Implications of the geographical distribution of the family Crocodylidae

# How differences in geography can lead to differences in species to BIN richness.

Thishanya Gunasekera

1011995

BINF\*6210

Introduction

Studies in species diversity are imperative for understanding the distribution, richness, and abundance of various species. Studying these metrics of species spread are important for various species level quantifications, particularly for species at risk. Thus, it is important to have proper species level information in addition to well documented geographical data. However, there are lots of factors that can lead to less-than ideal data to infer these metrics. For example, in species like those within the family Crocodylidae, almost half of the datapoints within the BOLD databases are missing geographical information. This is unfortunate, given that a majority of the species in this family are endangered (Valentine, 2022). So inferring species metrics is made more challenging.

Another important metric is species richness. With the BOLD databases, the creation of barcode index numbers allows for more definitive species differentiation. However, BINs are a species proxy, and this system of species differentiation has its own problems that may not always lead to a satisfying conclusion. Thus, inferring species richness with both species and BIN richness may lessen the degree of doubt that using one of these may. For the purposes of this paper, I will use both species name and BIN identification while examining the effects of missing geographical location on species diversity.

Results

Calendar, map

Description automatically generated

Fig 1: Map of total number of Crocodylidae data points obtained from African (top) and North American (bottom) countries. This graph visualizes the countries than have recorded data on members of the Crocodylidae family, with the US holding the largest amount of data point entries. Data points taken from BOLD databases.

Scatter chart

Description automatically generated with medium confidence

Fig 2: BINs and species data on members of the Crocodylidae family found in different countries based on data obtained in BOLD databases. Left graph contains all information present while right graph has been filtered to remove NA values from BIN and country data. These plots visualize the difference in species and BIN distribution amongst the countries that have sampled members of the Crocodylidae family. We can see the convergence and divergence between different BIN and species values.

