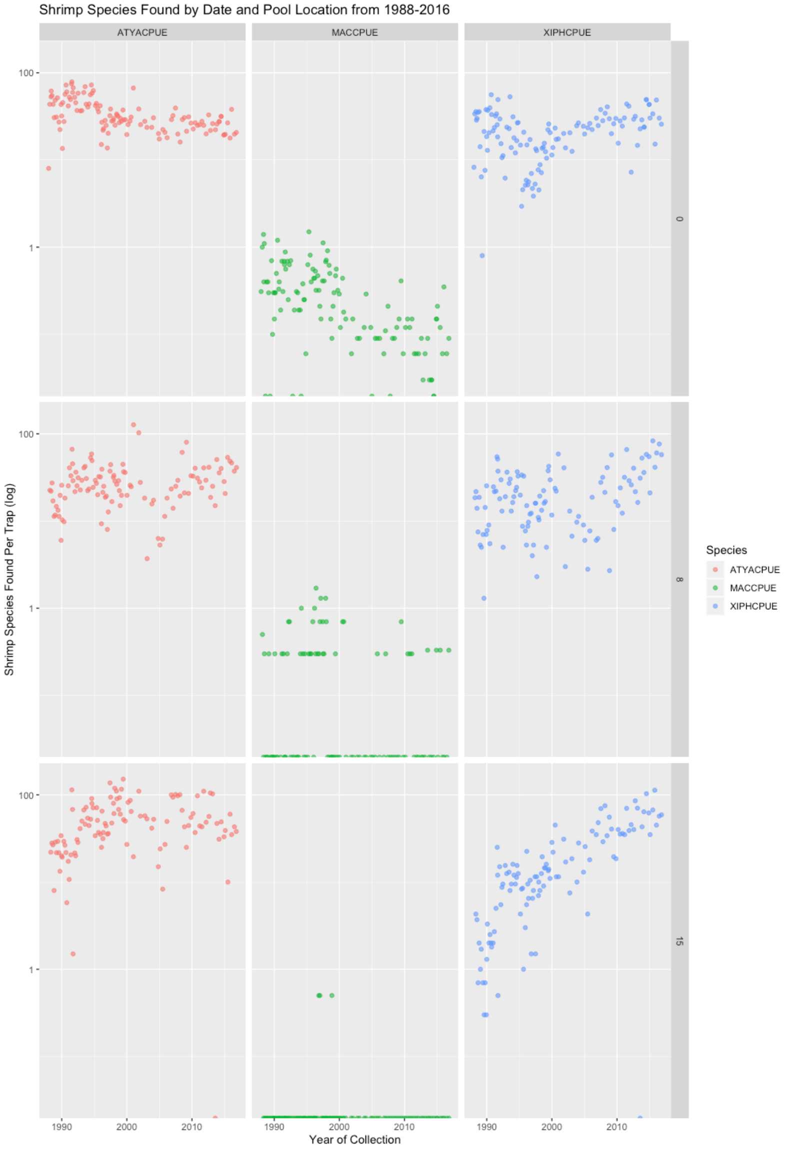

Figure 2. Logarithmic View of Shrimp Species Found by Date

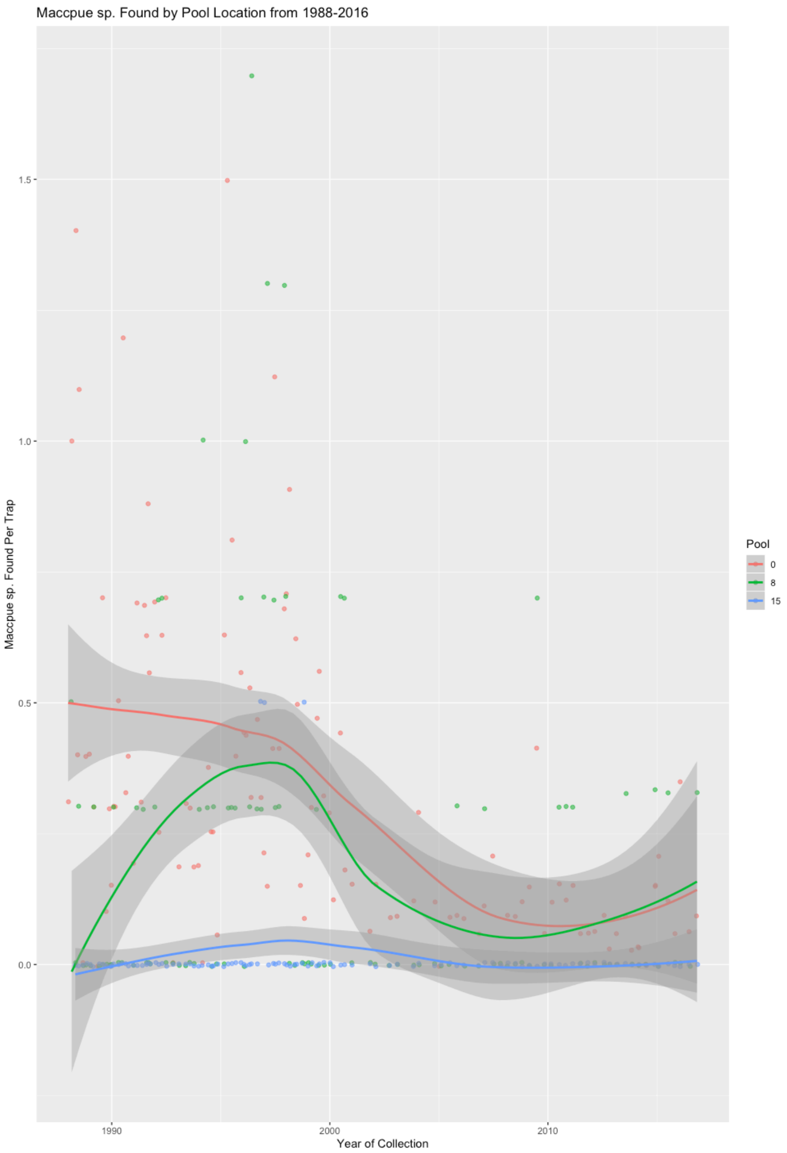

Figure 3. Macrobrachium spp Counts by Collection Year

Table 1. Overall Average Species Mean and Means of Average Species by Pool

Overall Average Species Means	***Atya lanipes***	***Xiphocaris elongate***	***Macrobrachium spp.***
	38.69	22.25	0.18 (≈ 1 as 0.18 of a living organism is physically impossible.)
Means of Average Species by Pool			
Pool 0	38.69	22.25	0.18 (≈ 1 as 0.18 of a living organism is physically impossible.)
Pool 8	38.69	22.25	0.18
Pool 15	38.69	22.25	0.18

