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")
library("dplyr")
source(file = "bioscan-functions.R")
bioscan <- CompleteBioscan()

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

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 16 rows (sites) of Malaise trap data, we create a new data frame with 16 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. We restrict the data in the iNaturalist data set to latitude and longitude boundaries in the bioscan data.

```
# Read in full data
inaturalist <- CleanINaturalist(bioscan.df = bioscan)
inaturalist.unclean <- CleanINaturalist(bioscan.df = bioscan, output = "data/iNaturalist-unclean-reduce
```

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`.

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

malaise.total <- sum(colSums(x = malaise[, species.cols]) > 0)
pollard.total <- sum(colSums(x = pollard[, species.cols]) > 0)
inaturalist.total <- length(unique(inaturalist$species))
```

```
# Set up size of bootstrap and data frame to collect results
num.samples <- 1000
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))
}

# Calculate means so we can report those
mean.bs.pollard <- mean(bootstrapped.df$Pollard)
mean.bs.malaise <- mean(bootstrapped.df$Malaise)
mean.bs.inaturalist <- mean(bootstrapped.df$iNaturalist)
```

Results

Looking at totals of species in the three data sources:

	Malaise	Pollard	iNaturalist
Observed richness	18	27	22
Bootstrapped estimate	16.39	24.113	20.117

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

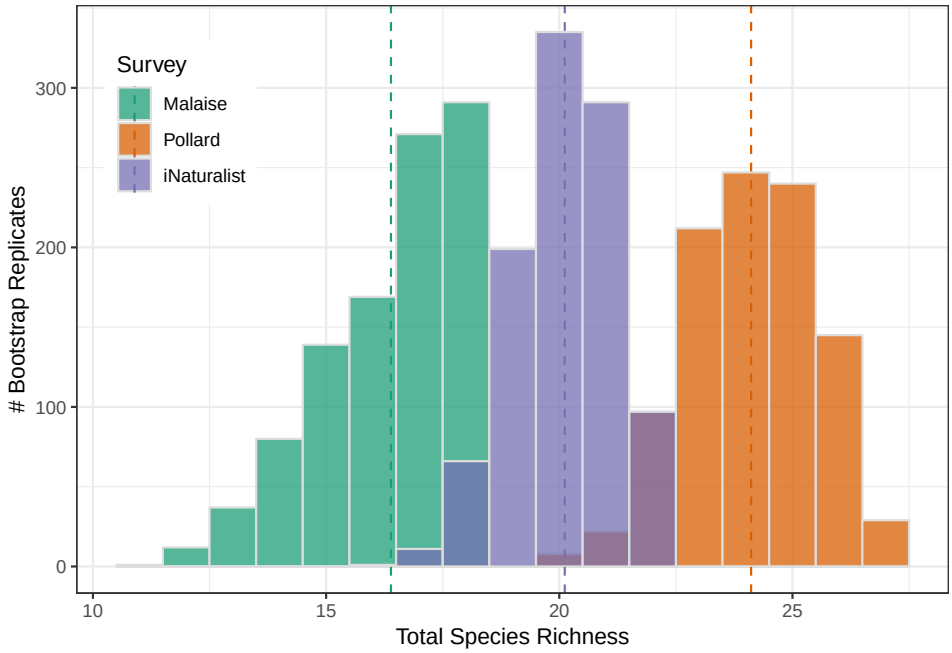
```
bootstrap.long <- gather(data = bootstrapped.df, key = "Collection.Method", value = "Richness", -id)

# Re-leveling the Collection.Method so it displays in preferred order
bootstrap.long$Collection.Method <- factor(bootstrap.long$Collection.Method,
```

```
                                        levels = c("Malaise", "Pollard", "iNaturalist"))

# Use group means for lines in histogram
means.df <- bootstrap.long %>%
  group_by(Collection.Method) %>%
  summarize(group.mean = mean(Richness))
# Re-level collection method to display as preferred
means.df$Collection.Method <-  factor(means.df$Collection.Method,
                                      levels = c("Malaise", "Pollard", "iNaturalist"))

richness.histogram <- ggplot(data = bootstrap.long, mapping = aes(x = Richness, fill = Collection.Metho
  geom_histogram(position = "identity", binwidth = 1, color = "#dddddd", alpha = 0.75) +
  scale_color_brewer(palette = "Dark2", name = "Survey") +
  scale_fill_brewer(palette = "Dark2", name = "Survey") +
  geom_vline(data = means.df,
             mapping = aes(xintercept = group.mean, color = Collection.Method),
             linetype = "dashed") +
  ylab(label = "# Bootstrap Replicates") +
  xlab(label = "Total Species Richness") +
  theme_bw() +
  theme(legend.position = c(0.12, 0.8))
print(richness.histogram)
```




```
ggsave(filename = "output/bootstrap-richness-histogram.png", plot = richness.histogram)
```

```
## Saving 6.5 x 4.5 in image
```

And do a quick t-test comparing iNaturalist to each method

```
bs.pollard.t <- t.test(x = bootstrapped.df$Pollard, y = bootstrapped.df$iNaturalist)
bs.malaise.t <- t.test(x = bootstrapped.df$Malaise, y = bootstrapped.df$iNaturalist)
```

Pollard walk vs. iNaturalist

$t = 70.194$
$p = 0$
Means = 24.113, 20.117 (Pollard, iNaturalist)

Malaise traps vs. iNaturalist

$t = -62.652$
$p = 0$
Means = 16.39, 20.117 (Malaise, iNaturalist)

Unreconciled iNaturalist

Looking at the effect of taxonomic reconciliation, and how skipping that step with the iNaturalist data may have affected our results.

