Fric et al. Data formatting

Vaughn Shirey & Elise Larsen

Current version 2-Dec-2020; initiated 9-Mar-2020

*

```
# load libraries
library(tidyverse)
library(ggplot2)
library(readxl)
library(lubridate)
```

*

Data Import and Formatting

data.csv file was downloaded from https://doi.org/10.6084/m9.figshare.9946934 (https://doi.org/10.6084/m9.figshare.9946934) (https://figshare.com/articles/Phenology_responses_of_temperate_butterflies_-_Supplementary_data/9946934 (https://figshare.com/articles/Phenology_responses_of_temperate_butterflies_-_Supplementary_data/9946934))

This cvs file contains the occurrence data used in Fric et al. (2020), which they downloaded from gbif. The file includes separate data tables for each dataset, which have been concatenated into one file. These data tables have the same fields but are not formatted as a single data table; individual datasets were all written into one data file, including headers and row indices in each dataset. This first set of code reformats the data & writes formatted data files.

```
all.data <- readLines("fric_supplements/data.csv")

#identify header rows
all.header.rows<-grep("decimalLongitude", all.data)

#check headers for consistency
uniqueheaders<-unique(all.data[all.header.rows])

# 2 versions! -> Get row numbers for "header 1"
header.rows1<-grep(uniqueheaders[1], all.data)
#Get row numbers for "header 2"
header.rows2<-setdiff(all.header.rows, header.rows1)

#Create row identifiers:
#0 is a header row, 1 is format 1 data, 2 is format 2 data
j<-rep(0,length(all.data))
for (i in all.header.rows) {
  #set index to the next header if it's not the last header; otherwise set to end of datafile +
 1
  if(i<max(all.header.rows)) {
    next_index<-min(all.header.rows[all.header.rows>i])
  }else { next_index<-length(all.data)+1 }

  #for data between header rows, set row index
  j[(i+1):(next_index-1)]<-ifelse(i%in%header.rows1,1,2)
}

#need to add a row index to the header text for new data files
newheader1<-paste('"row.index\",' ,uniqueheaders[1], sep="")
newheader2<-paste('"row.index\",' ,uniqueheaders[2], sep="")

#write data file
formatteddatafile1<-file("data/fric_data_header_1.txt")
writeLines(c(newheader1,all.data[which(j==1)]), formatteddatafile1)
close(formatteddatafile1)

formatteddatafile2<-file("data/fric_data_header_2.txt")
writeLines(c(newheader2,all.data[which(j==2)]), formatteddatafile2)
close(formatteddatafile2)
rm(list=ls())

#read back in the formatted data
data1<-read_csv("data/fric_data_header_1.txt")
```

```
## Parsed with column specification:
## cols(
##   row.index = col_double(),
##   name = col_character(),
##   decimalLongitude = col_double(),
##   decimalLatitude = col_double(),
##   year = col_double(),
##   month = col_double(),
##   country = col_character(),
##   day = col_double(),
##   SuccDay = col_double(),
##   rndLat = col_double(),
##   alt = col_double()
## )
```

```
data2<-read_csv("data/fric_data_header_2.txt")
```

```
## Parsed with column specification:
## cols(
##   row.index = col_double(),
##   name = col_character(),
##   decimalLongitude = col_double(),
##   decimalLatitude = col_double(),
##   year = col_double(),
##   month = col_double(),
##   day = col_double(),
##   country = col_character(),
##   SuccDay = col_double(),
##   rndLat = col_double(),
##   alt = col_double()
## )
```

```
paste( nrow(data1), "records in format 1;", nrow(data2), "records in format 2")
```

```
## [1] "49243 records in format 1; 233201 records in format 2"
```

```
alldata<-rbind(data1,data2)
rm(data1,data2)

##Fric et al includes different species names in results tables than found in data table. In the data curation folder, we match the data names to the results names and create the name_changes.csv file. Here we change names to match results tables:
name_changes<-read_csv("data/name_changes.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   result.name = col_character(),
##   data.name = col_character()
## )
```

```
table(alldata$name[which(alldata$name %in% name_changes$data.name)])
```

```
##
##    Agriades optilete    Callophrys polios    Fabriciana adippe
##                   86                   99                 5924
## Incisalia augustinus       Lethe eurydice       Lycaeides idas
##                    4                   72                   19
##      Maculinea arion Phyciodes campestris     Phyciodes tharos
##                  755                   42                  136
##   Plebejus saepiolus   Thymelicus lineola
##                  170                11179
```

```
for(namei in 1:nrow(name_changes)) {
  alldata$name[alldata$name==name_changes$data.name[namei]]<-name_changes$result.name[namei]
}

rm(name_changes)

##Fric et al identifies datasets by region (N. America, Europe), but the data file does not incl
ude this information. We label data by region using longitude:
## visualize data density by longitude
#hist(alldata$decimalLongitude, main="Data density by Longitude")
#We label everything East of -40 as Europe, the rest as N. America
alldata<-alldata %>%
  mutate(region=ifelse(decimalLongitude>=(-40),"Europe","N. America"))

#Fric et al removed all 1st of month observations and removed one species due to late season nes
ts
fricdata<-filter(alldata, day!=1, name!="Euphydryas aurinia")

summary(fricdata)
```

```
##    row.index         name           decimalLongitude   decimalLatitude
##  Min.   :    1   Length:275457      Min.   :-162.559   Min.   : 5.787
##  1st Qu.: 2340   Class :character   1st Qu.:  -2.676   1st Qu.:52.823
##  Median : 7074   Mode  :character   Median :   9.564   Median :55.775
##  Mean   :15039                      Mean   :   6.716   Mean   :56.354
##  3rd Qu.:20814                      3rd Qu.:  23.763   3rd Qu.:60.677
##  Max.   :85273                      Max.   :  59.333   Max.   :71.216
##
##       year          month        country               day
##  Min.   :1616   Min.   : 1.0   Length:275457      Min.   : 2.00
##  1st Qu.:1992   1st Qu.: 6.0   Class :character   1st Qu.: 9.00
##  Median :2002   Median : 7.0   Mode  :character   Median :16.00
##  Mean   :1996   Mean   : 6.5                      Mean   :16.19
##  3rd Qu.:2009   3rd Qu.: 7.0                      3rd Qu.:24.00
##  Max.   :2015   Max.   :12.0                      Max.   :31.00
##  NA's   :57
##     SuccDay          rndLat           alt              region
##  Min.   :  2.0   Min.   : 6.00   Min.   :-2666.74   Length:275457
##  1st Qu.:164.0   1st Qu.:53.00   1st Qu.:   23.25   Class :character
##  Median :186.0   Median :56.00   Median :   64.24   Mode  :character
##  Mean   :181.2   Mean   :56.29   Mean   :  113.64
##  3rd Qu.:201.0   3rd Qu.:61.00   3rd Qu.:  110.77
##  Max.   :361.0   Max.   :71.00   Max.   : 4305.17
##
```

```
#Save formatted and filtered occurrence data used by Fric et al.
save(alldata,file="data/occurrences.RData")
save(fricdata,file="data/occurrences_FricAnalysis.RData")
```

End of File.