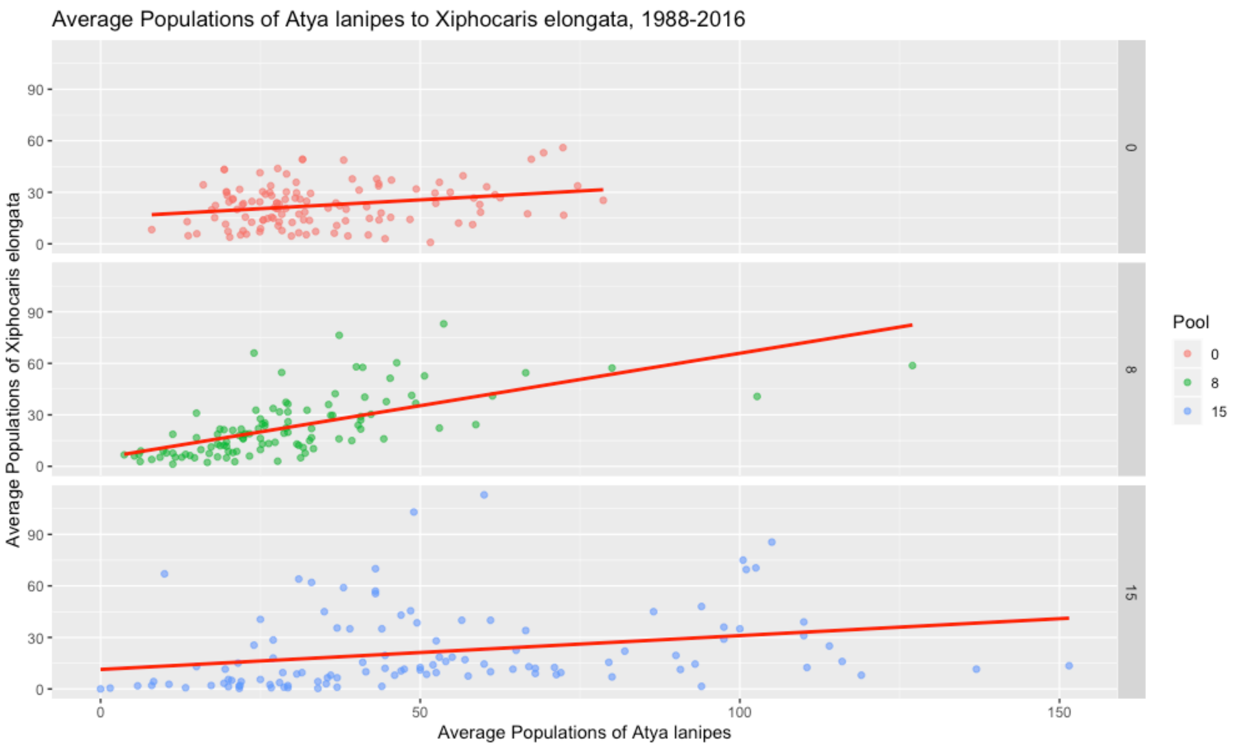

Figure 4. Comparison of Average Populations of Atya lanipes to Average Populations of Xiphocaris elongata by Pool

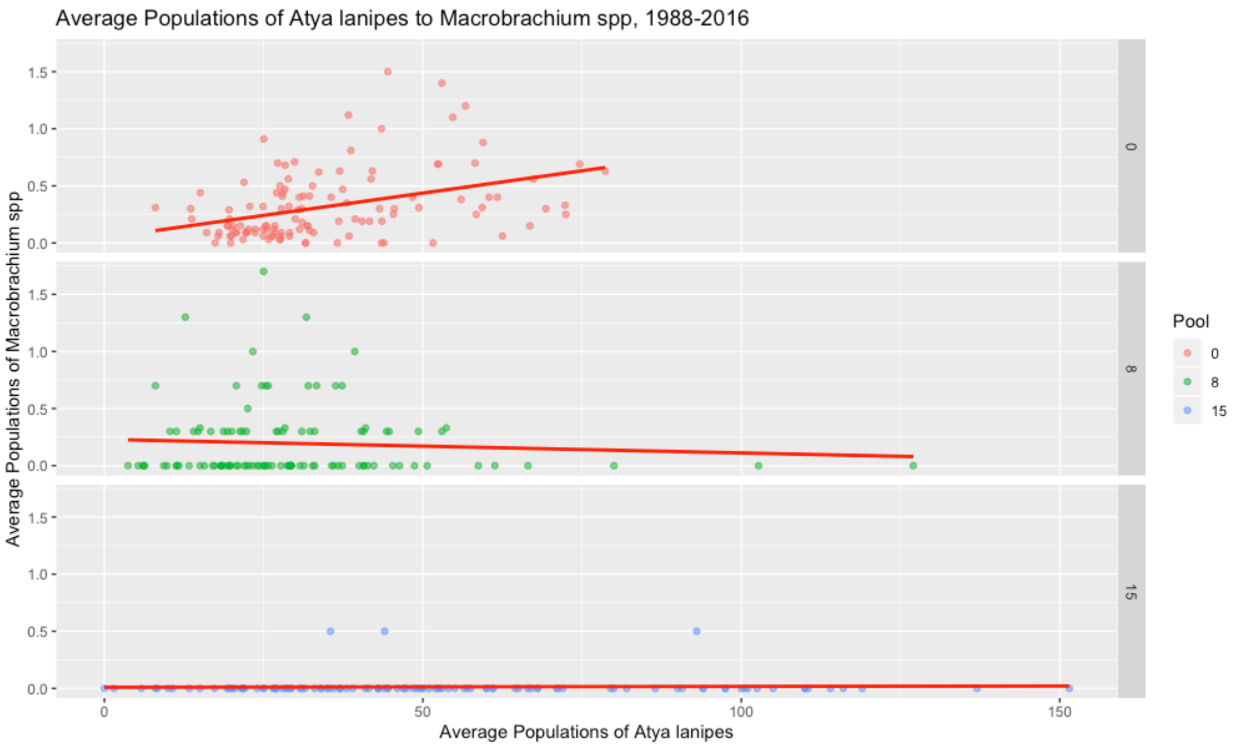

Figure 5. Comparison of Average Populations of Atya lanipes to Average Populations of Macrobrachium spp by Pool

