**BINF 6210 Assignment 1**

**Introduction**

In total, the genus *Oncorhynchus* is comprised of 12 species and consists of species such as rainbow trout (*Oncorhynchus mykiss*), Coho salmon (*Oncorhynchus kisutc*) and chinook salmon (*Oncorhynchus tshawytscha*) (Stearley and Smith, 1993). The *Oncorhynchus* genus is widely regarded as the most economically significant genus of fish and has a wide native range, ranging from the Bering Sea to Taiwan and Mexico (Crawford and Muir, 2008). However, many *Oncorhynchus* species have been introduced to a variety of locations and now play critical roles in maintaining equilibrium within those ecosystems (Crawford and Muir, 2008). The genus *Lepomis* consists of fish species such as pumpkinseed (*Lepomis gibbosus*), green sunfish (*Lepomis cyanellus*) and bluegill (*Lepomis macrochirus*) (Harris et al., 2005). In total, the genus *Lepomis* consists of 12 species (Harris et al., 2005). Generally, *Lepomis* is not considered to be as economically significant as *Oncorhynchus*. It is well established that bias occurs in any scientific field; however, with respect to biodiversity research, bias occurs due to two main factors; economic importance and geography. Martín-López et al. (2009) found that the number of publications pertaining to a species is significantly influenced by the economic significance of that species. Trimble, and van Aarde (2012) found that 75% of biodiversity studies focused on the Americas and Europe, and in contrast, Africa and Asia are critically understudied. Financial biases can lead to a lack of research pertaining to certain species, thus limiting our understanding of biodiversity. Geographical biases can lead to the development of false confidence with respect to our understanding of biodiversity.

With respect to the Barcode of Life Data (BOLD) system, it is vital that contributions occur from a variety of species and locations as this will result in a more complete and accurate understanding of biodiversity. Hence, the purpose of the study is to determine if financial and geographical biases are occurring in the BOLD database. The genera *Oncorhynchus* and *Lepomis* are investigated as both genera are relatively common and contain the same number of species. It is hypothesized that the genus *Oncorhynchus* will be more adequately represented in the BOLD database than *Lepomis* due to financial and geographical biases. In order to assess what genus is most accurately represented on the BOLD database two metrics are used: the number of records per species, and the number of contributing countries per species. If the genus *Oncorhynchus* is more adequately researched, then *Oncorhynchus* will have more records per species. Similarly, if the genus *Oncorhynchus* is better represented in the database, then the *Oncorhynchus* genus will have contributions from more countries per species.

**Code**

# Assignment 1

# Version: 2023.09.0+463

# Author: Eric Bonk

library(tidyverse)

# Load Data----

Oncorhynchus <- read\_tsv(file = "http://www.boldsystems.org/index.php/API\_Public/combined?taxon=Oncorhynchus&format=tsv")

Lepomis <- read\_tsv(file = "http://www.boldsystems.org/index.php/API\_Public/combined?taxon=Lepomis&format=tsv")

# Preliminary filtering of data ----

# NA Values need to be removed. Due to the nature of the analysis only high quality and complete data will be used.For this assignment we care about the country, bins, and species.

# Remove values with no country data for both data sets

Oncorhynchus<- Oncorhynchus%>% filter(! country %in% c(NA))

Lepomis<- Lepomis%>% filter(! country %in% c(NA))

# Remove values with no bin

Oncorhynchus<- Oncorhynchus%>% filter(! bin\_uri %in% c(NA))

Lepomis<- Lepomis%>% filter(! bin\_uri %in% c(NA))

# Remove NA for species

Oncorhynchus<- Oncorhynchus%>% filter(!species\_name %in% c(NA))

Lepomis<- Lepomis%>% filter(! species\_name %in% c(NA))

# Data Check----

# For this assignment we are working with the number of records and geological data

# We need to check to make sure there is not errors

# Check to make sure both genera have at least 10 bins

bins\_Oncorhynchus<-unique(Oncorhynchus$bin\_uri)

length(bins\_Oncorhynchus)

bins\_Lepomis<-unique(Lepomis$bin\_uri)

length(bins\_Lepomis)

# Check to make sure that species are actually species in the genus

unique(Oncorhynchus$species\_name)

unique(Lepomis$species\_name)

# For the geological data need to make sure the countries are actually countries

# Determine the countries for each genus

unique(Oncorhynchus$country)

unique(Lepomis$country)

# Oncorhynchus has "Exception - Culture" which needs to be removed

Oncorhynchus<- Oncorhynchus%>% filter(!country %in% c('Exception - Culture'))

# It is important to note that "Exception - Culture" was deliberately and critically removed now in order to ensure that both the records per species analysis and geographic analysis use the same data

