Introduction

Species richness is an important matrix for ecologists to better understand a community structure (Hewitt et al., 2016). However, many different ways to estimate species richness exist. In this study, my aim is to evaluate different methodologies to estimate species richness. I will be estimating species richness based on latitudinal bands to capture the latitudinal biodiversity gradient in each of my calculation groups. The latitudinal biodiversity gradient is one of the most prolific and widespread phenomena in biodiversity studies and thus will give me a good benchmark in comparing the different methodologies (Lawrence & Fraser, 2020). I will be using the specnumber function in the vegan package and the chao1 function in the fossil package as well as use taxonomic identification and bin identification to generate 4 different methodologies to calculate species richness: 1) Taxonomic identification and the specnumber function, 2) Taxonomic identification and the chao1 function, 3) Bin identification and the specnumber function, 4) Bin identification and the chao1 function. I will be gathering my species data using the BOLD database. I will use two large taxonomic groups that are widespread globally, the order *Avis* and the family *Nymphalidae.* I will subdivide the species richness calculations data into 6 bands of latitude to capture data pertaining to the northern hemisphere. If the different methodologies generate similar results for species richness estimation, then I would expect similar estimates of species richness in each latitudinal band. This will demonstrate that each calculation is capturing the gradient in similar ways. If the different methodologies generated different results for species richness estimates, then I would expect different estimates of species richness in each latitudinal band. This will demonstrate that each calculation is capturing the gradient in various ways.

Results and Discussion

The results from this study indicate that the methodology for estimating species richness can have an impact on the gradient. As you can see from figure 3 and figure 4, for both *Aves* and *Nymphalidae,* there is great variation in estimated species richness at the latitudinal band 0-15. However, with increasing latitude, the variation in estimation is mitigated. The great variation at the 0-15 latitudinal band could be attributed to the amount of singletons and doubletons in these bands as there is a greater overall number of unique species at these lower latitudes compared to the higher latitudes. Therefore, we see the chao1 estimates higher than the specnumber estimates due to chao1 making predictions about rare and undocumented species. It should be noted however, we see species richness using the specnumber function for taxonomic id in *Nymphalidae* that is comparable to the chao1 groups. This could be attributed to the high degree of taxonomic sampling of *Nymphalidae* in this region. This study demonstrated the importance for scientists to ensure they consider the proper methodology for estimating species richness that best fits their project when investigating biodiversity. It also highlights the need for scientists to consider geographical sampling biases when doing an analysis of species richness in a global context. As you can see from figures 1 and 2, there is a great bias in sampling in North America and Europe compared to the rest of the Northern Hemisphere. Future efforts need to be made to gather more data on species richness in under sampled regions of the globe. This will ensure a more accurate estimate of species richness which will serve to benefit conservation efforts globally. Future directions from this study can include investigating the Southern Hemisphere of the Earth to get a better picture of species richness estimates and sampling efforts in that region of the globe.

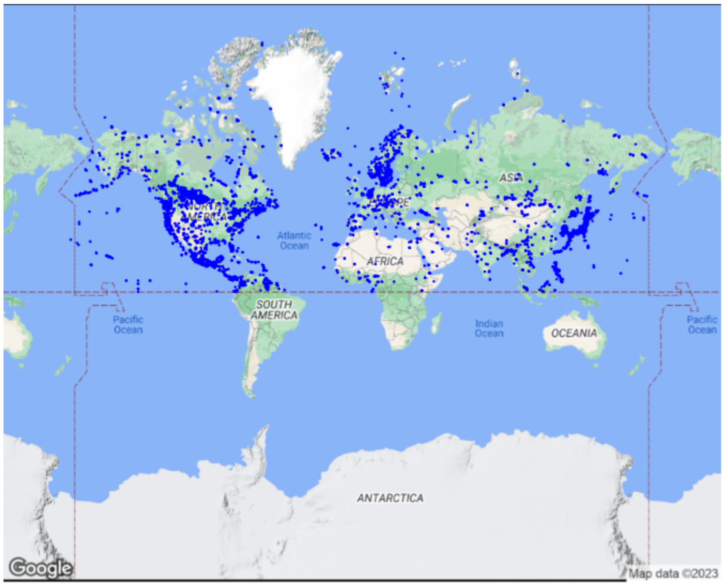

Figure 1. Sampling records of *Aves* in the Northern Hemisphere. Data was collected from the BOLD database. Each point represents a sample ID.

