Fric et al. critiques: data curation

Elise Larsen & Vaughn Shirey

11/23/2020

Here we explore the occurrence data from Fric et al. (2020)

Parts of this code are duplicated in Larsen-Shirey2020_v3; this gives a more detailed account of getting from the raw data provided by Fric et al. to our cleaned datasets.

```
rm(list=ls())
# load libraries
library(tidyverse)
library(readxl)
library(ggplot2)
#library(ggExtra)
library(gridExtra)
```

Data Input

```
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"
```

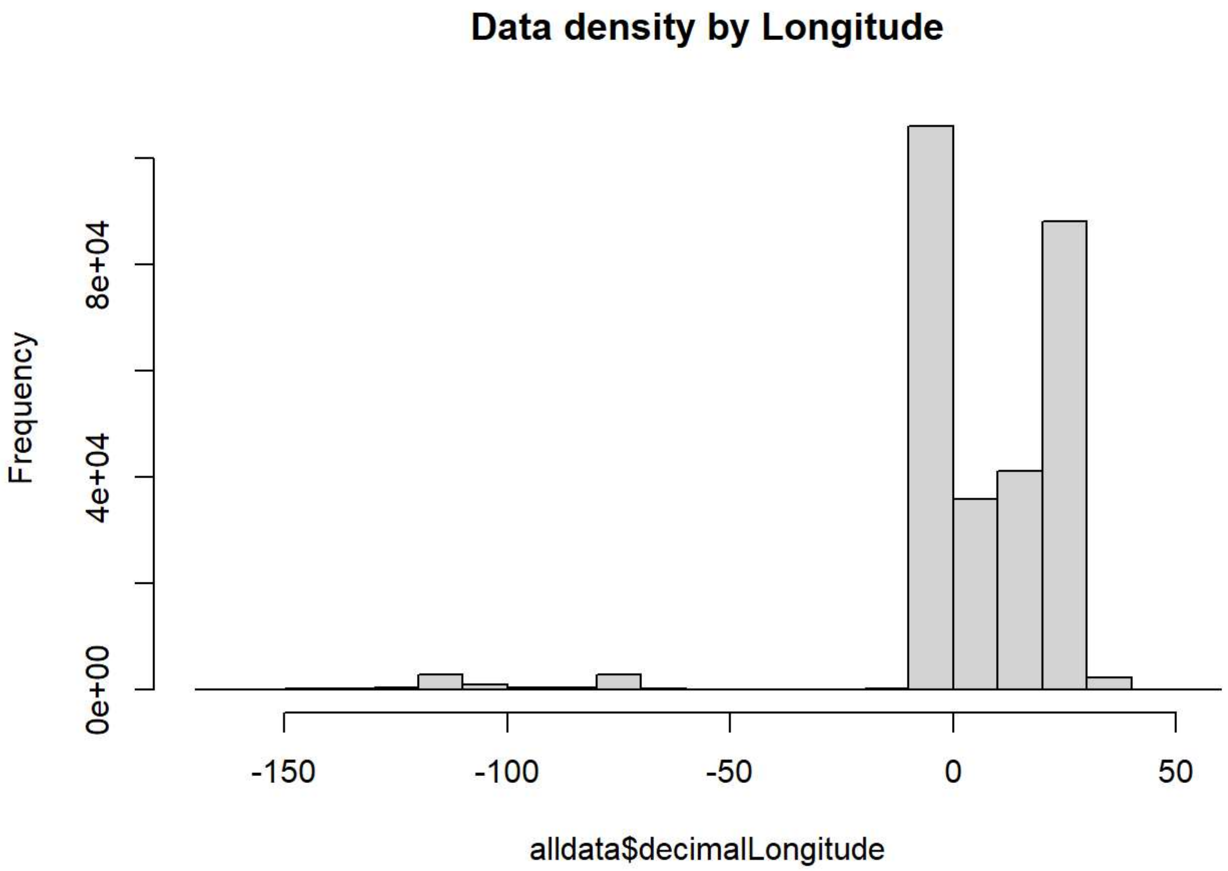
```
alldata<-rbind(data1,data2)
rm(data1,data2)
```

