Comparison Among Citizen Science Efforts

Jeff Oliver

August 28, 2018

Comparing estimates of species richness among the two sampling types in the Bioscan project with that of iNaturalist. For these comparisons we are interested in *area* estimates of species richness, rather than estimates for each site.

Methods

Setup

Loading dependencies for data wrangling (`tidyr`) and data visualization (`ggplot2`).

```
library("tidyr")
library("ggplot2")
bioscan <- read.csv(file = "data/BioScanData.csv")

# Drop any rows missing data
bioscan <- na.omit(bioscan)
```

We are thus going to calculate richness for the area based on Pollard walks, richness for the area based on Malaise traps, and richness for the area based on iNaturalist data. For all these calculations, we will only have a single number; to get an idea of variation, we'll use bootstrap resampling (with replacement) to get a mean richenss and some measure of uncertainty about this estimate. For the first two (Pollard walk and Malaise trap), the process will be:

1. Create a boostrapped sample of all sites for the collection method of interest. For example, if there are 17 rows (sites) of Malaise trap data, we create a new data frame with 17 rows of data sampled with replacement from the original data.
2. Perform richness calculation on that sample.
    1. Extract all columns with species counts
    2. Perform `colSums` calculation on columns with species counts
    3. Count all species with at least one individual present in at least one site
3. Store this richness value as the estimate of richness for that sample.
4. Repeat

For the iNaturalist data, will perform similar approach, but will need only to count number of unique values in species column to determine richness of bootstrapped sample. See `scripts/gbif-processing.sh` and `scripts/gbif-additional-processing.R` for details of the iNaturalist data. The data in the file of interest still needs to be geographically restricted. For now, we will just draw a rectangle that encapsulates the sites from bioscan and only include iNaturalist observations that are included in that rectangle.

```
# Read in full data
inaturalist <- read.csv(file = "data/iNaturalist-clean.csv")

# Drop NAs (should have been done already, but just in case)
inaturalist <- na.omit(inaturalist)

# Determine boundaries of rectangle from Bioscan data
max.lon <- max(bioscan$Longitude)
```

```r
min.lon <- min(bioscan$Longitude)
max.lat <- max(bioscan$Latitude)
min.lat <- min(bioscan$Latitude)

# Restrict iNaturalist data to that rectangle
inaturalist <- inaturalist[inaturalist$longitude >= min.lon &
                           inaturalist$longitude <= max.lon &
                           inaturalist$latitude >= min.lat &
                           inaturalist$latitude <= max.lat, ]

# We want to make species column in iNaturalist match the format as in bioscan
# data: Genus_species
inaturalist$species <- gsub(pattern = " ",
                            replacement = "_",
                            x = as.character(inaturalist$species))

# Turn it back into a factor, which also means we've dropped unused levels
inaturalist$species <- as.factor(inaturalist$species)

#TODO: Will need to reconcile potential taxonomic differences between
# iNaturalist and bioscan
```

Richness calculation

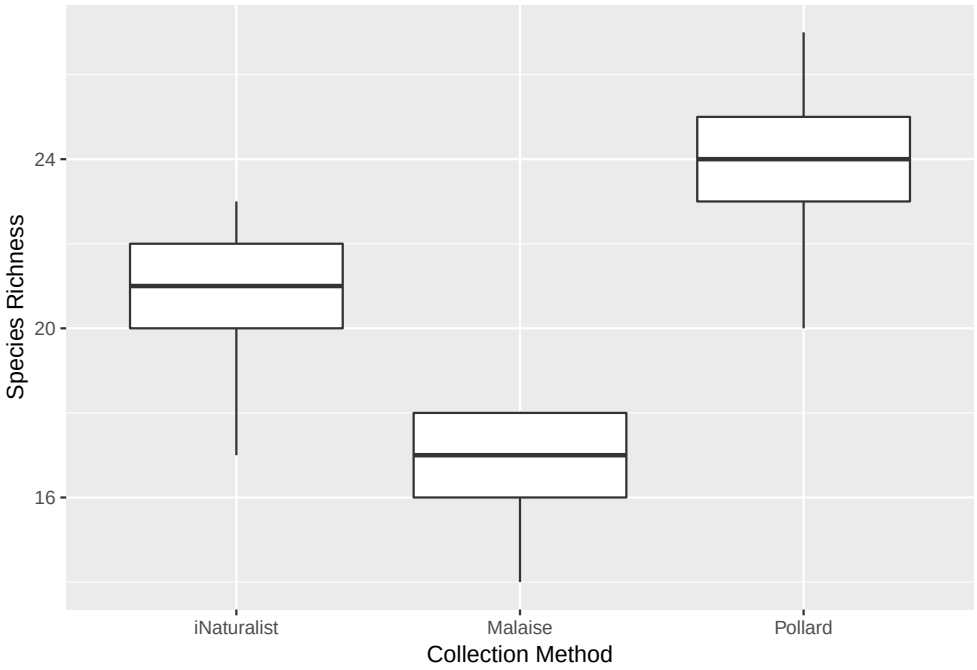
To calculate species richness, we count, for each site/collection method combination, the number of species for which at least one individual was observed. Our data are currently organized so that a single row represents a single site/collection method combination, so we can perform this operation once for each row and store the data in a new column called `richness`.

```r
# Identify those columns with species data
species.cols <- c(5:33)

# Split the two bioscan data into separate data frames for easier bootstrapping
malaise <- bioscan[bioscan$Collection.Method == "Malaise", ]
pollard <- bioscan[bioscan$Collection.Method == "Pollard Walk", ]

# Set up size of bootstrap and data frame to collect results
num.samples <- 100
bootstrapped.df <- data.frame(id = 1:num.samples,
                              Malaise = NA,
                              Pollard = NA,
                              iNaturalist = NA)

for (i in 1:num.samples) {
  bs.malaise <- malaise[sample(x = 1:nrow(malaise), size = nrow(malaise), replace = TRUE), ]
  bootstrapped.df$Malaise[i] <- sum(colSums(x = bs.malaise[, species.cols]) > 0)

  bs.pollard <- pollard[sample(x = 1:nrow(pollard), size = nrow(pollard), replace = TRUE), ]
  bootstrapped.df$Pollard[i] <- sum(colSums(x = bs.pollard[, species.cols]) > 0)

  bs.inaturalist <- inaturalist[sample(x = 1:nrow(inaturalist), size = nrow(inaturalist), replace = TRU
  bootstrapped.df$iNaturalist[i] <- length(unique(bs.inaturalist$species))
}
```

Results

We can compare the three sources of richness visually with a boxplot (after reshaping our data):

```
bootstrap.long <- gather(data = bootstrapped.df, key = "Collection.Method", value = "Richness", -id)

richness.boxplot <- ggplot(data = bootstrap.long, mapping = aes(x = Collection.Method, y = Richness)) +
  geom_boxplot() +
  xlab(label = "Collection Method") +
  ylab(label = "Species Richness")
print(richness.boxplot)
```



We also would like to glance at the overlap and differences among the three collections, a Venn diagram works best.