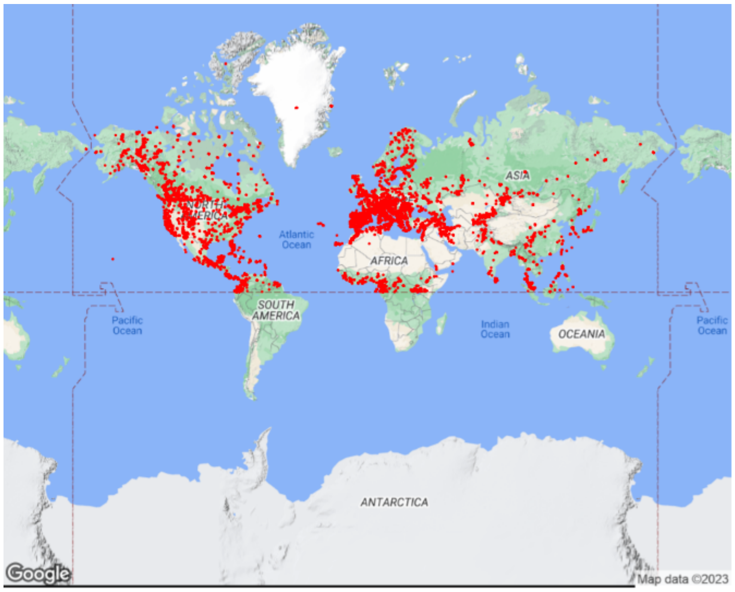

Figure 2. Sampling records of *Nymphalidae* in the Northern Hemisphere. Data was collected from the BOLD database. Each point represents a sample ID.

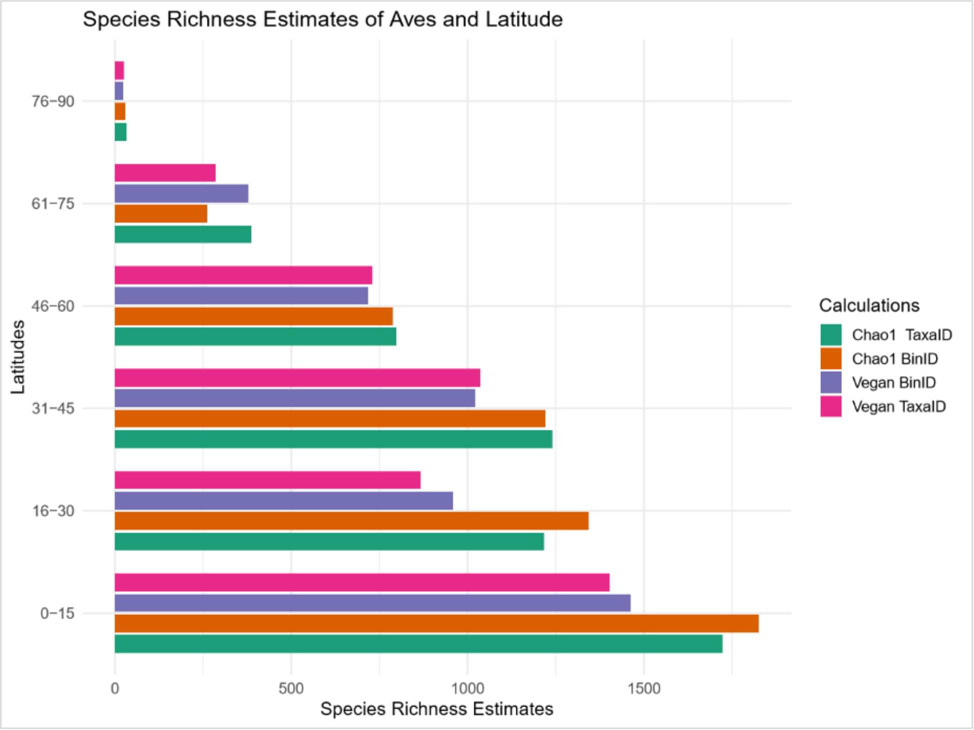

Figure 3. Species richness estimates of *Aves* at various latitudinal bands in the Northern hemisphere.