alldata now contains the raw data provided by Fric et al. in a usable format. ### Data exploration 1 Now we assign region, reconcile names that don’t match between the data file and results files provided in the original supplement, and filter the Fric dataset to remove first day of the month records to obtain the dataset used in Fric et al.

```
summary(alldata)
```

```
##    row.index          name           decimalLongitude   decimalLatitude
##  Min.   :    1   Length:282444      Min.   :-162.559   Min.   : 5.787
##  1st Qu.: 2367   Class :character   1st Qu.:  -2.782   1st Qu.:52.781
##  Median : 7006   Mode  :character   Median :   9.398   Median :55.628
##  Mean   :14816                      Mean   :   6.298   Mean   :56.267
##  3rd Qu.:20210                      3rd Qu.:  23.573   3rd Qu.:60.624
##  Max.   :85273                      Max.   :  59.333   Max.   :71.216
##
##       year          month          country               day
##  Min.   :1616   Min.   : 1.000   Length:282444      Min.   : 1.00
##  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   : 6.517                      Mean   :16.15
##  3rd Qu.:2009   3rd Qu.: 7.000                      3rd Qu.:24.00
##  Max.   :2015   Max.   :12.000                      Max.   :31.00
##  NA's   :58
##     SuccDay          rndLat           alt
##  Min.   :  2.0   Min.   : 6.00   Min.   :-2666.74
##  1st Qu.:163.0   1st Qu.:53.00   1st Qu.:   23.21
##  Median :186.0   Median :56.00   Median :   64.33
##  Mean   :181.6   Mean   :56.21   Mean   :  114.25
##  3rd Qu.:202.0   3rd Qu.:61.00   3rd Qu.:  111.09
##  Max.   :361.0   Max.   :71.00   Max.   : 4305.17
##
```

```
##Fric et al identifies datasets by region (N. America, Europe), but the data file does not include 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"))

#We expect 100 species names, based on the manuscript.
length(unique(alldata$name))
```

```
## [1] 108
```

```
#What are the names in the dataset?
data.names<-sort(unique(alldata$name))
#Which of these names shows up in the results?
result.names<-na.omit(read_excel("fric_supplements/Supplementary Table 2_final.xlsx", sheet="~la
titude", range="A3:A113"))
resultnames<-(strsplit(result.names$Species, " "))
result.names<-NULL
for(i in 1:length(resultnames)) {
  result.names<-c(result.names,paste(resultnames[[i]][1],resultnames[[i]][2],sep=" "))
}
which(data.names%in%result.names)
```

```
##  [1]   1   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
## [20]  21  22  23  25  26  27  28  29  31  32  33  34  35  36  37  38  39  40  41
## [39]  42  44  45  46  47  48  49  50  51  53  54  55  57  58  60  61  62  64  65
## [58]  67  68  69  70  72  73  74  75  76  78  79  80  81  82  83  84  85  86  87
## [77]  88  89  90  91  92  93  94  97  98 100 101 102 103 105 106 107
```

```
names_1<-data.names[which(!data.names%in%result.names)]
names_2<-result.names[which(!result.names%in%data.names)]

# We can link the following results names to similar data names
nmatch<-c(3,14,13,12,9,16,1,7)

#Of the remaining 8 names, Incisalia augustinus should be combined with Callophrys augustinus, L
ycaeides idas should be combined with Plebejus idas, Maculinea arion should be combined with Phe
ngaris arion. It is unclear if any others should be combined.
nmatch<-c(nmatch,8,10:11)
name_changes<-as.data.frame(cbind(result.name=c(names_2,sort(unique(result.names))[c(26,90,86
)]),data.name=c(names_1[nmatch])))
print(name_changes)
```

```
##              result.name            data.name
## 1       Callophrys polia    Callophrys polios
## 2     Icaricia saepiolus   Plebejus saepiolus
## 3       Phyciodes cocyta     Phyciodes tharos
## 4    Phyciodes pratensis Phyciodes campestris
## 5     Satyrodes eurydice       Lethe eurydice
## 6    Thymelicus lineolus   Thymelicus lineola
## 7    Vacciniina optilete    Agriades optilete
## 8        Argynnis adippe    Fabriciana adippe
## 9  Callophrys augustinus Incisalia augustinus
## 10         Plebejus idas       Lycaeides idas
## 11       Phengaris arion      Maculinea arion
```

```
write.csv(name_changes, file="data/name_changes.csv")
# this file can now be used for correcting names in the main file

for(namei in 1:nrow(name_changes)) {
  alldata$name[alldata$name==name_changes$data.name[namei]]<-name_changes$result.name[namei]
}
write.csv(alldata, file="data/all_data_formatted.csv")

fricdata<-alldata %>% filter(alldata$name %in% result.names)
rm(name_changes, resultnames, result.names, data.names, namei, names_1, names_2, nmatch)

#Fric et al removed all 1st of month observations.
fricdata<-filter(fricdata, day!=1)

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 ocurrence data used by Fric et al.
save(fricdata,file="data/occurrences_FricAnalysis.RData")
```