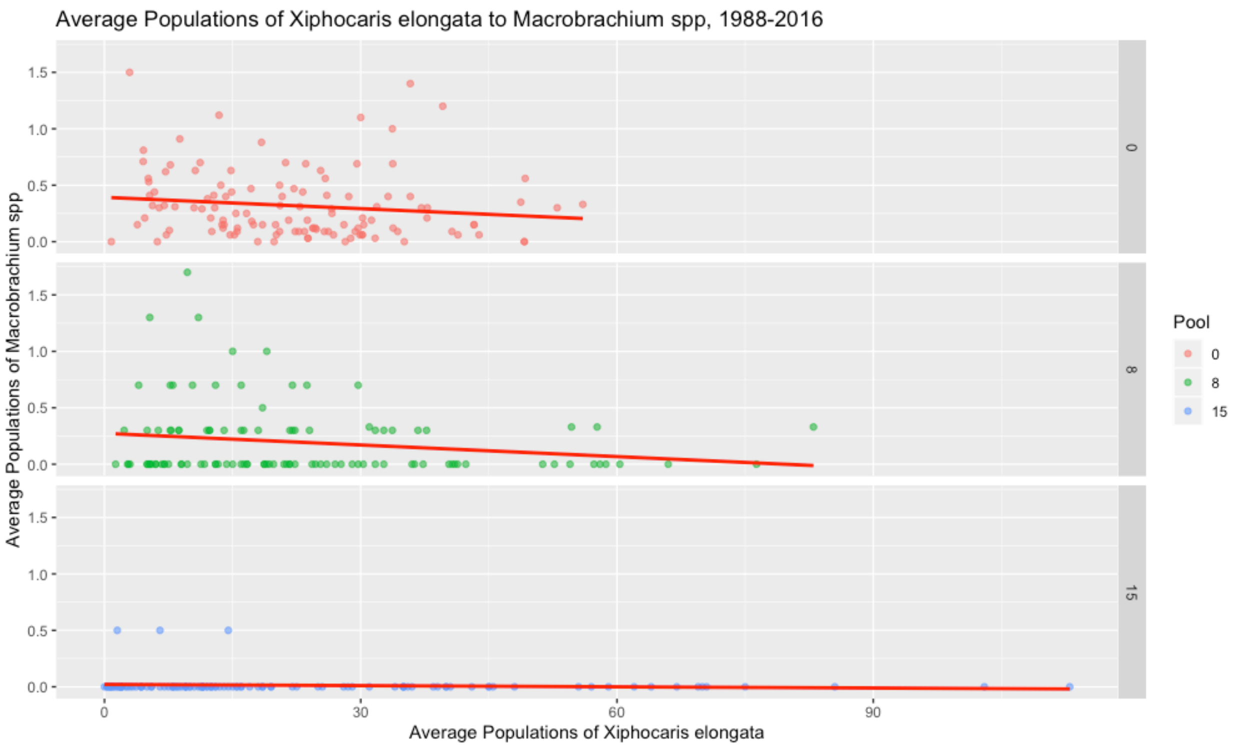

Figure 6. Comparison of Average Populations of Xiphocaris elongata to Average Populations of Macrobrachium spp by Pool

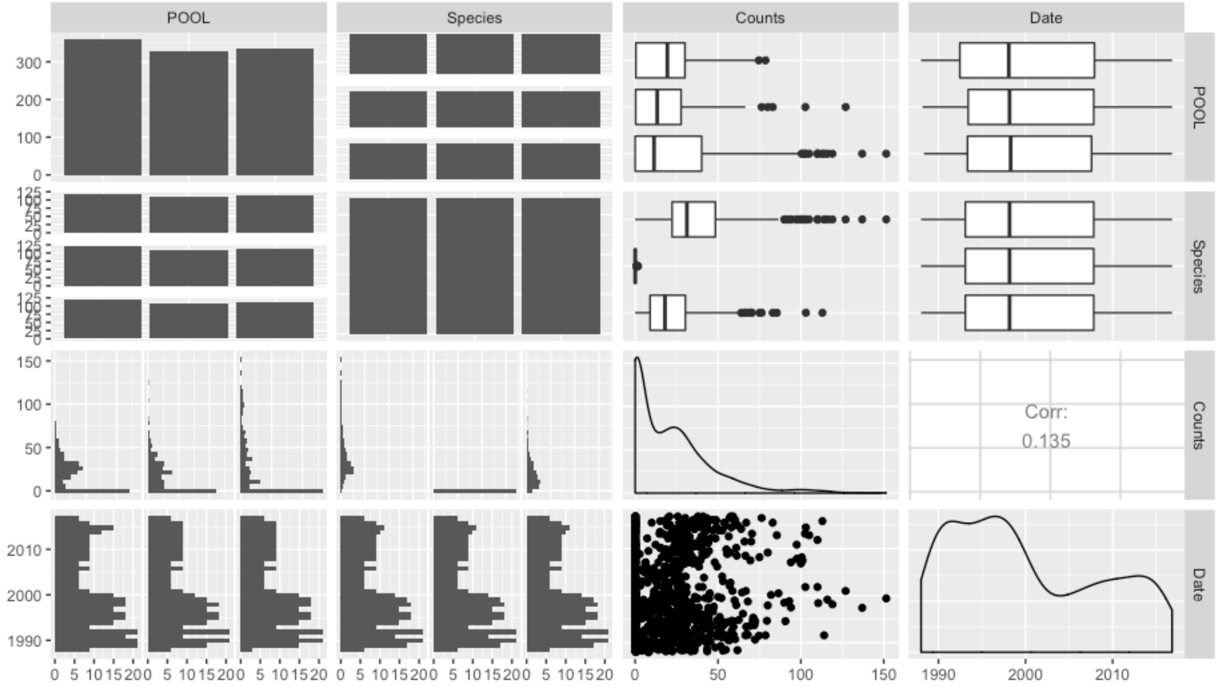

Figure 7. A correlogram of the shrimp dataset emphasizing the relationship between species and time.

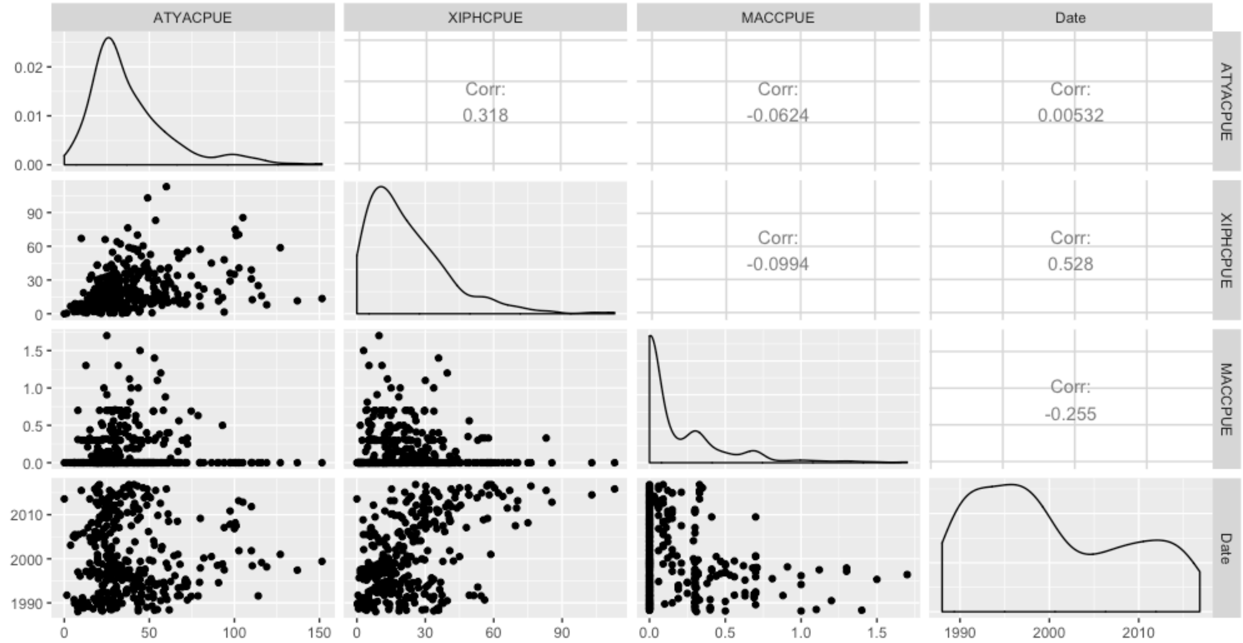

Figure 8. A correlogram of the dataset emphasizing relationships between species.