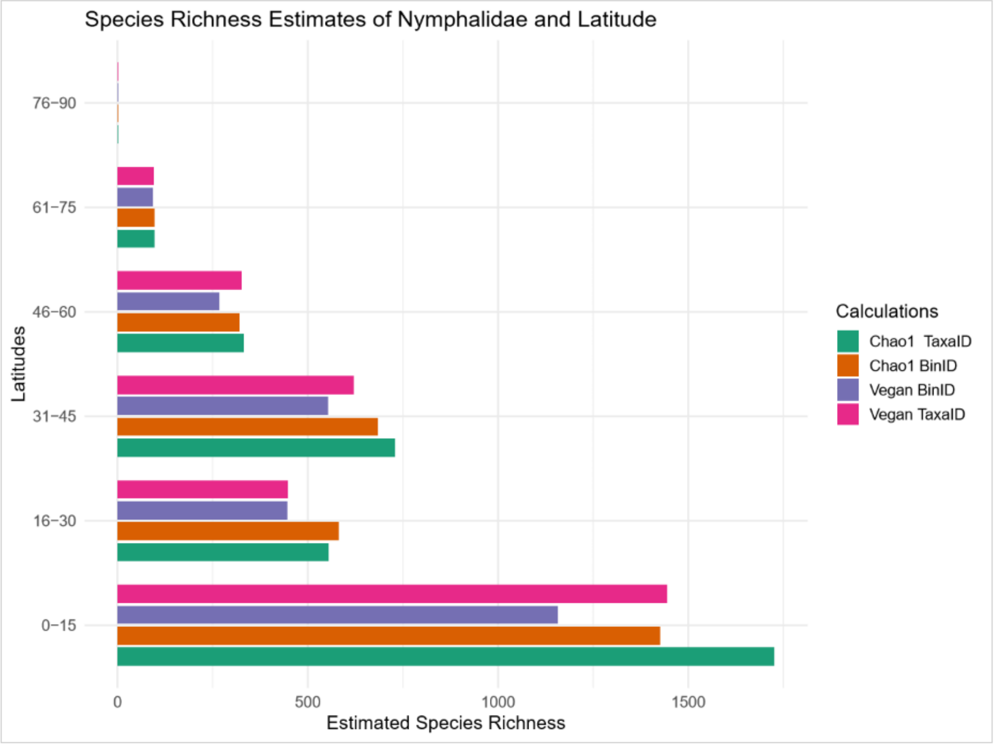

Figure 4. Species richness estimates of *Nymphalidae* at various latitudinal bands in the Northern hemisphere.

Code Documentation

#1.load the necessary libraries----

```
library("tidyverse")
library("vegan")
library("fossil")
library("ggmap")
```

#2. Load in Aves tsv----

#read in the aves data frame from BOLD

```
df_a <-
read_tsv('"http://www.boldsystems.org/index.php/API_Public/combined?taxon=Aves&format=tsv')
```

