Fric et al. Data formatting

Vaughn Shirey & Elise Larsen

Current version 3-Dec-2020; initiated Feb-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!
#In the data curation file, we write the data for each header to a separate text file. See that file for the detailed code.
#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<-bind_rows(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 ma
tch the data names to the results names and create the name_changes.csv file. Here we change names to match results tables:
names1<-read_csv("data/name_changes.csv")
```

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

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

```
# this file can now be used for correcting names in the main file

for(namei in 1:nrow(names1)) {
  alldata$name[alldata$name==names1$data.name[namei]]<-names1$result.name[namei]
}

rm(names1)

##Fric et al identifies datasets by region (N. America, Europe), but the data file does not include this information. We lab
el 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 nests
fricdata<-filter(alldata, day!=1, name!="Euphydryas aurinia")

summary(fricdata)
```

```
##    row.index         name           decimalLongitude   decimalLatitude
##  Min.   :    1   Length:257972      Min.   :-162.559   Min.   : 5.787
##  1st Qu.: 2341   Class :character   1st Qu.:  -2.676   1st Qu.:52.711
##  Median : 7274   Mode  :character   Median :   9.551   Median :55.638
##  Mean   :15624                      Mean   :   6.529   Mean   :56.296
##  3rd Qu.:22563                      3rd Qu.:  23.672   3rd Qu.:60.649
##  Max.   :85273                      Max.   :  59.333   Max.   :71.216
##
##       year          month          country               day
##  Min.   :1616   Min.   : 1.000   Length:257972      Min.   : 2.00
##  1st Qu.:1992   1st Qu.: 6.000   Class :character   1st Qu.: 9.00
##  Median :2002   Median : 7.000   Mode  :character   Median :16.00
##  Mean   :1996   Mean   : 6.519                      Mean   :16.19
##  3rd Qu.:2009   3rd Qu.: 7.000                      3rd Qu.:24.00
##  Max.   :2015   Max.   :12.000                      Max.   :31.00
##  NA's   :53
##     SuccDay          rndLat           alt              region
##  Min.   :  2.0   Min.   : 6.00   Min.   :-2666.74   Length:257972
##  1st Qu.:165.0   1st Qu.:53.00   1st Qu.:   23.25   Class :character
##  Median :187.0   Median :56.00   Median :   64.24   Mode  :character
##  Mean   :181.8   Mean   :56.23   Mean   :  114.26
##  3rd Qu.:202.0   3rd Qu.:61.00   3rd Qu.:  109.48
##  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.