Megachile wheeleri Nest cell leaves - Morella & Museums

steppe & Em ?

Contents

List of Figures

List of Tables

1 A Beginning

While measurements for Solidago spathulata and Erigeron glaucus leaves for the purposes of identifying Megachile wheeleri nest cells were obtained via field work at Lanphere dunes, Morella californica records come from Herbarium specimens.

2 Methods

2.1 Data Acquisition

The California Consortium of Herbaria 2 website was searched on October, 2nd 2021, at ca. 4:00 PM CST. The following search terms and toggles were used to filter the database:

Taxonomic Criteria : (1) Include Synonyms (2) Morella californica

Specimen Criteria: (1) 'Limit to Specimens with data' (2) 'Limit to Specimens with Geocoordinates'

All other settings were maintained as defaults. The results, in 'Table Display' were then sorted by 'Collection Date' in an ascending order. All data were downloaded as Darwin Core, CSV, with UTF-8 encoding, with all 'Data Extensions'.

The Consortium of Pacific Northwest Herbaria was searched on October, 2nd 2021, at ca. 10:00 PM CST, filtered to only 'specimens with images'. A CSV file including longitude and latitiude was downloaded using a user profile which allowed for access to coordinates of collections in British Columbia where this is a sensitive taxon, and coordinates are generally scrubbed.

THe downloaded text file was loaded into LibreOffice Calc to load headers, which were problematic on import to R, and resaved as a comma seperated values document.

2.2 Sample Selection - Data Wrangling

To ensure an equal number of samples of this taxon relative to *Erigeron glaucus* we subset the 91 possible observations down to 54.

Data attributes were acquired in the same manner as are reflected in the primary script for analysis.

```
files <- list.files(path = "Morella_californica_CCH2_symbiota",pattern = "csv$")
files <- paste0("Morella_californica_CCH2_symbiota", sep = "/", files)
occurrence_data <- read.csv(files[4], stringsAsFactors = F, na.strings=c("","NA"))
image_data_link <- read.csv(files[2], stringsAsFactors = F, na.strings=c("","NA"))

file <- list.files(pattern = ".csv")
data_cpnwh <- read.csv(file[1], header = TRUE, stringsAsFactors = F, na.strings=c("","NA"))[,1:100]

set.seed(1125)
rm(files, file)
```

We join the neccessary specimen information and direct links to images of the specimens here from CCH2.

```
occurrence_data <- occurrence_data[,c('id', 'institutionCode','scientificName', 'eventDate', 'year', 'c
image_data_link <- image_data_link[,c('coreid','goodQualityAccessURI','Owner','associatedSpecimenRefere

names_herb <- names(occurrence_data)
n <- names(image_data_link)
n[1] <- "id"
names(image_data_link) <- n

data <- merge(occurrence_data, image_data_link, by = "id")

rm(occurrence_data, image_data_link, n)
```

We will make our CPNWH data confluent with the CCH2 data.

```
data_cpnwh <- data_cpnwh[c(1:32,34:40),]
data_cpnwh$eventDate <- paste0(data_cpnwh$Year.Collected, sep = "-", data_cpnwh$Month.Collected, sep =
data_cpnwh <- data_cpnwh[,c(1,3,14,101, 50, 54,61,62, 64, 45)]
names(data_cpnwh) <- names_herb
data_cpnwh[c("goodQualityAccessURI", "Owner", "associatedSpecimenReference", "MetadataDate")] <- NA

data <- rbind(data, data_cpnwh)
data$scientificName <- "Morella_californica"
i <- c(1,5,7,8)
data[ , i] <- apply(data[ , i], 2,            # Specify own function within apply
                    function(x) as.numeric(as.character(x)))

rm(names_herb, data_cpnwh, i)
```

2.3 Sample Selection - Sample Selection

Let's see if any of these images are duplicates from the same collection. I am sure I am not the only person who collects in multiplicate from the same individual when it comes to shrubs and trees... We will use the collection date, the collector, and the county of collection together to remove records.

```
data$county <- gsub("County","", data$county) # the application of county is inconsistent across the re
data$county <- gsub(" ", "_", data$county) # we will remove white spaces and replace with underscores s
data$county <- gsub("_$", "", data$county) # there were some trailing white spaces which are now uncers

data <- data[!duplicated(data[,c('eventDate','recordedBy', 'county')]),]
```

We are left with 124 records.

As mentioned we want up to 54 measurements, and have 86 records. Let's remove some of the older material which tends to be in rougher condition.

```
hist(data$year, xlab = "Year", ylab = "Number of Collections", main = "Imaged Herbarium Collections of
```