#3.Number of samples based on latitude----

#The function serves as a tool for exploratory analysis of my data. It selects sampleid and lat columns from df_a and filters the desired latitudinal range. It then groups the data by latitude and counts the number of samples at each latitude group. It then pulls the count column and then calculates the sum of this column to get the number of samples made of the taxa at the inputted latitudinal band.

```
num_of_samples_aves <- function(minimum_latitude, max_latitude) {
  sample_count <- df_a %>%
    select(sampleid, lat) %>%
    filter(!is.na(lat) & lat >= minimum_latitude & lat < max_latitude) %>%
    group_by(lat) %>%
    summarise(count = n()) %>%
    pull(count) %>%
    sum() %>%
    return(sample_count) }

aves_sampling_effort <- c(num_of_samples_aves(0, 15), num_of_samples_aves(16, 30),
num_of_samples_aves(31, 45), num_of_samples_aves(46, 60), num_of_samples_aves(61, 75),
num_of_samples_aves(76, 90))

latitude <- c('0-15', '16-30', '31-45', '46-60', '61-75', '76-90')

aves_sampling <- data.frame(latitude, aves_sampling_effort)
```

#4. Creating a map of sampling efforts for Aves----

#Sample data frame with latitude and longitude

```
df_a_map <- df_a %>%
  select(sampleid, lat, lon) %>%
  filter(!is.na(lat), !is.na(lon)) %>%
  filter(lat >= 0)

#Inputting API key
register_google(key = "AIzaSyBxWwRVQigeDNM_MhkZlR2fhXHN8e8Vqbk")

#create map plot of sampling efforts for Aves
map_aves <- ggmap(get_googlemap(center = c(lon = 0, lat = 40), zoom = 1, maptype = 'terrain', color = 'color')) +
  geom_point(data = df_a_map, aes(x=lon, y=lat), color = 'blue', size = 0.1) +
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
  )

print(map_aves)
```

#5.Function that estimates species richness using vegan and taxonomic identification ----

#Here, I create a function that will select species_name and lat and filter out any na's. It will then filter the desired latitudinal range of the user. Next, it will then group by species_name and count the number of species_name groups. It will then reformat the data so that it can be used for the specnumber() function in the vegan package. The function will then return the estimated species richness.

```
speciesrichness_name_f <- function(min_lat, max_lat) {
  c_a <- df_a %>%
    select(species_name, lat) %>%
    filter(!is.na(species_name), !is.na(lat)) %>%
    filter(lat >= min_lat & lat < max_lat) %>%
    group_by(species_name) %>%
    count(species_name) %>%
    pivot_wider(names_from = species_name, values_from = n) %>%
    vegan::specnumber() %>%
    return(c_a)
}
#calling the function with my latitudinal bands as input
name_vegan_015 <- speciesrichness_name_f(0, 15)
name_vegan_1630 <- speciesrichness_name_f(16, 30)
name_vegan_3145 <- speciesrichness_name_f(31, 45)
name_vegan_4660 <- speciesrichness_name_f(46, 60)
name_vegan_6175 <- speciesrichness_name_f(61, 75)
```

```
name_vegan_7690 <- speciesrichness_name_f(76, 90)
```

#6.Function that estimates species richness using chao1 and bin identification----

#Here, I create a function that will select bin_uri and lat and filter out any NA's. It will then filter the desired latitudinal range of the user. Next, it will then group by bin_uri and count the number of bin_uri groups. It will then reformat the data so that it can be used for the chao1 function in the fossil package. The function will then return the estimated species richness.

```
chao1f <- function(min_lat, max_lat) {
  c_a <- df_a %>%
    select(bin_uri, lat) %>%
    filter(!is.na(bin_uri), !is.na(lat)) %>%
    filter(lat >= min_lat & lat < max_lat) %>%
    group_by(bin_uri) %>%
    count(bin_uri) %>%
    pivot_wider(names_from = bin_uri, values_from = n) %>%
    fossil::chao1(taxa.row = TRUE) %>%
    return(c_a)
  }
```

#calling the function with my latitudinal bands as input

```
chao1f_015 <- chao1f(0, 15)
chao1f_1630 <- chao1f(16, 30)
chao1f_3145 <- chao1f(31, 45)
chao1f_4660 <- chao1f(46, 60)
chao1f_6175 <- chao1f(61, 75)
chao1f_7690 <- chao1f(76, 90)
```

#7. Function that estimates species richness using vegan and bin identification----