Table 2. Evaluation of Correlation Against Null Model

Linear Model (lm)	**Coefficients**		**Standard Error**	**Root Mean Squared Error (RMSE)**	**Null Model**	**Coefficient of Determination/ Percent of Variability (R^2)**	**Interpretation of Model**
	Intercept	**Measured Value**					
Species ~ Time (lm_spyr)	9.12	0.001	2.70	23.43 on 1021 df	20.37	0.02	Time is not a good determinant of size of species population size (Figure 9).
Xiph ~ Atya	13.3	0.232	1.72	17.08 on 339 df	22.25	0.101	The size of the *Atya* population is not a good determinant for the size of the *Xiphocaris* species (Figure 10).
Macc ~ Atya	0.21	-0.0007	0.03	0.29 on 339 df	0.179	0.004	The size of the *Atya* population is not a good determinant for the size of the Maccrobrachium species (Figure 11).
Macc ~ Xiph	0.21	-0.0016	0.02	0.29 on 339 df	0.179	0.010	The size of the *Xiphocaris* population is not a good determinant for the size of the Maccrobrachium species (Figure 12)

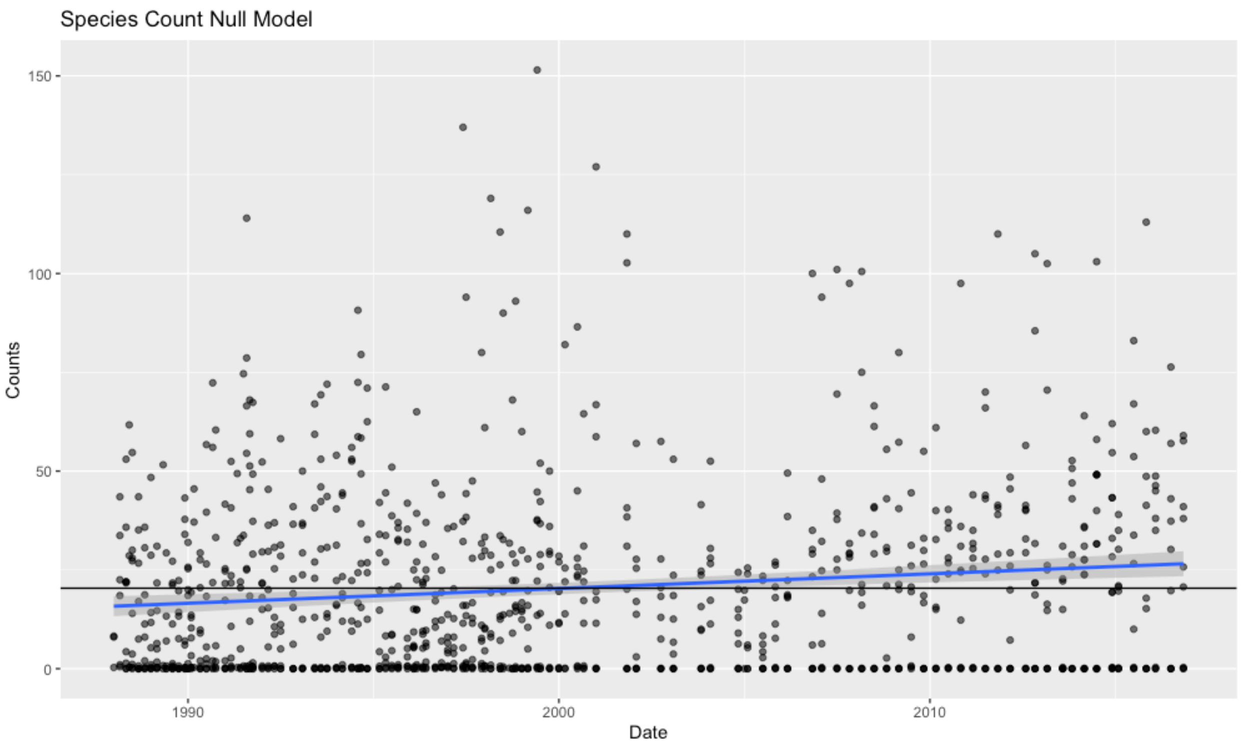

Figure 9. Species Count ~ Time Null Model

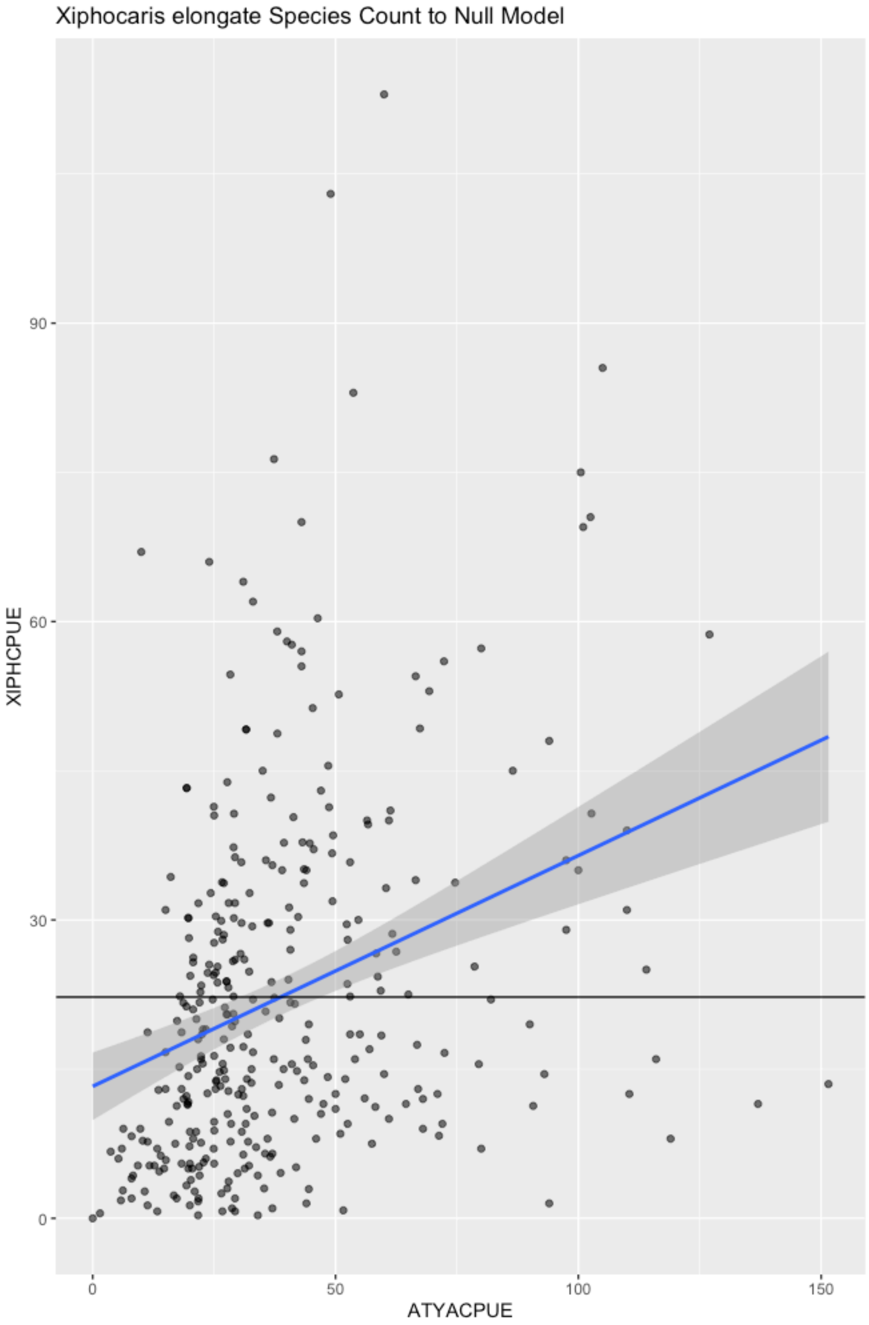

Figure 10. Xiphocaris~Atya Species Count to Null Model

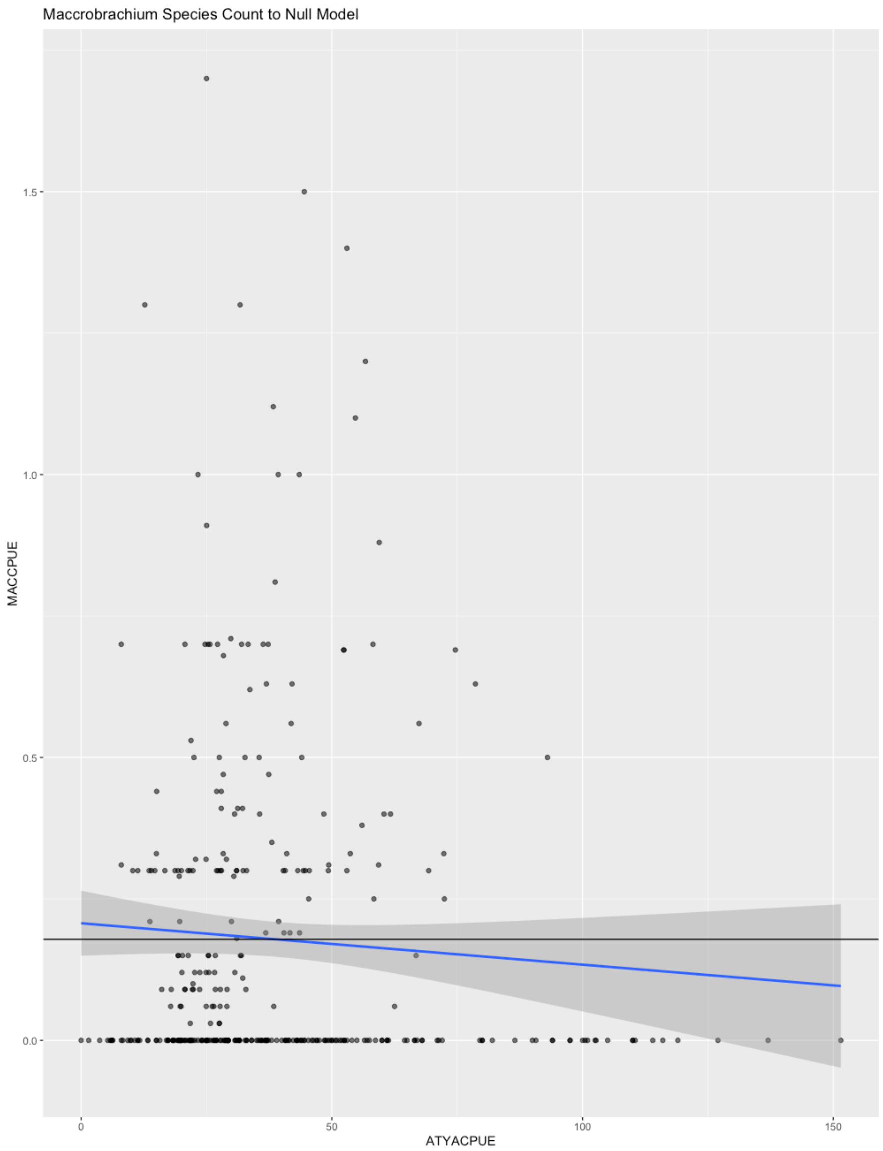

Figure 11. Maccrobrachium spp ~ Atya Species Count to Null Model

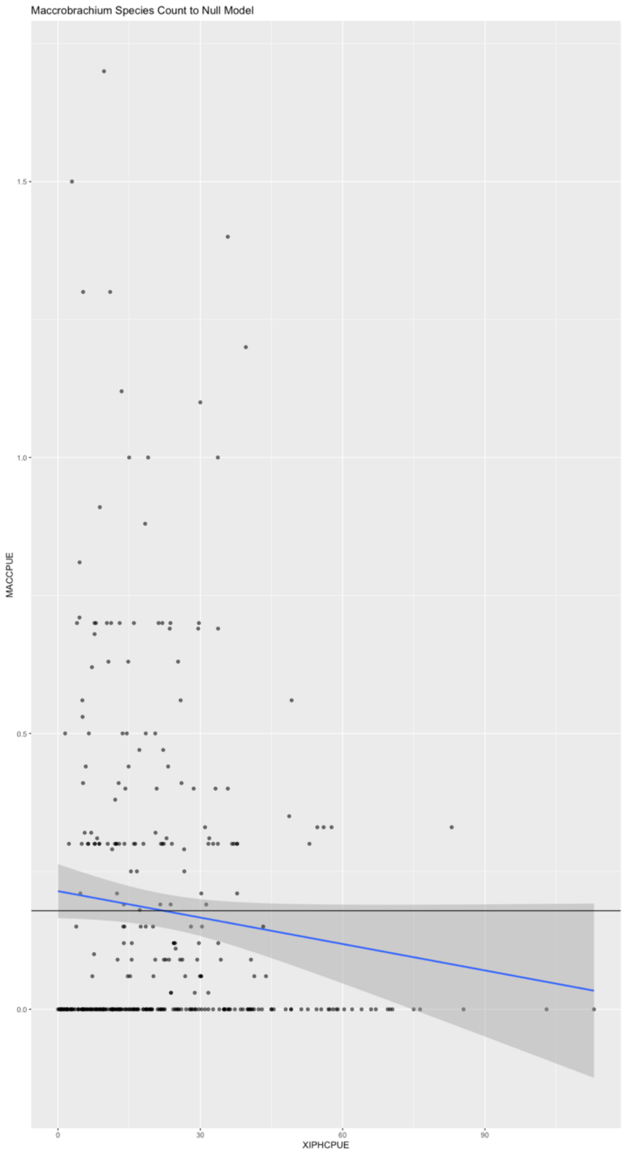

Figure 12. Maccrobrachium spp ~ Xiphocaris Species Count to Null Model

Appendix B: R Code

```
#Pratt Info 640 Fall 2019
#Diedre Brown; dbrow207@pratt.edu
#Predictive Data Analysis Project - Due 22 Oct 2019


#call libraries
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
library(broom) #broom helps clean things up and remerge dataframes
library(GGally) #GGally helps run multiple pair-wise correlations


####Import Datasets####
#Shrimp populations in Quebrada Prieta (Pools 0, 8, 9, 15) (El Verde)
#Source: Crowl T. 2010. Shrimp populations in Quebrada Prieta (Pools 0, 8, 9, 15) (El Verde). Environmental Data Initiative.
#https://doi.org/10.6073/pasta/f6c8497c780ecf619053dcd020d371f2. Dataset accessed 10/22/2019.
#Creator: Crowl, Todd
#Creator Publication Date: 2010-11-27
#Creator's Abstract:Freshwater shrimp from the Quebrada Prieta (a tributary to the Sonadora in the Espiritu Santu drainage, have been censused 6 times yearly since 1988.
#Atya lanipes, Xiphocaris elongata and Macrobrachium spp. are regularly captured and comprise the species in this data base.


#NOTES ON DATA:
#On further inspection of the creator's notes, I found that the count figure represents:
#Total number of freshwater species of shrimps captured divided by the number of traps from the corresponding pool in Quebrada Prieta and then released.
```

```
#Number of traps in each pool can vary but usually are 34, 3, and 2 for Pools 0, 8, 15
       respectively. Record is missing when data is missing.

#

#Though biannual and weekly figures were available, I just want an overall picture to evaluate
       for correlation and future study, so I will only use the biannual data.

#Only weekly data since 1993 on Pool 9 was available and therefore not included.

#



#biannual data for pools 0, 8, and 15 from 1988-2016

shrimppool_0_bia <- read.csv("Datasets/knb-lter-luq.54.945757/ShrimpPool-0-biannual-1988-
       2016.csv")

shrimppool_8_bia <- read.csv("Datasets/knb-lter-luq.54.945757/ShrimpPool-8-biannual-1988-
       2016.csv")

shrimppool_15_bia <- read.csv("Datasets/knb-lter-luq.54.945757/ShrimpPool-15-biannual-1988-
       2016.csv")



#view data for each pool and check for NA's

#shrimp pool 0

class(shrimppool_0_bia)

head(shrimppool_0_bia)

str(shrimppool_0_bia)

summary(shrimppool_0_bia)

sum(is.na(shrimppool_0_bia))

#shrimp pool 8

class(shrimppool_8_bia)

head(shrimppool_8_bia)

str(shrimppool_8_bia)

summary(shrimppool_8_bia)

sum(is.na(shrimppool_8_bia))

#shrimp pool 15
```

```
class(shrimppool_15_bia)

head(shrimppool_15_bia)

str(shrimppool_15_bia)

summary(shrimppool_15_bia)

sum(is.na(shrimppool_15_bia))




####Join All Biannual Dataframes into 1 Wide Dataset and 1 Long Dataset####

shrimppool_temp_bia = full_join(shrimppool_0_bia, shrimppool_8_bia, by=c("YEAR", "Month", "POOL",
       "ATYACPUE", "XIPHCPUE", "MACCPUE" ), copy=FALSE)

shrimppool_all_bia = full_join(shrimppool_temp_bia, shrimppool_15_bia, by=c("YEAR", "Month",