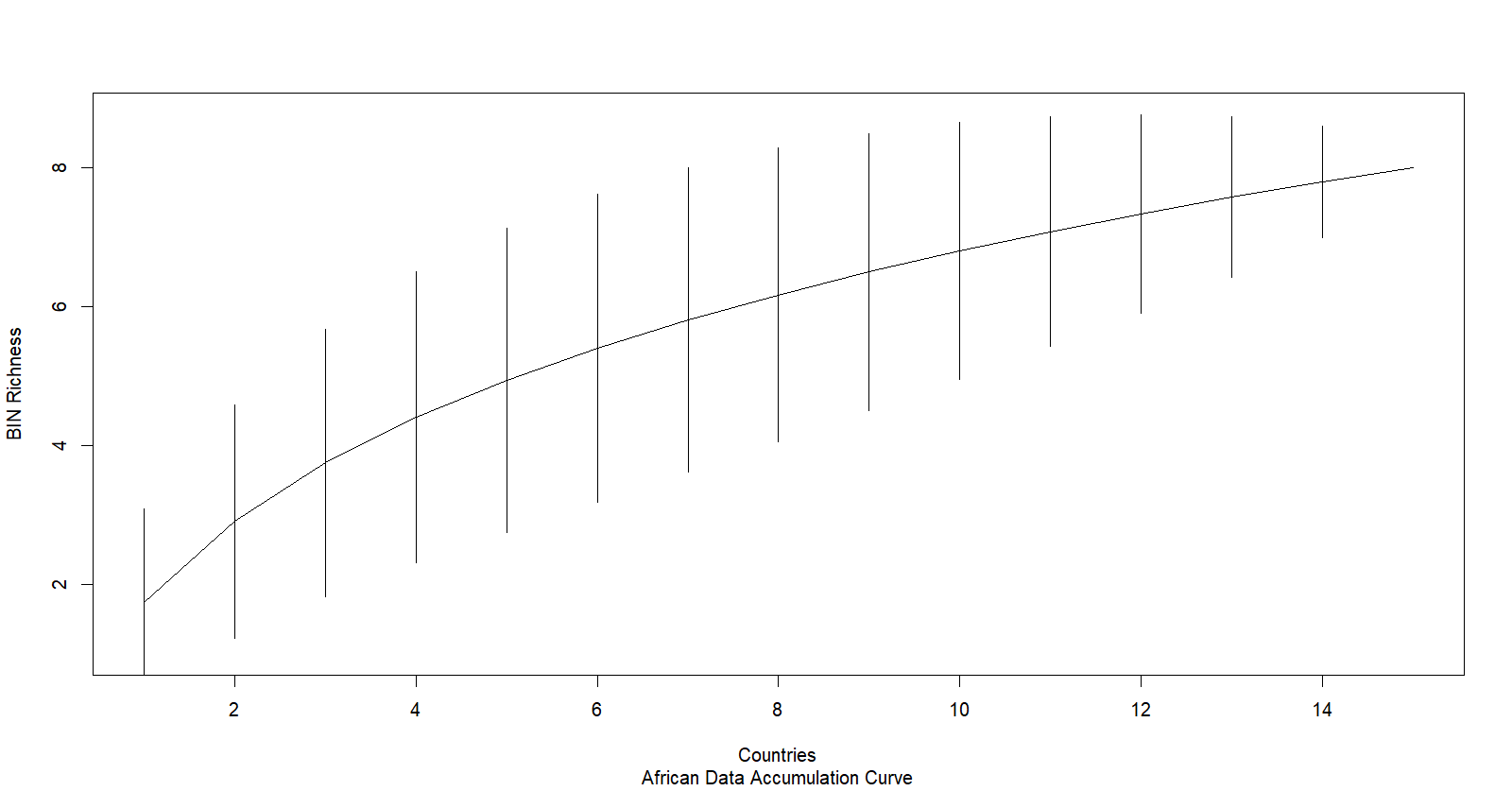


Fig 3: African BIN accumulation curve for members of the Crocodylidae family. Looks at BIN richness as BINs are sampled throughout the African countries. Given that the top end of the graph has not levels out quite yet, there seems to still be a need for more sampling to find the greater species richness within the African continent.

Chart, line chart

Description automatically generated

(This graph was added for completions sake, knowing that it will not be graded).

Fig 4: American BIN accumulation curve for members of the Crocodylidae family. Looks at BIN richness as BINs are sampled throughout countries in the Americas. Given the near linear relationship seen in this graph, we can assume that more sampling is required to find the greater species richness within the Americas.

Discussion

This data is severely lacking in geographical information. Problems presented by this lack of information are exacerbated when considering where data is lacking. That is to say, because we don’t have reliable geographical data, we will interpret the results incorrectly. For the Crocodylidae data, this lack of geographical information can lead to such conclusions as “There are only documented Crocodylidae species in the Americans and in Africa, with rare instances of Crocodylidae species in other countries, like Japan and India”. This is, however, incorrect. There are Crocodylidae species all over Asia and Oceanic regions (Valentine, 2022). The data did not compute this. Thus, there is a need for more comprehensive geographically relevant information with respect to the regions that are underrepresented in this dataset. That is to say, we need more information on species in regions that were not documented in this data set.

Disregarding NA values, there was a larger number of BINs in the Africa data than in the Americas, but there were less species for these BINs (4 species for 8 BINs) in Africa than in the Americas (6 BINs for 8 species). Despite there being more data for Africa and despite there being a larger area of study (16 African countries vs 4 countries in the Americas) there are less species indicated in the African data. There is a differential species-to-BIN richness presented between the two geographical locations of interest. There are multiple reasons this may be the case. In Africa, there are more BINs than there are species, so there is the potential that diversification of some form is being witnessed. It is also possible that the species in Africa are not as well documented and taxonomically differentiated as species in the Americas. Considering the size of the area being considered, and the historical scope, it would make sense that there is a greater species richness in Africa than what is presented in these studies. But more research into this would be required to determine how this has occurred.

References

Valentine, L. (2022). Conservation In L. Valentine (Ed.), *Reproductive Biology of the Crocodylia* (pp. 177-119). Academic Press. [https://doi.org/10.1016/B978-0-12-821801-3.00013-3](https://doi-org.subzero.lib.uoguelph.ca/10.1016/B978-0-12-821801-3.00013-3)

R file:

library(tidyverse)

library(vegan)

library(tmap)

data("World")

library(gridExtra)

library(vegan)

# downloaded on September 26th 2022

croc <- read\_tsv("bold\_data.txt")

# Start with a quick overview of the information available. We start with the names and a summary of the columns in this data frame.

str(croc)

summary(croc)

names(croc)

# We then look at some of the variables of interest. Here we find that there are no recorded values of 'habitat' (all values are NA) so this variable will be discarded

length(unique(croc$habitat))

length(unique(croc$bin\_uri))

length(unique(croc$species\_name))

length(unique(croc$lat))

length(unique(croc$lon))

length(unique(croc$country))

# With the information we gathered, we create a new data frame containing the variables of interest and summarize the new data frame

croc\_simp <- croc[c(8, 20, 22, 47, 48, 55)]

names(croc\_simp)

summary(croc\_simp)