#Here, I create a function that will select bin_uri and lat and filter out any NA's. It will then filter the desired latitudinal range of the user. Next, it will then group by bin_uri and count the number of bin_uri groups. It will then reformat the data so that it can be used for the specnumber() function in the vegan package. The function will then return the estimated species richness.

```
bin_speciesrichness_f <- function(min_lat, max_lat) {
  c_a <- df_a %>%
    select(bin_uri, lat) %>%
    filter(!is.na(bin_uri), !is.na(lat)) %>%
    filter(lat >= min_lat & lat < max_lat) %>%
    group_by(bin_uri) %>%
    count(bin_uri) %>%
    pivot_wider(names_from = bin_uri, values_from = n) %>%
    vegan::specnumber() %>%
    return(c_a)
}
#calling the function with latitudinal bands as inputs
bin_vegan_015 <- bin_speciesrichness_f(0, 15)
bin_vegan_1630 <- bin_speciesrichness_f(16, 30)
bin_vegan_3145 <- bin_speciesrichness_f(31, 45)
bin_vegan_4660 <- bin_speciesrichness_f(46, 60)
bin_vegan_6175 <- bin_speciesrichness_f(61, 75)
bin_vegan_7690 <- bin_speciesrichness_f(76, 90)
```

#8. Function that estimates species richness using Chao1 and taxonomic identification----

#Here, I create a function that will select species_name and lat and filter out any na's. It will then filter the desired latitudinal range of the user. Next, it will then group by species_name and count the number of species_name groups. It will then reformat the data so that it can be used for the chao1 function in the fossil package. The function will then return the estimated species richness.

```
chao1_name_f <- function(min_lat, max_lat) {
  c_a <- df_a %>%
    select(species_name, lat) %>%
    filter(!is.na(species_name), !is.na(lat)) %>%
    filter(lat >= min_lat & lat < max_lat) %>%
    group_by(species_name) %>%
    count(species_name) %>%
    pivot_wider(names_from = species_name, values_from = n) %>%
    fossil::chao1(taxa.row = TRUE) %>%
    return(c_a)
}

#calling function with latitudinal bands as inputs
chao1_namef_015 <- chao1_name_f(0, 15)
chao1_namef_1630 <- chao1_name_f(16, 30)
chao1_namef_3145 <- chao1_name_f(31, 45)
chao1_namef_4660 <- chao1_name_f(46, 60)
chao1_namef_6175 <- chao1_name_f(61, 75)
chao1_namef_7690 <- chao1_name_f(76, 90)
```

#9.creating a bar plot to visualize the results across the calculations.----

```
#creating a data frame to organize my results across the calculations
calculation_groups <- c("Chao1  TaxaID", "Vegan TaxaID", "Chao1 BinID", "Vegan BinID")
```

```
species_richness_aves <- c(chao1_namef_015, name_vegan_015,chao1f_015,bin_vegan_015,
chao1_namef_1630,name_vegan_1630,chao1f_1630, bin_vegan_1630,
chao1_namef_3145,name_vegan_3145,chao1f_3145, bin_vegan_3145, chao1_namef_4660,
name_vegan_4660, chao1f_4660,bin_vegan_4660,
chao1_namef_6175,name_vegan_6175,bin_vegan_6175, chao1f_6175,
chao1_namef_7690,name_vegan_7690, chao1f_7690, bin_vegan_7690)
```

#creating the data frame

```
df_aves_species_richness_calculations <- data.frame(latitude, calculation_groups, species_richness_aves)
```

#creating a bar plot to visualize the results across the calculations. I am using ggplot to create a grouped bar chart of my data so that I can visualize each species richness estimate at each of the latitudinal bands. I am ensuring the bar plot colours are colour blind friendly by using 'Dark2' palette.

```
aves_plot <- df_aves_species_richness_calculations %>%
  ggplot(aes(x = latitudes, y = species_richness_aves, fill = calculation_groups))+
  geom_bar(stat = 'identity', position = position_dodge(width = 0.8), width = 0.7)+
  labs(
    title = 'Species Richness Estimates of Aves and Latitude',
    x = 'Latitudes',
    y = 'Species Richness Estimates',
    fill = 'Calculations') +
  scale_fill_brewer(palette = 'Dark2') +
  theme_minimal() +
  theme(
    legend.key.size = unit(0.5, 'cm'),
    legend.text = element_text(size = 9)) +
  coord_flip()

aves_plot
```