       "POOL", "ATYACPUE", "XIPHCPUE", "MACCPUE" ), copy=FALSE)

head(shrimppool_all_bia)

str(shrimppool_all_bia)

glimpse(shrimppool_all_bia)

summary(shrimppool_all_bia)

#clean up dataset

#Since POOL is a location description and not a number, let's change its type to factor

shrimppool_all_bia$POOL <- as.factor(shrimppool_all_bia$POOL)

glimpse(shrimppool_all_bia)

#let's make a wide dataset from shrimppool_all_bia that has year and month in one column to use
       for later

W1shrimppool <- shrimppool_all_bia

W1shrimppool$Date <- paste(W1shrimppool$YEAR, W1shrimppool$Month, "1", sep = "-")

W1shrimppool$Date <- ymd(W1shrimppool$Date)

wide_shrimppool <- W1shrimppool%>%

  group_by(Date, POOL, ATYACPUE, XIPHCPUE, MACCPUE)%>%

  select(-YEAR, -Month)

wide_shrimppool
```

```
#let's make a long dataset shrimppool_all_bia
#The species are also factors, let's make 2 columns:
#one for species as a factor varialbles
#one for the count values currently stored in the individual species columns
T_shrimppool<- gather(shrimppool_all_bia,Species,Counts,-YEAR, -Month, -POOL)
head(T_shrimppool)
tail(T_shrimppool)
T_shrimppool$Species <- as.factor(T_shrimppool$Species)
glimpse(T_shrimppool)
#Let's clean up the date. As no sample day was given, we will assume the first of the month
#make a date out of the columns
T_shrimppool$Date <- paste(T_shrimppool$YEAR, T_shrimppool$Month, "1", sep = "-")
glimpse(T_shrimppool)
head(T_shrimppool)
#format the date column
T_shrimppool$Date <- ymd(T_shrimppool$Date)
glimpse(T_shrimppool)
head(T_shrimppool)
#make another table that eliminates the YEAR and Month column
shrimppool_fin <- T_shrimppool %>%
  group_by(Date, POOL, Species)%>%
  select(-YEAR, -Month)
head(shrimppool_fin)
glimpse(shrimppool_fin)

####EDA-Visualizations to graphically understand data####
```

```
shrimppool_fin %>% arrange(shrimppool_fin$Date)

#scatterplot

ggplot(shrimppool_fin, aes(x = shrimppool_fin$Date, y = shrimppool_fin$Counts,
       color=shrimppool_fin$Species))+

  geom_jitter(alpha = 0.6) +

  stat_smooth(method = "lm", se=FALSE, col = "red") +

  scale_y_continuous("Shrimp Species Found Per Trap") +

  scale_x_date("Year of Collection") +

  facet_grid(rows = vars(shrimppool_fin$POOL), cols = vars(shrimppool_fin$Species)) +

  labs(title = "Shrimp Species Found by Date and Pool Location from 1988-2016", col = "Species")

#between the zero counts and low counts Maccpue sp. seems to show no trends. let's plot on log
       scale and alone to see if more info is revealed.

#log plot

ggplot(shrimppool_fin, aes(x = shrimppool_fin$Date, y = shrimppool_fin$Counts,
       color=shrimppool_fin$Species))+

  geom_jitter(alpha = 0.6) +

  scale_y_log10("Shrimp Species Found Per Trap (log)") +

  scale_x_date("Year of Collection") +

  facet_grid(rows = vars(shrimppool_fin$POOL), cols = vars(shrimppool_fin$Species)) +

  labs(title = "Shrimp Species Found by Date and Pool Location from 1988-2016", col = "Species")



#Maccpue sp. by date and location

maccpuectplot <- shrimppool_fin %>%

  filter(Species == "MACCPUE") %>%

  ggplot(aes(x = Date, y = Counts, color= POOL)) +

    geom_jitter(alpha = 0.6) +

    stat_smooth() +

    labs(x="Year of Collection", y="Maccpue sp. Found Per Trap", title = "Maccpue sp. Found by
       Pool Location from 1988-2016", col = "Pool")
```

```
maccpuectplot



ggpairs(data = shrimppool_fin, columns = 1:4)

#The only evidence of a poor correlation is between the number of date and the number of counts.

#Corr 0.135 particularly in the late 1990s and 2010s

#looks like the collection dates are inconsistent per pool and species. ideally, this should be
       where i create a series of loops to compute the average counts for each species
       biannually.

#i will come back to that; however, in the interest of time for this assignment, i will leave the
       data as is.



#we want to see if there's correlation between species over time so let's look at three more
       plots:

#calculate means of each species with all pools assummed equal

mean_Asp <- mean(wide_shrimppool$ATYACPUE)

mean_Asp #38.69164

mean_Xsp <- mean(wide_shrimppool$XIPHCPUE)

mean_Xsp #22.24613

mean_Msp <- mean(wide_shrimppool$MACCPUE)

mean_Msp #0.1786804

#calculate means of each species by pool