# Number of records analysis----

# Filter data

# Determine the number of records per species

Oncorhynchus\_simplified<-Oncorhynchus %>%

group\_by(species\_name) %>%

summarise(count = length(processid))

Lepomis\_simplified<-Lepomis %>%

group\_by(species\_name) %>%

summarise(count = length(processid))

# Summarization of key variables for analysis----

# Determine the total number of records for each genus

sum(Oncorhynchus\_simplified$count)

sum(Lepomis\_simplified$count)

# Determine mean number of records for each genus

mean(Oncorhynchus\_simplified$count)

mean(Lepomis\_simplified$count)

# Determine standard error for the mean number of records for each genus

library(plotrix)

std.error(Oncorhynchus\_simplified$count)

std.error(Lepomis\_simplified$count)

# Determine range of records for each genus

range(Oncorhynchus\_simplified$count)

range(Lepomis\_simplified$count)

# Determine if there is a statistically significant difference----

# Test assumptions

shapiro.test(Lepomis\_simplified$count)

shapiro.test(Oncorhynchus\_simplified$count)

# Levene test

library(car)

leveneTest(Lepomis\_simplified$count,Lepomis\_simplified$species\_name)

leveneTest(Oncorhynchus\_simplified$count,Oncorhynchus\_simplified$species\_name)

# The Levene's test is NA due to there being only one replicate per species so in order to further assess the distribution of the data a histogram needs to be made

hist(Oncorhynchus\_simplified$count) # Positive skew

hist(Lepomis\_simplified$count) # Positive skew

# Data is Non-parametric so Wilcox test

wilcox.test(Lepomis\_simplified$count, Oncorhynchus\_simplified$count)

# Statistically significant difference occurs

# Make bar plot for the mean number of records per species for each genus----

# Make data frame to conform to ggplot requirements

records\_per\_species<-data.frame(count=c(Lepomis\_simplified$count,Oncorhynchus\_simplified$count ),

Genus=c(rep("Lepomis", length(Lepomis\_simplified$count)),

rep("Oncorhynchus", length(Oncorhynchus\_simplified$count))))

# Plot

library(ggplot2)

ggplot(records\_per\_species, aes(x=factor(Genus), y=count)) +

geom\_bar(stat="summary")+

stat\_summary(fun.data = mean\_se,

geom = "errorbar") +

labs(y="The Mean Number of Records per Species", x="Genus")+

theme(axis.text.x = element\_text(size=7,face = "bold"))+

theme(axis.title = element\_text(size = 8,face = "bold"))

# Geological analysis ----

# Determine contributing countries for each genus

countries\_Oncorhynchus<-unique(Oncorhynchus$country)

countries\_Lepomis<-unique(Lepomis$country)

# Filter data

# Determine the number of contributing countries per species

Oncorhynchus\_simplified\_geo<-Oncorhynchus %>%

group\_by(species\_name) %>%

summarise(count = length(unique(country)))

Lepomis\_simplified\_geo<-Lepomis %>%

group\_by(species\_name) %>%

summarise(count = length(unique(country)))

# Summarization of key variables for analysis----

# Determine how many contries have contributed to each genus

length(countries\_Oncorhynchus)

length(countries\_Lepomis)

# Are there any countries shared between the 2 genera

intersect(countries\_Oncorhynchus,countries\_Lepomis) # 11 Countries in common

# Determine mean number of countries per species for each genus

mean(Oncorhynchus\_simplified\_geo$count)

mean(Lepomis\_simplified\_geo$count)

# Determine standard error for the mean number of records for each genus

std.error(Oncorhynchus\_simplified\_geo$count)

std.error(Lepomis\_simplified\_geo$count)

# Determine range of number of countries for each genus

range(Oncorhynchus\_simplified\_geo$count)

range(Lepomis\_simplified\_geo$count)

# Determine if there is a statistically significant difference----

# Test assumptions

shapiro.test(Lepomis\_simplified\_geo$count)

shapiro.test(Oncorhynchus\_simplified\_geo$count)

# Levene test

leveneTest(Lepomis\_simplified\_geo$count,Lepomis\_simplified\_geo$species\_name)

leveneTest(Oncorhynchus\_simplified\_geo$count,Oncorhynchus\_simplified\_geo$species\_name)

# The Levene's test is NA due to there being only one replicate per species so in order to further assess the distribution of the data a histogram needs to be made

hist(Oncorhynchus\_simplified\_geo$count) # Positive skew

hist(Lepomis\_simplified\_geo$count) # Positive skew

# Data is Non-parametric so Wilcox test

wilcox.test(Lepomis\_simplified\_geo$count, Oncorhynchus\_simplified\_geo$count)

# Statistically significant difference occurs

# Make map of countries that have contributed to data----

library(RgoogleMaps)

library(ggmap)