#1.Load the necessary libraries----

```
library("tidyverse")
library("vegan")
library("fossil")
library("ggplot2")
library("ggmap")
```

#2.Load in the Nymphalidae tsv----

#read in the Nymphalidae data frame from BOLD

```
df_n <-
read_tsv('http://www.boldsystems.org/index.php/API_Public/combined?taxon=Nymphalidae&format=tsv'
)
```

#3.Looking at the number of samples at each latitudinal range----

```
#The function serves as a tool for exploratory analysis of my data. It selects sampleid and lat columns from df_n and filters the desired latitudinal range. It then groups the data by latitude and counts the number of samples at each latitude group. It then pulls the count column and then calculates the sum of this column to get the number of samples made of the taxa at the inputted latitudinal band
num_of_samples_nymph <- function(minimum_latitude, max_latitude) {
  sample_count <- df_n %>%
    select(sampleid, lat) %>%
    filter(!is.na(lat) & lat >= minimum_latitude & lat < max_latitude) %>%
    group_by(lat) %>%
    summarise(count = n()) %>%
    pull(count) %>%
    sum()
  return(sample_count) }

#creating a vector of function calls to be used in the nymph_sampling data frame.
nymph_sampling_effort <- c(num_of_samples_nymph(0, 15), num_of_samples_nymph(16, 30), num_of_samples_nymph(31, 45), num_of_samples_nymph(46, 60), num_of_samples_nymph(61, 75), num_of_samples_nymph(76, 90))

#creating a vector containing my latitudinal bands to be used in the nymph_sampling data frame.
latitude <- c('0-15', '16-30', '31-45', '46-60', '61-75', '76-90')

#creating nymph_sampling data frame from latitude and nymph_sampling_effort
nymph_sampling <- data.frame(latitude, nymph_sampling_effort)
```

#4.Creating a map of sampling efforts----

```
#Sample data frame with latitude and longitude
df_n_map <- df_n %>%
  select(sampleid, lat, lon) %>%
  filter(!is.na(lat), !is.na(lon)) %>%
  filter(lat >= 0)
#Inputting API key
register_google(key = "AIzaSyBxWwRVQigeDNM_MhkZlR2fhXHN8e8Vqbk")

#create map plot
map_nymph <- ggmap(get_googlemap(center = c(lon = 0, lat = 40), zoom = 1, maptype = 'terrain', color = 'color')) +
  geom_point(data = df_n_map, aes(x=lon, y=lat), color = 'red', size = 0.1) +
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
  )
```

```
print(map_nymph)
```

#5.Creating a function to calculate species richness using taxonomic id and vegan----

#Here, I create a function that will select species_name and lat and filter out any na's. It will then filter the desired latitudinal range of the user. Next, it will then group by species_name and count the number of species_name groups. It will then reformat the data so that it can be used for the specnumber() function in the vegan package. The function will then return the estimated species richness.

```
speciesrichness_taxid_vegan_nymph <- function(min_lat, max_lat) {
  c_n <- df_n %>%
    select(species_name, lat) %>%
    filter(!is.na(species_name), !is.na(lat)) %>%
    filter(lat >= min_lat & lat < max_lat) %>%
    group_by(species_name) %>%
    count(species_name) %>%
    pivot_wider(names_from = species_name, values_from = n) %>%
    vegan::specnumber() %>%
    return(c_n)
}
```

#calling the function based on my desired latitudinal ranges

```
n_taxid_vegan_015 <- speciesrichness_taxid_vegan_nymph(0, 15)
n_taxid_vegan_1630 <- speciesrichness_taxid_vegan_nymph(16, 30)
n_taxid_vegan_3145 <- speciesrichness_taxid_vegan_nymph(31, 45)
n_taxid_vegan_4660 <- speciesrichness_taxid_vegan_nymph(46, 60)
n_taxid_vegan_6175 <- speciesrichness_taxid_vegan_nymph(61, 75)
n_taxid_vegan_7690 <- speciesrichness_taxid_vegan_nymph(76, 90)
```