#Pool 0 by species

mean_bia0A <- shrimppool_all_bia %>%

  filter(POOL == '0') %>%

  summarize(mean0A = mean(shrimppool_all_bia$ATYACPUE), mean0X =
       mean(shrimppool_all_bia$XIPHCPUE), mean0M = mean(shrimppool_all_bia$MACCPUE))

#Pool 8 by species

mean_bia8A <- shrimppool_all_bia %>%

  filter(POOL == '8') %>%

  summarize(mean8A = mean(shrimppool_all_bia$ATYACPUE),mean8X =
       mean(shrimppool_all_bia$XIPHCPUE), mean8M = mean(shrimppool_all_bia$MACCPUE))
```

```
#Pool 15 by species

mean_bia15A <- shrimppool_all_bia %>%

  filter(POOL == '15') %>%

  summarize(totmean15A = mean(shrimppool_all_bia$ATYACPUE), mean15X =
       mean(shrimppool_all_bia$XIPHCPUE), mean15M = mean(shrimppool_all_bia$MACCPUE))



#use the wide dataset wide_shrimppool to compare ATYACPUE-XIPHCPUE, ATYACPUE-MACCPUE, XIPHCPUE-
       MACCPUE across all pools

#as ATYACPUE had the largest average observations per pool, we will assume it to be the
       independent variable (x) in all species comparisons.

#as MACCPUE had the smallest average observations per pool, we will assume it to be the dependent
       variable (y) in all species comparisons.

#as the average observations per pool for XIPHCPUE were <ATYACPUE and >MACCPUE, we will assume it
       to be the dependent variable (x) in comparison to ATYACPUE,

#and the independent variable (x) in comparison to MACCPUE

#x=ATYACPUE-y=XIPHCPUE

ggplot(wide_shrimppool, aes(x=wide_shrimppool$ATYACPUE, y=wide_shrimppool$XIPHCPUE,
       color=wide_shrimppool$POOL))+

  geom_point(alpha=0.6) +

  stat_smooth(method = "lm", se=FALSE, col = "red") +

  scale_y_continuous("Average Populations of Xiphocaris elongata") +

  scale_x_continuous("Average Populations of Atya lanipes") +

  facet_grid(wide_shrimppool$POOL) +

  labs(title = "Average Populations of Atya lanipes to Xiphocaris elongata, 1988-2016",
       color="Pool")

#x=ATYACPUE-y=MACCPUE

ggplot(wide_shrimppool, aes(x=wide_shrimppool$ATYACPUE, y=wide_shrimppool$MACCPUE,
       color=wide_shrimppool$POOL))+

  geom_point(alpha=0.6) +

  stat_smooth(method = "lm", se=FALSE, col = "red") +

  scale_y_continuous("Average Populations of Macrobrachium spp") +

  scale_x_continuous("Average Populations of Atya lanipes") +
  facet_grid(wide_shrimppool$POOL) +

  labs(title = "Average Populations of Atya lanipes to Macrobrachium spp, 1988-2016",
       color="Pool")

#x=XIPHCPUE-y=MACCPUE

ggplot(wide_shrimppool, aes(x=wide_shrimppool$XIPHCPUE, y=wide_shrimppool$MACCPUE,
       color=wide_shrimppool$POOL))+

  geom_point(alpha=0.6) +

  stat_smooth(method = "lm", se=FALSE, col = "red") +

  scale_y_continuous("Average Populations of Macrobrachium spp") +

  scale_x_continuous("Average Populations of Xiphocaris elongata") +

  facet_grid(wide_shrimppool$POOL) +

  labs(title = "Average Populations of Xiphocaris elongata to Macrobrachium spp, 1988-2016",
       color="Pool")



head(wide_shrimppool)

#ggpairs(data = wide_shrimppool, columns = 1:5)

ggpairs(data = wide_shrimppool, columns = 2:5)



#standard deviation between species

by (wide_shrimppool$ATYACPUE, wide_shrimppool$XIPHCPUE, wide_shrimppool$MACCPUE, sd)

#standard deviation between species and time

by (shrimppool_fin$Date, shrimppool_fin$Counts, sd)



####Linear Models####

#lm(y~x,data), where y is the dependent variable, x is the independent variable

#create a unified wide dataframe with all original data, all explanatory variables, and residuals

#use null model to find out how well our model performed



##ALL SPECIES TO TIME##
```

```
lm_spyr <- lm(Counts ~ Date, data=shrimppool_fin)

lm_spyr #intercept=9.117668; Date= 0.001019

summary (lm_spyr)

coef(lm_spyr)

#vector with all the fitted values (y'), which will tell us what the model predicted

fitted_mx <- fitted.values(lm_spyr)

#residuals from fitted values, which tells the difference between the actual, measured value and
       the predicted (fitted) values

res_spyr <- residuals(lm_spyr)

#Residuals:

#Min      1Q  Median      3Q     Max

#-26.555 -17.724  -5.415  10.595 131.431

#Coefficients:

#  Estimate Std. Error t value Pr(>|t|)

#(Intercept) 9.117668   2.695591   3.382 0.000746 ***

#  Date        0.001019   0.000235   4.338 1.58e-05 ***

#  ---

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