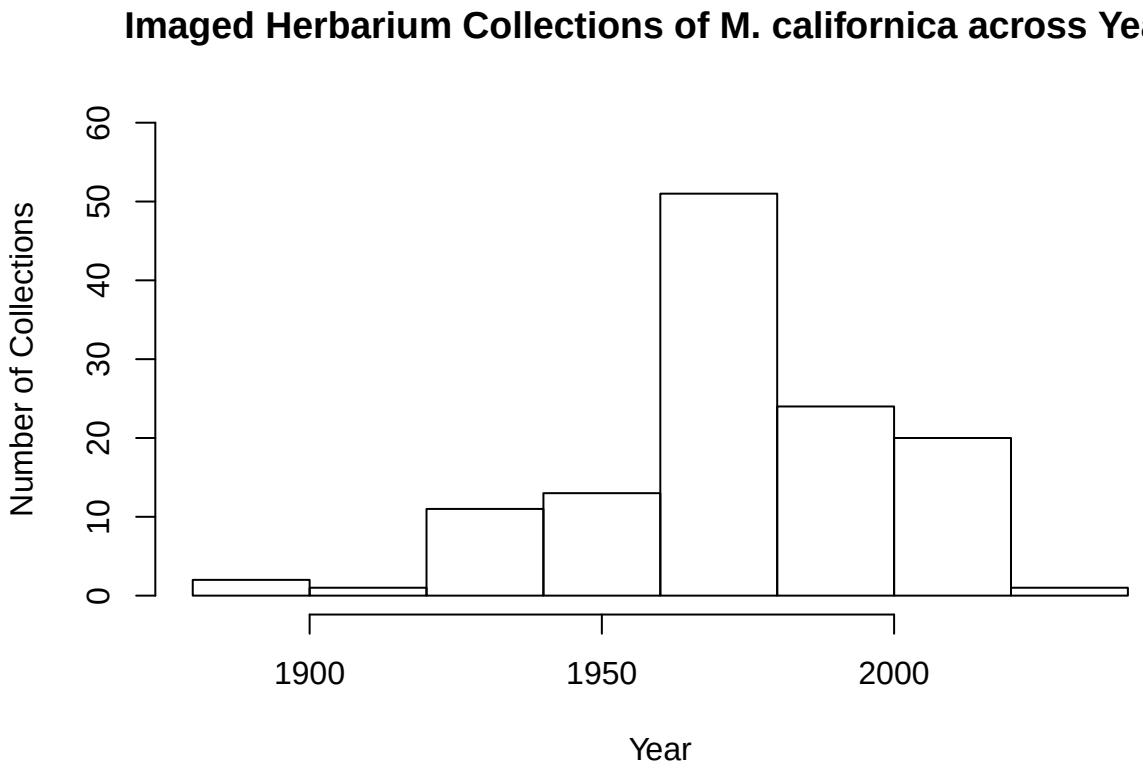

Figure 1: Records from the California Consortium of Herbaria 2 (CCH2), and Consortium of Pacific Northwest Herbaria, across time

We will remove records before 1940.

```
data <- data[data$year >= 1950,]
data <- data[!is.na(data$id),]
```

We are now left with 102 records.

Let's also see if many of these records are stacked on top of each other.

```
t <- as.data.frame(table(data$county))
names(t) <- c('county', 'specimens')
knitr::kable(t,
             caption = "Distribution of Herbarium Records by County")
```

Table 1: Distribution of Herbarium Records by County

county	specimens
Coos	7
Cowlitz	1
Curry	4
Del_Norte	1
Douglas	1
Grays_Harbor	6
Humboldt	2
Lake	1
Lane	7
Lincoln	4
Mendocino	19
Merced	1
Multnomah	1
Pacific	1
San_Francisco	2
San_Luis_Obispo	15
San_Mateo	2
Santa_Barbara	13
Santa_Cruz	4
Sonoma	2
Tillamook	1

```
rm(t)
```

I do not like that so many of the records are from Santa Barbara and San Luis Obispo county. We will downsample those so that there are only 7 possible records from those areas. We will leave Mendocino at 19, since those records are close to the study area.

```
SLOcal <- data[data$county == "San_Luis_Obispo",]
SoCal <- data[data$county == "Santa_Barbara",]
SoCal <- rbind(SLOcal, SoCal)

data <- data[data$county != "San_Luis_Obispo",]
data <- data[data$county != "Santa_Barbara",]


SoCal <- do.call(rbind,
        lapply(split(SoCal, SoCal$county),
               function(x) x[sample(nrow(x), 7, replace = F), ]))

data <- rbind(data, SoCal)
rownames(data) <- NULL

rm(SoCal, SLOcal)
```

Alright well, the draw is the draw. We have 81 samples going in. We need 54 for balanced sampling, let's pull an extra 16 samples in case some of our first draw do not have enough detail (i.e. poor imaging resolution), to be used.

```
data <- data[!is.na(data$id),]
oversample <- data[sample(nrow(data), 74), ] # possible oversamples
sample <- oversample[sample(nrow(oversample), 54),] # samples from here to get the selected mamterial.

oversample <- data.table::setDT(oversample)[!sample, on = "id"] # just figured out I do have data.table

rm(data)
```

2.4 Sample Selection - Export

We will now 'order' the oversamples, i.e. they will be sampled in this order until we have the adequate sample size.

```
oversample$draw <- sample(20, size = nrow(oversample), replace = F)
oversample <- oversample[order(draw)]

sample$draw <- NA
```

Doesn't it suck when you randomly sample away your own collections? Damn 'seeds' cannot even act like I have to re-run the script.

Write out the list of specimens to sample.

```
sample <- rbind(sample, oversample)

write.csv(sample, "MoCa_herb_samples.csv")
rm(sample, oversample)
```