#6.Creating a function to calculate species richness using bin id and vegan----

#Here, I create a function that will select bin_uri and lat and filter out any NA's. It will then filter the desired latitudinal range of the user. Next, it will then group by bin_uri and count the number of bin_uri groups. It will then reformat the data so that it can be used for the specnumber() function in the vegan package. The function will then return the estimated species richness.

```
speciesrichness_binid_vegan_nymph <- function(min_lat, max_lat) {
  c_n <- df_n %>%
    select(bin_uri, lat) %>%
    filter(!is.na(bin_uri), !is.na(lat)) %>%
    filter(lat >= min_lat & lat < max_lat) %>%
    group_by(bin_uri) %>%
    count(bin_uri) %>%
    pivot_wider(names_from = bin_uri, values_from = n) %>%
    vegan::specnumber() %>%
    return(c_n)
}
```

#calling the function based on my desired latitudinal ranges

```
n_binid_vegan_015 <- speciesrichness_binid_vegan_nymph(0, 15)
n_binid_vegan_1630 <- speciesrichness_binid_vegan_nymph(16, 30)
n_binid_vegan_3145 <- speciesrichness_binid_vegan_nymph(31, 45)
n_binid_vegan_4660 <- speciesrichness_binid_vegan_nymph(46, 60)
n_binid_vegan_6175 <- speciesrichness_binid_vegan_nymph(61, 75)
n_binid_vegan_7690 <- speciesrichness_binid_vegan_nymph(76, 90)
```

#7.Creating a function to calculate species richness using taxonomic id and chao1----

#Here, I create a function that will select species_name and lat and filter out any na's. It will then filter the desired latitudinal range of the user. Next, it will then group by species_name and count the number of species_name groups. It will then reformat the data so that it can be used for the chao1 function in the fossil package. The function will then return the estimated species richness.

```
speciesrichness_taxid_chao1_nymph <- function(min_lat, max_lat) {
  c_n <- df_n %>%
    select(species_name, lat) %>%
    filter(!is.na(species_name), !is.na(lat)) %>%
    filter(lat >= min_lat & lat < max_lat) %>%
    group_by(species_name) %>%
    count(species_name) %>%
    pivot_wider(names_from = species_name, values_from = n) %>%
    fossil::chao1(taxa.row = TRUE) %>%
    return(c_n)
}
```

#calling function with latitudinal bands as inputs

```
n_taxid_chao1_015 <- speciesrichness_taxid_chao1_nymph(0, 15)
n_taxid_chao1_1630 <- speciesrichness_taxid_chao1_nymph(16, 30)
n_taxid_chao1_3145 <- speciesrichness_taxid_chao1_nymph(31, 45)
n_taxid_chao1_4660 <- speciesrichness_taxid_chao1_nymph(46, 60)
n_taxid_chao1_6175 <- speciesrichness_taxid_chao1_nymph(61, 75)
n_taxid_chao1_7690 <- speciesrichness_taxid_chao1_nymph(76, 90)
```