Data exploration: altitude (elevation)

(We defer to the Fric et al use of “altitude” for clarity)

Early on in data exploration we were concerned with the range of altitude values in the data. One aspect of our data exploration for altitude involved examining outliers and spot-checking specific occurrence records in GBIF, which were either below 0m or in the top quartile of altitudes. Looking at these records led us to understand that

- 1. GIS coordinates had often been assigned by placename, or were otherwise inaccurate, and

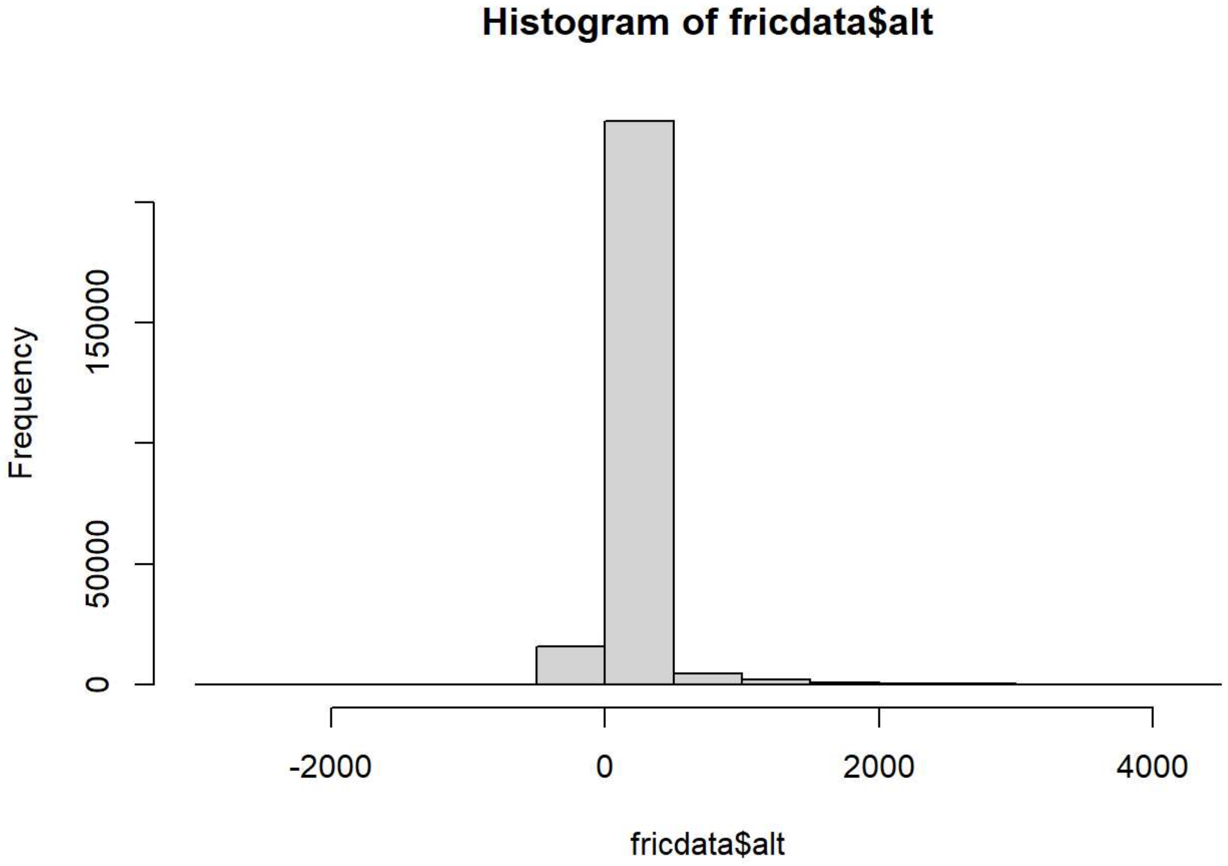
2. 2. altitudes obtained by using the Google API to extract altitude for coordinates did not provide reliable altitudes for the underlying occurrences.