```
# Set up size of bootstrap and data frame to collect results
num.samples <- 1000
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.unclean[sample(x = 1:nrow(inaturalist.unclean),
                                               size = nrow(inaturalist.unclean),
                                               replace = TRUE), ]
  bootstrapped.df$iNaturalist[i] <- length(unique(bs.inaturalist$species))
}

# Calculate means so we can report those
mean.bs.pollard <- mean(bootstrapped.df$Pollard)
mean.bs.malaise <- mean(bootstrapped.df$Malaise)
mean.bs.inaturalist <- mean(bootstrapped.df$iNaturalist)
```

Looking at totals of species in the three data sources:

	Malaise	Pollard	iNaturalist
Observed richness	18	27	22
Bootstrapped estimate	16.402	24.011	20.795

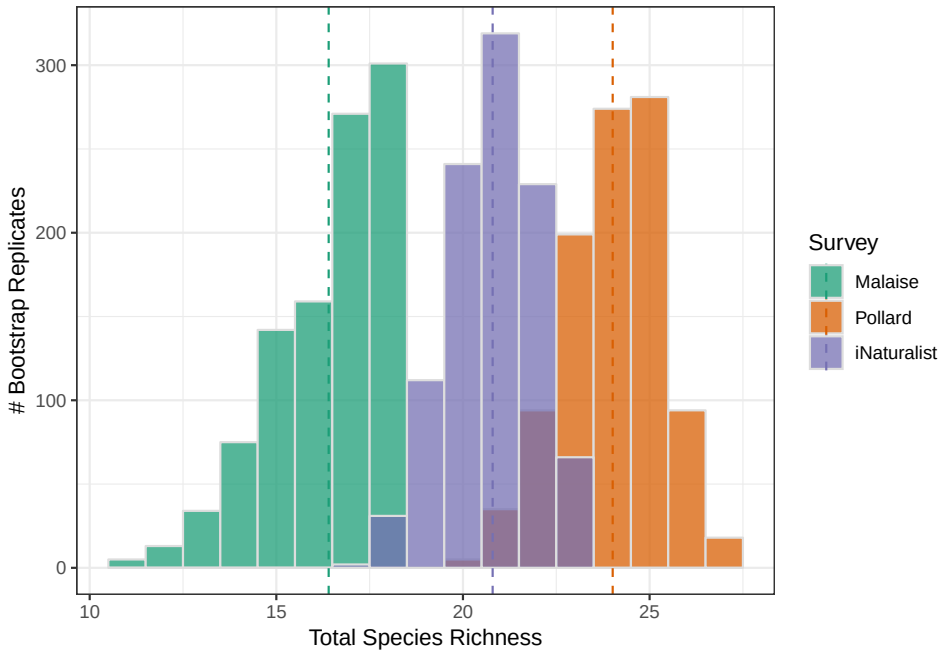
Comparing the results of the bootstrapping in a histogram:

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

```
# Re-leveling the Collection.Method so it displays in preferred order
bootstrap.long$Collection.Method <- factor(bootstrap.long$Collection.Method,
                                           levels = c("Malaise", "Pollard", "iNaturalist"))

# Use group means for lines in histogram
means.df <- bootstrap.long %>%
  group_by(Collection.Method) %>%
  summarize(group.mean = mean(Richness))
# Re-level collection method to display as preferred
means.df$Collection.Method <-  factor(means.df$Collection.Method,
                                      levels = c("Malaise", "Pollard", "iNaturalist"))

richness.histogram <- ggplot(data = bootstrap.long, mapping = aes(x = Richness, fill = Collection.Metho
  geom_histogram(position = "identity", binwidth = 1, color = "#dddddd", alpha = 0.75) +
  scale_color_brewer(palette = "Dark2", name = "Survey") +
  scale_fill_brewer(palette = "Dark2", name = "Survey") +
  geom_vline(data = means.df,
             mapping = aes(xintercept = group.mean, color = Collection.Method),
             linetype = "dashed") +
  ylab(label = "# Bootstrap Replicates") +
  xlab(label = "Total Species Richness") +
  theme_bw()
print(richness.histogram)
```




```
ggsave(filename = "output/bootstrap-richness-histogram-unclean.png", plot = richness.histogram)
```

```
## Saving 6.5 x 4.5 in image
```

And do a quick t-test comparing iNaturalist to each method

```
bs.pollard.t <- t.test(x = bootstrapped.df$Pollard, y = bootstrapped.df$iNaturalist)
bs.malaise.t <- t.test(x = bootstrapped.df$Malaise, y = bootstrapped.df$iNaturalist)
```

Pollard walk vs. iNaturalist

$t = 56.472$
$p = 0$
Means = 24.011, 20.795 (Pollard, iNaturalist)

Malaise traps vs. iNaturalist

$t = -70.807$
$p = 0$
Means = 16.402, 20.795 (Malaise, iNaturalist)