#Residual standard error: 23.43 on 1021 degrees of freedom

#Multiple R-squared:  0.0181, Adjusted R-squared:  0.01714

#F-statistic: 18.82 on 1 and 1021 DF,  p-value: 1.577e-05

#unified dataframe for all species to year

lm_spyr_un <- broom::augment(lm_spyr)

head(lm_spyr_un)

lm_spyr_un %>%

  arrange(desc(.resid))%>%

  head() %>%
  tail()
lm_spyr_un$.resid_abs<-abs(lm_spyr_un$.resid_abs)
lm_spyr_un %>%
  arrange(desc(.resid_abs)) %>%
  head()
#inspect outlier Date==1999-06-01
shrimppool_fin%>%
  filter(Date == '1999-06-01') #152 Atyacpue species in Pool 15. As the other species were only counted
#at 13.5 and 0 in Pool 15 on that date, a question for further examination would be what is causing Atya sp to proliferate.
#create null model
spyr_null <- lm(Counts ~1, data = shrimppool_fin)
spyr_null #intercept = 20.37
#verify null model
mean_spcount <- mean(shrimppool_fin$Counts)
mean_spcount #20.37215
ggplot(data = shrimppool_fin, aes(x=Date, y=Counts))+
  geom_point(alpha=0.6)+
  geom_hline(yintercept = mean_spcount) +
  labs(title = "Species Count Null Model")
ggplot(data = shrimppool_fin, aes(x=Date, y=Counts))+
  geom_point(alpha=0.6)+
  stat_smooth(method = "lm")+

  geom_hline(yintercept = mean_spcount) +

  labs(title = "Species Count Null Model")

#assess error = Multiple R-squared

summary(lm_spyr) #0.0181 which is not good as most points are
     outside the the standard error so the model cannot
     accurately predict the amount of species related to year




##XIPHCPUE-ATYACPUE##

lm_xa <- lm(XIPHCPUE ~ ATYACPUE, data=wide_shrimppool)

lm_xa #intercept= 13.268; ATYACPUE=0.232

summary (lm_xa)

coef(lm_xa)

#vector with all the fitted values (y'), which will tell us what
     the model predicted

fitted_xa <- fitted.values(lm_xa)

#residuals from fitted values, which tells the difference between
     the actual, measured value and the predicted (fitted) values

res_xa <- residuals(lm_xa)

#Residuals:

#  Min      1Q  Median      3Q     Max

#-34.921 -12.015  -3.749   8.552  85.810

#Coefficients:
```

```
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept)  13.2685     1.7237   7.697 1.53e-13 ***
#  ATYACPUE      0.2320     0.0376   6.171 1.93e-09 ***
#  ---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#Residual standard error: 17.08 on 339 degrees of freedom
#Multiple R-squared:  0.101,  Adjusted R-squared:  0.09835
#F-statistic: 38.09 on 1 and 339 DF,  p-value: 1.931e-09
#unified dataframe for all species to year
lm_xa_un <- broom::augment(lm_xa)
head(lm_xa_un)
lm_xa_un %>%
  arrange(desc(.resid))%>%
  head() %>%
  tail()
#inspect outlier XIPHCPUE==113 & ATYACPUE==60
wide_shrimppool %>%
  filter(XIPHCPUE==113 & ATYACPUE==60) #This outlier occured on 2015-11-01. The  XIPHCPUE==113 > ATYACPUE==60
#Contrary to our calculated mean  XIPHCPUE==113 > ATYACPUE==60. What caused this change in population dynamic?
#create null model
xa_null <- lm(XIPHCPUE~1, data = wide_shrimppool)
xa_null #intercept = 22.25

#verify null model

mean_Xsp

ggplot(data = wide_shrimppool, aes(x=ATYACPUE, y=XIPHCPUE))+

  geom_point(alpha=0.6)+

  geom_hline(yintercept = mean_Xsp) +

  labs(title = "Xiphocaris elongate Species Count to Null Model")

ggplot(data = wide_shrimppool, aes(x=ATYACPUE, y=XIPHCPUE))+

  geom_point(alpha=0.6)+

  stat_smooth(method = "lm")+

  geom_hline(yintercept = mean_Xsp) +

  labs(title = "Xiphocaris elongate Species Count to Null Model")

#assess error = Multiple R-squared

summary(lm_xa) #0.0181 which is not good as most points are
     outside the the standard error so the model cannot
     accurately predict the amount of species related to year

##MACCPUE-ATYACPUE##

lm_ma <- lm(MACCPUE~ATYACPUE, data=wide_shrimppool)

lm_ma #intercept=0.207005614; ATYACPUE=-0.000732077

summary (lm_ma)

coef(lm_ma)

#vector with all the fitted values (y'), which will tell us what
     the model predicted

fitted_ma <- fitted.values(lm_ma)
```

```
#residuals from fitted values, which tells the difference between
     the actual, measured value and the predicted (fitted) values

res_ma <- residuals(lm_ma)

#Residuals:

#Min      1Q  Median      3Q     Max

#-0.2070 -0.1831 -0.1382  0.1128  1.5113

#Coefficients:

#  Estimate Std. Error t value Pr(>|t|)

#(Intercept)  0.2070056  0.0291442   7.103 7.23e-12 ***