Here we examine broad patterns and specific outlier cases.

```
#basic range & frequency in data
summary(fricdata$alt)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -2666.74    23.25    64.24   114.26   109.48  4305.17
```

```
hist(fricdata$alt)
```



```
#how many records below 0?
print(paste(nrow(filter(fricdata,alt<0)),"records below sea level represent", round(nrow(filter
(fricdata,alt<0))/nrow(fricdata)*100,2),"percent of all ocurrence records. We examined lat/long
 for many of these records and all examined locations were in bodies of water.",sep=" "))
```

```
## [1] "9974 records below sea level represent 3.87 percent of all ocurrence records. We examine
d lat/long for many of these records and all examined locations were in bodies of water."
```

```
#how many records are above 500m?
print(paste(nrow(filter(fricdata,alt>500)),"records above 500m represent", round(nrow(filter(fri
cdata,alt>500))/nrow(fricdata)*100,2),"percent of all ocurrence records. We examined lat/long an
d location for a small subset of high altitude records and found vague place names had been used
for geolocation.",sep=" "))
```

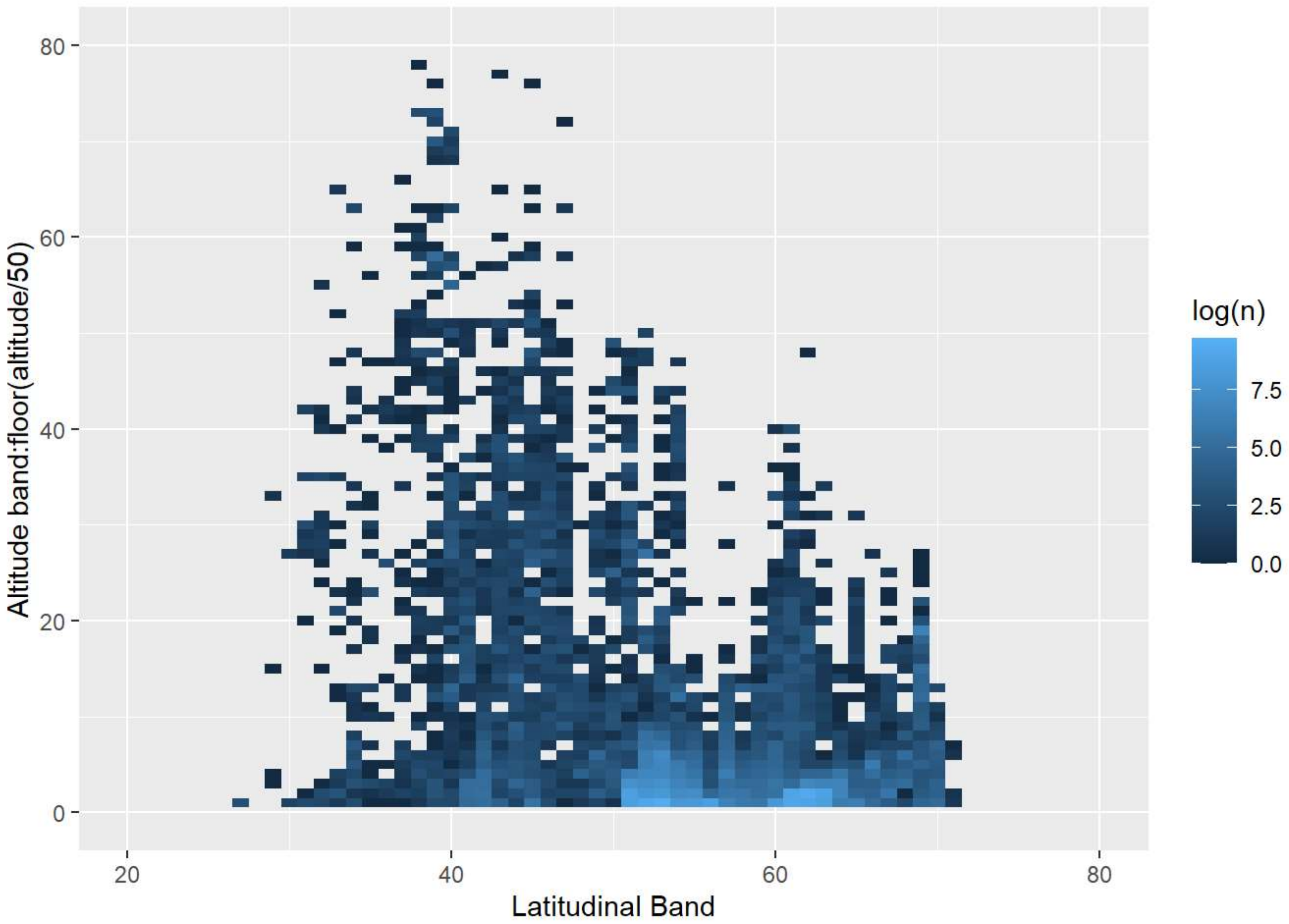
```
## [1] "8629 records above 500m represent 3.34 percent of all ocurrence records. We examined la
t/long and location for a small subset of high altitude records and found vague place names had
been used for geolocation."
```

```
#How many in the 0-500m range
print(paste(nrow(filter(fricdata,between(alt,0,500))),"records  within 0-500m represent", round
(nrow(filter(fricdata,between(alt,0,500)))/nrow(fricdata)*100,2),"percent of all ocurrence recor
ds. For reanalysis, we can constrain data to these records with minimal impact on data density.
 ",sep=" "))
```

```
## [1] "239369 records  within 0-500m represent 92.79 percent of all ocurrence records. For rean
alysis, we can constrain data to these records with minimal impact on data density. "
```

```
altdata<-fricdata %>% mutate(alt.grp=floor(alt/50))  %>%
  group_by(alt.grp, rndLat) %>% tally()
# Heatmap
ggplot(altdata, aes(rndLat, alt.grp, fill= log(n))) +
  geom_tile() + labs(x="Latitudinal Band", y="Altitude band:floor(altitude/50)") +
  xlim(20,80) + ylim(0,80)
```

```
## Warning: Removed 37 rows containing missing values (geom_tile).
```