# continuing the information summarization by looking at the number of NA values in each column of the new data frame that were not seen in the summary

croc\_simp %>%

count(is.na(bin\_uri))

croc\_simp %>%

count(is.na(species\_name))

croc\_simp %>%

count(is.na(country))

# Created charts to see if there were any clear distinctions in geographical location based on latitude and longitude. We find a large gap between longitudes that are lower than -60 and higher than -20. This is the separation between information obtained in the Americas and Africa. We then plot the latitude and longitude values together to get the general distribution.

hist(croc\_simp$lat)

hist(croc\_simp$lon)

plot(croc$lon, croc$lat)

# Create a table that displays the number of bins associated with each species when NA is removed from the bin data

view(croc\_simp %>%

group\_by(species\_name) %>%

count(species\_name, sort = TRUE))

# Create a table that looks at the total number of data points within each listed country

view(croc\_simp %>%

group\_by(country) %>%

count(country))

# Create a table that looks at the number of bins in each country after removing all NA values in both columns

view(croc\_simp %>%

filter(!is.na(country)) %>%

filter(!is.na(bin\_uri)) %>%

group\_by(country) %>%

count(bin\_uri))

# Create 2 new data frames; one with the American data and one with the African data, grouped b country.

Amer <- croc\_simp[which(croc\_simp$country == 'Colombia'),] %>%

add\_row(croc\_simp[which(croc\_simp$country == 'Cuba'),] %>%

add\_row(croc\_simp[which(croc\_simp$country == 'Mexico'),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == 'United States'),]))

Afr <- croc\_simp[which(croc\_simp$country == c('Cameroon')),] %>%

add\_row(croc\_simp[which(croc\_simp$country == c("Cote d'Ivoire")),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Democratic Republic of the Congo')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Egypt')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Gabon')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Gambia')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Ghana')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Guinea')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Madagascar')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Mauritania')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Nigeria')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Republic of the Congo')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Senegal')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('South Africa')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Uganda')),]) %>%

add\_row(croc\_simp[which(croc\_simp$country == c('Zimbabwe')),])

# Now we will create a map with the data we have for each country

# Create a data frame that groups the country data and counts the number of data points in each country. This gives us information on which countries have done the most amount of research on this subject based on the information given.

new\_col <- croc\_simp %>%

group\_by(country) %>%

count(country) %>%

filter(!is.na(country))

# We now create a new data frame that contains the world data on Africa. We add to this data frame a new column called 'Data\_Points' which will house the number of data points each country has. This data frame will be used to create a map of Africa using tmap that visualizes the number of data points each country contributed to the study of Crocodylidae

Afr\_data <- World[which(World$continent == "Africa"),] %>%

add\_column(Data\_Points = NA, .after = "sovereignt")

# We change the name of some of the countries in the new\_col dataframe so that they can be compared to the names in the Afr\_data data frame.

new\_col$country <- str\_replace(string = new\_col$country, pattern = "Republic of the Congo", replacement = "Republic of Congo")

new\_col$country <- str\_replace(string = new\_col$country, pattern = "Cote d'Ivoire", replacement = "Ivory Coast")

# We create a for loop that goes through the countries in Afr\_data, and another for loop that goes through the names in new\_col. The country names will be compared and, if they are the same, the value of n in new col (the total number of data points) will be added to the Data\_Points column in Afr\_data

for(name in Afr\_data$sovereignt){

for(n in new\_col$country){

if(name == n){

Afr\_data$Data\_Points[which(Afr\_data$sovereignt == name)] <- new\_col$n[which(new\_col$country == n)]}

}

}

# After physically running through the Data\_Points column in Afr\_data, it was discovered that the values for 'Democratic Republic of the Congo' ws not inputted. This was a result of the way the name was added to the data frame, so this value was added manually.

Afr\_data$Data\_Points[which(Afr\_data$sovereignt == "Democratic Republic of the Congo")] <- new\_col$n[which(new\_col$country =="Democratic Republic of Congo")]

# Repeat all of the previous steps that created Afr\_data, but now for the American data.

Amer\_data <- World[which(World$continent == "North America"),]%>%

add\_column(Data\_Points = NA, .after = "sovereignt")

for(name1 in Amer\_data$name){

for(n1 in new\_col$country){

if(name1 == n1){

Amer\_data$Data\_Points[which(Amer\_data$name == name1)] <- new\_col$n[which(new\_col$country == n1)]}

}

}

# From the 2 newly created data frames, we use tmap to make map shapes and fill the countries using the Data\_Points column that was added. These maps are then outputted together using tmap\_arrange.