# In order to increase the readability of the map the coordinates in the Bold Data sets can not be used as each country would have many data points in slightly different locations

# Need to use a central coordinate for each country

# Must retrieve general country coordinates from the function map\_data

country.maps <- map\_data("world")

# Filter the data frame for the countries needed for each genus

country\_cord\_Oncorhynchus<- country.maps%>% filter(region %in% c(countries\_Oncorhynchus))

country\_cord\_Lepomis<- country.maps%>% filter(region %in% c(countries\_Lepomis))

# These coordinates are the border coordinates so we need to take the mean of the lat and long for each country to get a roughly central coordinate

country\_cord\_Oncorhynchus<-country\_cord\_Oncorhynchus%>%

group\_by(region) %>%

summarise(long = mean(long), lat = mean(lat))

country\_cord\_Lepomis<-country\_cord\_Lepomis%>%

group\_by(region) %>%

summarise(long = mean(long), lat = mean(lat))

#Change name of region column for convienve

country\_cord\_Oncorhynchus<-country\_cord\_Oncorhynchus%>%

rename(Country=region)

country\_cord\_Lepomis<-country\_cord\_Lepomis%>%

rename(Country=region)

# Load base map

# Note: Depending on the computer and the experience someone has with the get\_map function, an initial attempt to load a world map using this function may not work. The function get\_map archives map tiles as you use them and since I use the package frequently I have many tiles already archived. The package often won't run on a map this size if there are no archives. What you need to do is start with a smaller range of coordinates and slowly increase the range of coordinates. For me I simply had to use these lines of code:

# base = get\_map(location=c(-180,-85,180,85), zoom=3, maptype="terrain-background")

# map1 = ggmap(base)

# But in order to make sure the base map loads properly I have uploaded the ggmap file to my GitHub. For convenience I will walk through how to access the file. Note: This file was created by me using the code in lines 156 and 157 and then using the code save(map1, file = "map.RData"). I then uploaded and published the file on my github.

load(url("https://github.com/EricABonk/BINF-Assignment-1/releases/download/v1.0.0/map.RData"))

map1

# Check to make sure base map has loaded correctly and zoom is at an appropriate level

map1

# Apply genus data to base map for each genus

library(viridis) # For colors

map\_Oncorhynchus<-map1 + geom\_point(data=country\_cord\_Oncorhynchus, aes(x=long, y=lat, colour= Country), shape=20, size=2)+

scale\_color\_viridis(discrete = TRUE,option = "H")+

labs(x="Latitude (\u00B0)", y="Longitude (\u00B0)") +

theme\_bw() + theme(axis.text = element\_text(size = rel(0.75),face = "bold"),

axis.text.x = element\_text(angle=45, vjust=0.5,face = "bold"),

axis.title = element\_text(face = "bold"))+

theme(legend.text =element\_text(size = 9,face = "bold"))+

theme(legend.key.size = unit(.2, 'cm')) +

theme(legend.title = element\_text(face = "bold"))+

guides(colour = guide\_legend(ncol = 1))

map\_Lepomis<-map1 + geom\_point(data=country\_cord\_Lepomis, aes(x=long, y=lat,colour=Country), shape=20, size=2)+

scale\_color\_viridis(discrete = TRUE,option = "H")+

labs(x="Latitude (\u00B0)", y="Longitude (\u00B0)") +

theme\_bw() + theme(axis.text = element\_text(size = rel(0.75),face = "bold"),

axis.text.x = element\_text(angle=45, vjust=0.5,face = "bold"),

axis.title = element\_text(face = "bold"))+

theme(legend.text =element\_text(size = 9,face = "bold"))+

theme(legend.key.size = unit(.2, 'cm')) +

theme(legend.title = element\_text(face = "bold"))

# Place the two maps side by side

library(ggpubr)

ggarrange(map\_Oncorhynchus,map\_Lepomis, ncol=1,labels = c("A", "B"),font.label = list(size=10))

**Plots**

**A graph with a number of squares

Description automatically generated with medium confidence**

**Figure 1.** The mean number of records per species for the genera *Lepomis* and *Oncorhynchus* in the BOLD database. The *Oncorhynchus* and *Lepomis* genus had a mean number of records per species of 147.667 + 40.499 and 34.917 + 10.968. Error bars are +/- SE.

**A map of the world with different countries/regions

Description automatically generated**

**Figure 2**. The countries with contributions to the BOLD database for the genera *Oncorhynchus* (Figure A) and *Lepomis* (Figure B). The *Oncorhynchus* and *Lepomis* genera has 28 and 18 contributing countries. The map was created using the R package “RgoogleMaps”, following the methodology of Loecher and Ropkins (2015).