Outliers appear to be a problem with altitude. Reviewing GBIF records, this appears to be primarily due to the assumption by Fric et al. that the GIS coordinates are precise and that the google API would provide accurate and reliable altitude metrics. Based on the records we spot-checked, when GBIF includes elevation, the values do not match those used in the analysis.

A few examples including the lowest and highest alt records, as well as some additional records selected arbitrarily from the extreme quantiles of altitude:

- 1953 Anthocharis sara record (row.index 166; altitude -525.96m) is from https://www.gbif.org/occurrence/1039154960 (https://www.gbif.org/occurrence/1039154960); geocoordinates were assigned via vertnet in 2015. These coordinates are located in the ocean. The GBIF record traces to https://collections.peabody.yale.edu/search/Record/YPM-ENT-729028 (https://collections.peabody.yale.edu/search/Record/YPM-ENT-729028) which simply gives a locality of

"North America; USA; California; Los Angeles County; Rolling Hills". Rolling Hills, CA is ~10km east of the given lat/long according to our estimation using googlemaps.
- 1991 Parnassius smintheus record (row.index 38; altitude 4048m) is from https://www.gbif.org/occurrence/1039027733 (https://www.gbif.org/occurrence/1039027733) (which gives elevation of 3810m). The GBIF record traces to https://collections.peabody.yale.edu/search/Record/YPM-ENT-430824 (https://collections.peabody.yale.edu/search/Record/YPM-ENT-430824) which gives a locality of "North America; USA; Colorado; Summit County; Loveland Pass, 3810 m". The actual collection altitude is provided by the source, and is different than that used in the analysis.
- 1918 Euphydryas chalcedona record (row.index 139; altitude 4305m) is the highest record in the data. It's from https://www.gbif.org/occurrence/1039181223 (https://www.gbif.org/occurrence/1039181223). The GBIF record traces to https://collections.peabody.yale.edu/search/Record/YPM-ENT-819202 (https://collections.peabody.yale.edu/search/Record/YPM-ENT-819202) which gives a locality of "North America; USA; California; Siskiyou County; Mount Shasta" There is a city named Mount Shasta, CA that incorporated in 1905 that is at elevation 1100m and the peak of Mount Shasta is 4320. It is unclear whether the locality refers to the mountain or to the city; either way it is unlikely that an altitude so close to the peak of the mountain is the best choice for this specimen.