afr\_plot <- tm\_shape(Afr\_data, name = Afr$country)+

tm\_borders() +

tm\_fill("Data\_Points")

amer\_plot <- tm\_shape(Amer\_data, name = Amer$country)+

tm\_borders() +

tm\_fill("Data\_Points")

tmap\_arrange(afr\_plot, amer\_plot)

# We then summarize these data frames.

summary(Amer)

summary(Afr)

# We also want to determine the number of BINs and species in each set. The Amer set has 7 BINs (including NA) and 8 species while the Afr set has 9 BINs (including NA) and 4 species.

length(unique(Amer$bin\_uri))

length(unique(Amer$species\_name))

unique(Amer$bin\_uri)

unique(Amer$species\_name)

length(unique(Afr$bin\_uri))

length(unique(Afr$species\_name))

unique(Afr$bin\_uri)

unique(Afr$species\_name)

# Here we create tables the show the instances of each BIN number with each species for both data set.

table(Amer$bin\_uri, Amer$species\_name)

table(Afr$bin\_uri, Afr$species\_name)

# We visualize a table that counts the number of BINs in each country. One table for each data set

view(Amer %>%

group\_by(country) %>%

count(bin\_uri))

view(Afr %>%

group\_by(country) %>%

count(bin\_uri))

# We visualize a table that counts the number of BINs for each species

view(Amer %>%

group\_by(species\_name) %>%

count(bin\_uri))

view(Afr %>%

group\_by(species\_name) %>%

count(bin\_uri))

# Plots

# Create a list of colours to be used in the next plot

cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7", "#990099", "#33FF33", "#990000", "#F5793A", "#99CC00", "#CC6600", "#663300", "#0F2080", "#85C080")

# Create a scatter plot that has bin number as the x axis and country as the y axis. Different coloured points represent different species

total\_data <- ggplot(data = croc\_simp) +

geom\_point(mapping = aes(x = bin\_uri, y = country, colour = species\_name), size = 5) +

scale\_colour\_manual(values = cbbPalette) +

theme(axis.text.x=element\_text(angle=90, vjust=0.5))

# Create a new data frame that filters out the NAs from country and BIN columns, then remake the above graph with the new dataframe

filtered\_croc\_simp <- croc\_simp %>%

filter(!is.na(bin\_uri)) %>%

filter(!is.na(country))

filtered\_data <- ggplot(data = filtered\_croc\_simp) +

geom\_point(mapping = aes(x = bin\_uri, y = country, colour = species\_name), size = 5) +

scale\_colour\_manual(values = cbbPalette) +

theme(axis.text.x=element\_text(angle=50, vjust=0.5))

# Visualize the 2 graphs together

grid.arrange(total\_data, filtered\_data, ncol=2)

# Accumulation curves to determine the level of sampling in both geographical locations.

# Create new data frames with the count of the BINs in each country and filter out the NAs from country and BIN columns.

afr\_country\_count <- Afr %>%

group\_by(country, bin\_uri) %>%

count(bin\_uri) %>%

filter(!is.na(country)) %>%

filter(!is.na(bin\_uri))

amer\_country\_count <- Amer %>%

group\_by(country, bin\_uri) %>%

count(bin\_uri) %>%

filter(!is.na(country)) %>%

filter(!is.na(bin\_uri))

# Spread the BIN data to count the number of BINs in each country. Then set any NA values to 0. We then set the row names to country name rather and ensure that there is no seperate column with country information.

afr\_spread <- pivot\_wider(data = afr\_country\_count, names\_from = bin\_uri, values\_from = n)

afr\_spread[is.na(afr\_spread)] <- 0

afr\_spread <- afr\_spread %>%

remove\_rownames %>%

column\_to\_rownames(var = "country")

# Do the same as above for the American data

amer\_spread <- pivot\_wider(data = amer\_country\_count, names\_from = bin\_uri, values\_from = n)

amer\_spread[is.na(amer\_spread)] <- 0

amer\_spread <- amer\_spread %>%

remove\_rownames %>%

column\_to\_rownames(var = "country")

# Plot both accumulation curves

AccumCurve\_afr <- specaccum(afr\_spread)

Afr\_accum\_curve <- plot(AccumCurve\_afr, xlab="Countries", ylab= "BIN Richness", sub = "African Data Accumulation Curve")

AccumCurve\_amer <- specaccum(amer\_spread)

Amer\_accum\_curve <- plot(AccumCurve\_amer, xlab = "Countries", ylab = "BIN Richness", sub = "American Data Accumulation Curve")