#8.Creating a function to calculate species richness using bin id and chao1----

#Here, I create a function that will select bin_uri and lat and filter out any NA's. It will then filter the desired latitudinal range of the user. Next, it will then group by bin_uri and count the number of bin_uri groups. It will then reformat the data so that it can be used for the chao1() function in the fossil package. The function will then return the estimated species richness.

```
speciesrichness_binid_chao1_nymph <- function(min_lat, max_lat) {
  c_n <- df_n %>%
    select(bin_uri, lat) %>%
    filter(!is.na(bin_uri), !is.na(lat)) %>%
    filter(lat >= min_lat & lat < max_lat) %>%
    group_by(bin_uri) %>%
    count(bin_uri) %>%
    pivot_wider(names_from = bin_uri, values_from = n) %>%
    fossil::chao1(taxa.row = TRUE) %>%
    return(c_n)
}

#calling function with latitudinal bands as inputs
n_binid_chao1_015 <- speciesrichness_binid_chao1_nymph(0, 15)
n_binid_chao1_1630 <- speciesrichness_binid_chao1_nymph(16, 30)
n_binid_chao1_3145 <- speciesrichness_binid_chao1_nymph(31, 45)
n_binid_chao1_4660 <- speciesrichness_binid_chao1_nymph(46, 60)
n_binid_chao1_6175 <- speciesrichness_binid_chao1_nymph(61, 75)
n_binid_chao1_7690 <- speciesrichness_binid_chao1_nymph(76, 90)
```

#9.Creating a group bar plot to visualize species richness and calculation groups across latitude----

```
#organizing my data into the proper data frame format to make it compatible with a group bar chart
latitudes <- c('0-15','0-15', '0-15', '0-15', '16-30','16-30', '16-30', '16-30', '31-45','31-45', '31-45',
'31-45','46-60','46-60', '46-60', '46-60', '61-75','61-75', '61-75', '61-75', '76-90', '76-90', '76-90', '76-90')

calculation_groups <- c("Chao1  TaxaID", "Vegan TaxaID", "Chao1 BinID", "Vegan BinID")

species_richness_calcgroups_nymph <- c(n_taxid_chao1_015, n_taxid_vegan_015,n_binid_chao1_015,
n_binid_vegan_015, n_taxid_chao1_1630, n_taxid_vegan_1630,n_binid_chao1_1630,
n_binid_vegan_1630, n_taxid_chao1_3145, n_taxid_vegan_3145,n_binid_chao1_3145,
n_binid_vegan_3145, n_taxid_chao1_4660, n_taxid_vegan_4660, n_binid_chao1_4660,
n_binid_vegan_4660, n_taxid_chao1_6175,n_taxid_vegan_6175,n_binid_chao1_6175,
n_binid_vegan_6175, n_taxid_chao1_7690,n_taxid_vegan_7690, n_binid_chao1_7690,
n_binid_vegan_7690)


#creating the data frame
```

```
df_nymph_species_richness_calculations <- data.frame(latitudes, calculation_groups,
species_richness_calcgroups_nymph)


#creating a bar plot to visualize the results across the calculations. I am using ggplot to create a grouped
bar chart of my data so that I can visualize each species richness estimate at each of the latitudinal bands.
I am ensuring the bar plot colours are colour blind friendly by using 'Dark2' palette.
Nymph_plot <- df_nymph_species_richness_calculations %>%
  ggplot(aes(x = latitudes, y = species_richness_calcgroups_nymph, fill = calculation_groups ))+
  geom_bar(stat = 'identity', position = position_dodge(width = 0.8), width = 0.7)+
  labs(
    title = 'Species Richness Estimates of Nymphalidae and Latitude',
    x = 'Latitudes',
    y = 'Estimated Species Richness',
    fill = 'Calculations') +
  scale_fill_brewer(palette = 'Dark2') +
  theme_minimal() +
  theme(
    legend.key.size = unit(0.5, 'cm'),
    legend.text = element_text(size = 9)) +
  coord_flip()

Nymph_plot
```

Citations

Hewitt, J. E., Thrush, S. F., & Ellingsen, K. E. (2016). The role of time and species identities in spatial patterns of species richness and conservation. *Conservation Biology*, *30*, 1080–1088.

Lawrence, E. R., & Fraser, D. J. (2020). Latitudinal biodiversity gradients at three levels: Linking species richness, population richness and genetic diversity. *Global Ecology and Biogeography*, *29*, 770–788.

YouTube. (2021, June 23). *R tutorial: Mapping data on to a google map using ggmap (part 1)*. YouTube.https://www.youtube.com/watch?v=SdvGzbOZ-Qs&ab_channel=StatisticsGuideswithDrPaulChristiansen