So far those examples are all North America - does this problem exist in Europe too?

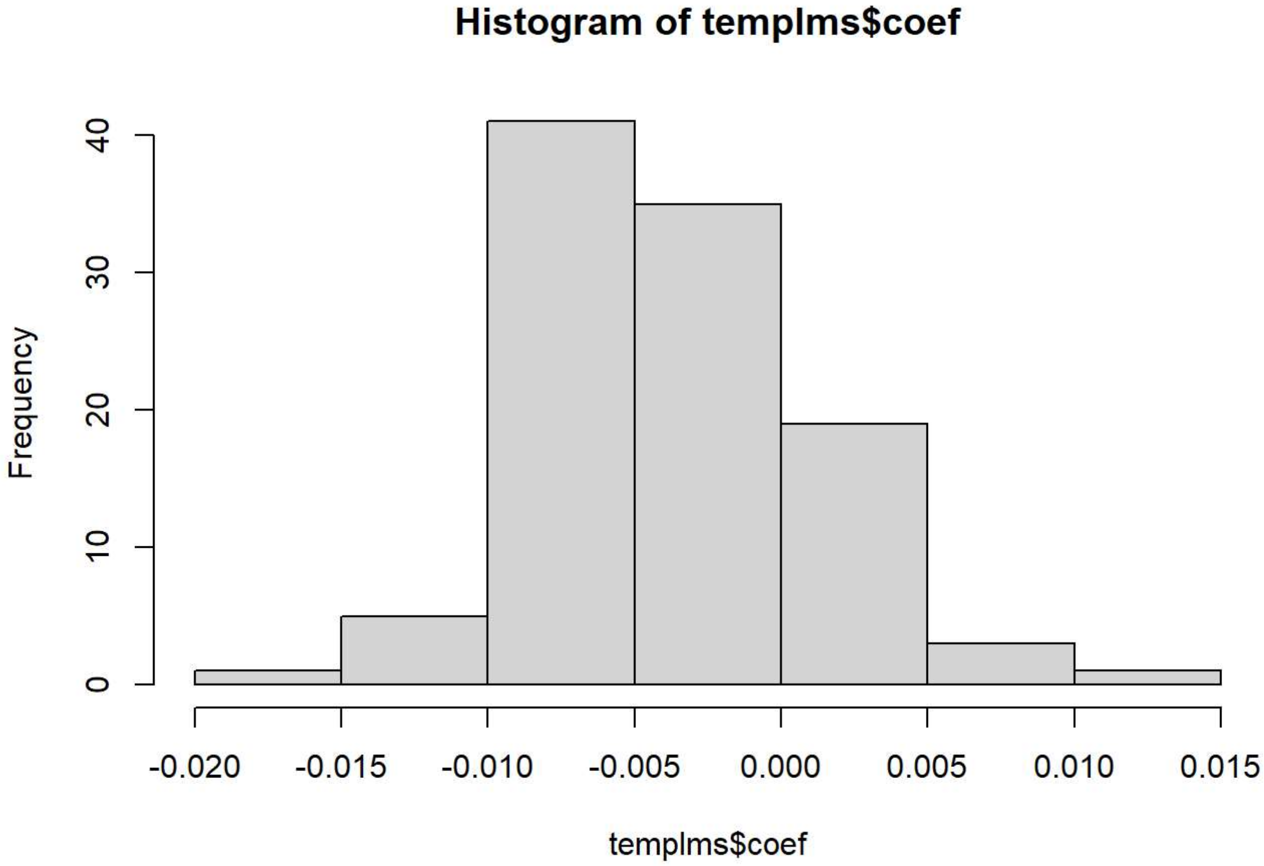
- A Lycaena hippothoe record from 1995 (row.index 2160; altitude 3274m) is from https://www.gbif.org/occurrence/2570253925 (https://www.gbif.org/occurrence/2570253925) which lists an inferred elevation of 2000m.
- A Lycaena virgaureae record from 2002 (row.index 4501; altitude -85.8m) appears to match https://www.gbif.org/occurrence/173651704 (https://www.gbif.org/occurrence/173651704) which is located in the Gulf of Bothnia, though GBIF assigns an elevation of 0m. Considering the lat/long are (65,23) most likely those coordinates are imprecise.