**Discussion**

In total the *Oncorhynchus* and *Lepomis* genera has 1329 and 419 records respectively. Interestingly, the number of records per species for the *Oncorhynchus* and *Lepomis* genera had a range of 3 to 379 and 3 to 125 respectively. Ultimately, the results of the record analysis show that a statistically significant difference in the mean number of records per species occurs between the two genera (p= 0.025, Figure 1). The *Oncorhynchus* genus had a mean number of records per species (mean= 147.667 + 40.499, Figure 1) 4.229 times larger than the mean number of records per species for the *Lepomis* genus (mean=34.917 + 10.968, Figure 1). Hence, the results suggest that economic bias is occurring in the BOLD database. In total, the *Oncorhynchus* and *Lepomis* genera has 28 and 18 contributing countries respectively (Figure 2). Interestingly, the number of countries per species for the *Oncorhynchus* and *Lepomis* genera had a range of 1 to 24 and 1 to 15 respectively. Ultimately, the results of the geographical analysis show that a statistically significant difference in the mean number of countries per species occurs between the two genera (p= 0.039). The *Oncorhynchus* genus had a mean number of contributing countries per species (mean= 7.111 + 2.312) 2.246 times larger than the mean number of contributing countries per species for the *Lepomis* genus (mean=3.167 + 1.134). Therefore, the results suggest that geographical bias is occurring in the BOLD database. Ultimately, the results of the analyses support the hypothesis that the *Oncorhynchus* genus is more adequately represented in the BOLD database than *Lepomis* due economical and geographical biases.

The results of the study align with the findings of Martín-López et al. (2009) and Titley et al. (2017) which found significant taxonomic bias occurs in scientific literature due to the economic significance of a species. Moreover, the results of the study support the results of Boakes et al. (2010) and Rocha-Ortega et al. (2021) which found that significant geographical biases occur with respect to biodiversity research. It should be noted that the *Lepomis* genus is known to hybridize, and that DNA barcoding cannot differentiate between hybrids and pure breeds which may have resulted in certain organisms being falsely classified as a specific species (Kneenleyside, 1966). The results of the study are relevant as biases and limitations with respect to biodiversity data must be fully understood in order to truly increase our knowledge of biodiversity. Future research should investigate other taxonomic groups as it is hypothesized that further economical and geographical biases exist within the BOLD database and will continue to occur in the database without diligent and dedicated research.

References

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-Qing, D., Clark, N.E., O’Connor, K., Mace, G.M., 2010. Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. PLoS Biol 8, e1000385. https://doi.org/10.1371/JOURNAL.PBIO.1000385

Crawford, S.S., Muir, A.M., 2008. Global introductions of salmon and trout in the genus Oncorhynchus: 1870-2007. Rev Fish Biol Fish 18, 313–344. https://doi.org/10.1007/S11160-007-9079-1/METRICS

Harris, P.M., Roe, K.J., Mayden, R.L., 2005. A Mitochondrial DNA Perspective on the Molecular Systematics of the Sunfish Genus Lepomis (Actinopterygii: Centrarchidae). Copeia 2005, 340–346. https://doi.org/10.1643/CG-04-035R1

Keenleyside, M.H.A., 1967. Behavior of Male Sunfishes (Genus Lepomis) Towards Females of Three Species. Evolution (N Y) 21, 688. https://doi.org/10.2307/2406766

Loecher, M., Ropkins, K., 2015. RgoogleMaps and loa: Unleashing R graphics power on map tiles. J Stat Softw 63. https://doi.org/10.18637/jss.v063.i04

Martín-López, B., Montes, C., Ramírez, L., Benayas, J., 2009. What drives policy decision-making related to species conservation? Biol Conserv 142, 1370–1380. https://doi.org/10.1016/J.BIOCON.2009.01.030

Rocha-Ortega, M., Rodriguez, P., Córdoba-Aguilar, A., 2021. Geographical, temporal and taxonomic biases in insect GBIF data on biodiversity and extinction. Ecol Entomol 46, 718–728. https://doi.org/10.1111/EEN.13027

Stearley, R.F., Smith, G.R., 1993. Phylogeny of the Pacific Trouts and Salmons ( Oncorhynchus ) and Genera of the Family Salmonidae . Trans Am Fish Soc 122. https://doi.org/10.1577/1548-8659(1993)122<0001:potpta>2.3.co;2

Titley, M.A., Snaddon, J.L., Turner, E.C., 2017. Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. PLoS One 12, e0189577. https://doi.org/10.1371/JOURNAL.PONE.0189577

Trimble, M.J., Aarde, R.J. van, 2012. Geographical and taxonomic biases in research on biodiversity in human-modified landscapes. Ecosphere 3, 1–16. https://doi.org/10.1890/ES12-00299.1