#  ATYACPUE    -0.0007321  0.0006357  -1.152     0.25

#---

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

#Residual standard error: 0.2887 on 339 degrees of freedom

#Multiple R-squared:  0.003897,    Adjusted R-squared:  0.0009587

#F-statistic: 1.326 on 1 and 339 DF,  p-value: 0.2503

lm_ma_un <- broom::augment(lm_ma)

head(lm_ma_un)

lm_ma_un %>%

  arrange(desc(.resid))%>%

  tail()

#inspect outlier MACCPUE==1.7 & ATYACPUE==25

wide_shrimppool %>%
  filter(MACCPUE==1.7 & ATYACPUE==25) #This outlier occured on 1996-06-01.What caused this small count in Atyacpue in Pool 8?

#inspect outlier MACCPUE==0 & ATYACPUE==6

wide_shrimppool %>%

  filter(MACCPUE==0 & ATYACPUE==6) #This outlier occured on 1989-12-01. What caused this small count for all species in Pool 8?

#create null model

ma_null <- lm(MACCPUE~1, data = wide_shrimppool)

ma_null #intercept = 0.1787

#verify null model

mean_Msp

ggplot(data = wide_shrimppool, aes(x=ATYACPUE, y=MACCPUE))+

  geom_point(alpha=0.6)+

  geom_hline(yintercept = mean_Msp) +

  labs(title = "Maccrobrachium spp Species Count to Null Model")

ggplot(data = wide_shrimppool, aes(x=ATYACPUE, y=MACCPUE))+

  geom_point(alpha=0.6)+

  stat_smooth(method = "lm")+

  geom_hline(yintercept = mean_Msp) +

  labs(title = "Maccrobrachium Species Count to Null Model")

#assess error = Multiple R-squared

summary(lm_ma) #0.029  the species are not correlated
```

```
##MACCPUE-XIPHCPUE##

lm_mx <- lm(MACCPUE~XIPHCPUE, data=wide_shrimppool)

lm_mx #intercept=0.2141961; XIPHCPUE=-0.0015965

summary (lm_mx)

coef(lm_mx)

#vector with all the fitted values (y'), which will tell us what
     the model predicted

fitted_mx <- fitted.values(lm_mx)

#residuals from fitted values, which tells the difference between
     the actual, measured value and the predicted (fitted) values

res_mx <- residuals(lm_mx)

#Residuals:

#  Min      1Q  Median      3Q     Max

#-0.2142 -0.1914 -0.1299  0.1090  1.5013

#Coefficients:

#  Estimate Std. Error t value Pr(>|t|)

#(Intercept)  0.2141961  0.0248170   8.631  2.4e-16 ***

#  XIPHCPUE    -0.0015965  0.0008681  -1.839   0.0668 .

#---

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

#Residual standard error: 0.2878 on 339 degrees of freedom
```

```
#Multiple R-squared:  0.009879,    Adjusted R-squared:  0.006958

#F-statistic: 3.382 on 1 and 339 DF,  p-value: 0.06677

lm_mx_un <- broom::augment(lm_mx)

head(lm_mx_un)

lm_mx_un %>%

  arrange(desc(.resid))%>%

  head() %>%

  tail()

#inspect outlier MACCPUE==0.31 & Xiphcpue==8.25

wide_shrimppool %>%

  filter(MACCPUE==0.31 & XIPHCPUE==8.25) #This outlier occured on
     1988-01-01 in Pool0.What caused the counts of A&X to be
     nearly equal?

#inspect outlier MACCPUE==1.7 & Xiphcpue==9.7

wide_shrimppool %>%

  filter(MACCPUE==1.7 & XIPHCPUE==9.7) #This outlier occured in
     Pool 8 on  1996-06-01.

#create null model

mx_null <- lm(MACCPUE~1, data = wide_shrimppool)

mx_null #intercept = 0.1787

#verify null model

mean_Msp

ggplot(data = wide_shrimppool, aes(x=XIPHCPUE, y=MACCPUE))+

  geom_point(alpha=0.6)+
  geom_hline(yintercept = mean_Msp) +
  labs(title = "Maccrobrachium spp Species Count to Null Model")
ggplot(data = wide_shrimppool, aes(x=XIPHCPUE, y=MACCPUE))+
  geom_point(alpha=0.6)+
  stat_smooth(method = "lm")+
  geom_hline(yintercept = mean_Msp) +
  labs(title = "Maccrobrachium Species Count to Null Model")
#assess error = Multiple R-squared
summary(lm_mx) #0.009879 which is a horizontal line indicating the species are not correlated
```

Bibliography

ABOUT US | Luquillo LTER. (n.d.). Retrieved October 25, 2019, from https://luq.lter.network/about

Carrington, D. (2018, March 12). What is biodiversity and why does it matter to us? *The Guardian*. Retrieved from https://www.theguardian.com/news/2018/mar/12/what-is-biodiversity-and-why-does-it-matter-to-us.

Crowl, T. (2017). *Shrimp populations in Quebrada Prieta (Pools 0, 8, 9, 15) (El Verde)* [Data set]. https://doi.org/10.6073/PASTA/F6C8497C780ECF619053DCD020D371F2.

Recommended Data Repositories | Scientific Data. (n.d.). Retrieved October 25, 2019, from https://www.nature.com/sdata/policies/repositories

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18