Altitude ~ Latitude collinearity

Fric et al. used regression of residuals for corrected analyses. Regression of residuals is not recommended, particularly if there could be collinearity among explanatory variables. We examined the collinearity between altitude and latitude, which would indicate the regression of residuals analysis would produce biased parameter estimates.

```
#Additional issues with altitude
#Given the use of regression of residuals, we were concerned that collinearity among independent
variables could have led to biased results.

#How many datasets have significant collinearity between altitude and latitude?
templms<-NULL
datasets<-fricdata %>% group_by(name, region)  %>% tally()
for (spi in 1:nrow(datasets)) {
  tempdata<-fricdata %>% filter(name==datasets$name[spi],region==datasets$region[spi])
  spilm<-summary(lm(rndLat~alt, data=tempdata))
  templms<-rbind(templms,c(nrow(tempdata), spilm$coefficients[2,1],  spilm$coefficients[2,4], sp
ilm$r.squared))
}
templms<-as.data.frame(templms)
names(templms)<-c("n","coef","pval","r2")
hist(templms$coef)
```



```
summary(templms)
```

```
##        n              coef                pval              r2           
##  Min.   :   15   Min.   :-0.019376   Min.   :0.00000   Min.   :0.0000222  
##  1st Qu.:   78   1st Qu.:-0.006861   1st Qu.:0.00000   1st Qu.:0.0280076  
##  Median :  189   Median :-0.004516   Median :0.00000   Median :0.1909175  
##  Mean   : 2457   Mean   :-0.003832   Mean   :0.06654   Mean   :0.2824787  
##  3rd Qu.: 1067   3rd Qu.:-0.001088   3rd Qu.:0.00851   3rd Qu.:0.5261002  
##  Max.   :51819   Max.   : 0.014635   Max.   :0.86050   Max.   :0.8487862
```

```
round(nrow(filter(templms,pval<0.05))/nrow(templms),2)
```

```
## [1] 0.85
```

```
#How many datasets have significant collinearity
print(paste(nrow(filter(templms,pval<0.05)),"datasets have significant collinearity, representin
g", round(nrow(filter(templms,pval<0.05))/nrow(templms)*100,1),"percent of all datasets. For dat
asets with significant collinearity, the mean coefficient is",round(mean(templms$coef[templms$pv
al<0.05]),3),"(which translates to a slope of", round(1/mean(templms$coef[templms$pval<0.05]),0
),"meters per degree latitude) and mean r-squared is",round(mean(templms$r2[templms$pval<0.05]),
3)," - therefore regression of residuals is likely producing bias parameters.",sep=" "))
```

```
## [1] "89 datasets have significant collinearity, representing 84.8 percent of all datasets. Fo
r datasets with significant collinearity, the mean coefficient is -0.004 (which translates to a
slope of -224 meters per degree latitude) and mean r-squared is 0.33  - therefore regression of
residuals is likely producing bias parameters."
```

Data exploration: data density

- In Fric et al. (2020), datasets were analysed with as few as 15 ocurrence records.
- We examine the prevalence of singleton ocurrences, when just one ocurrence was available in a latitudinal band.

```
lat.summary1<-fricdata %>%
  group_by(name, region, rndLat) %>%
  summarize(lat.samplesize=n(),singleton=ifelse(lat.samplesize==1,1,0),dur=max(SuccDay)-min(Succ
Day))
```

```
## `summarise()` regrouping output by 'name', 'region' (override with `.groups` argument)
```

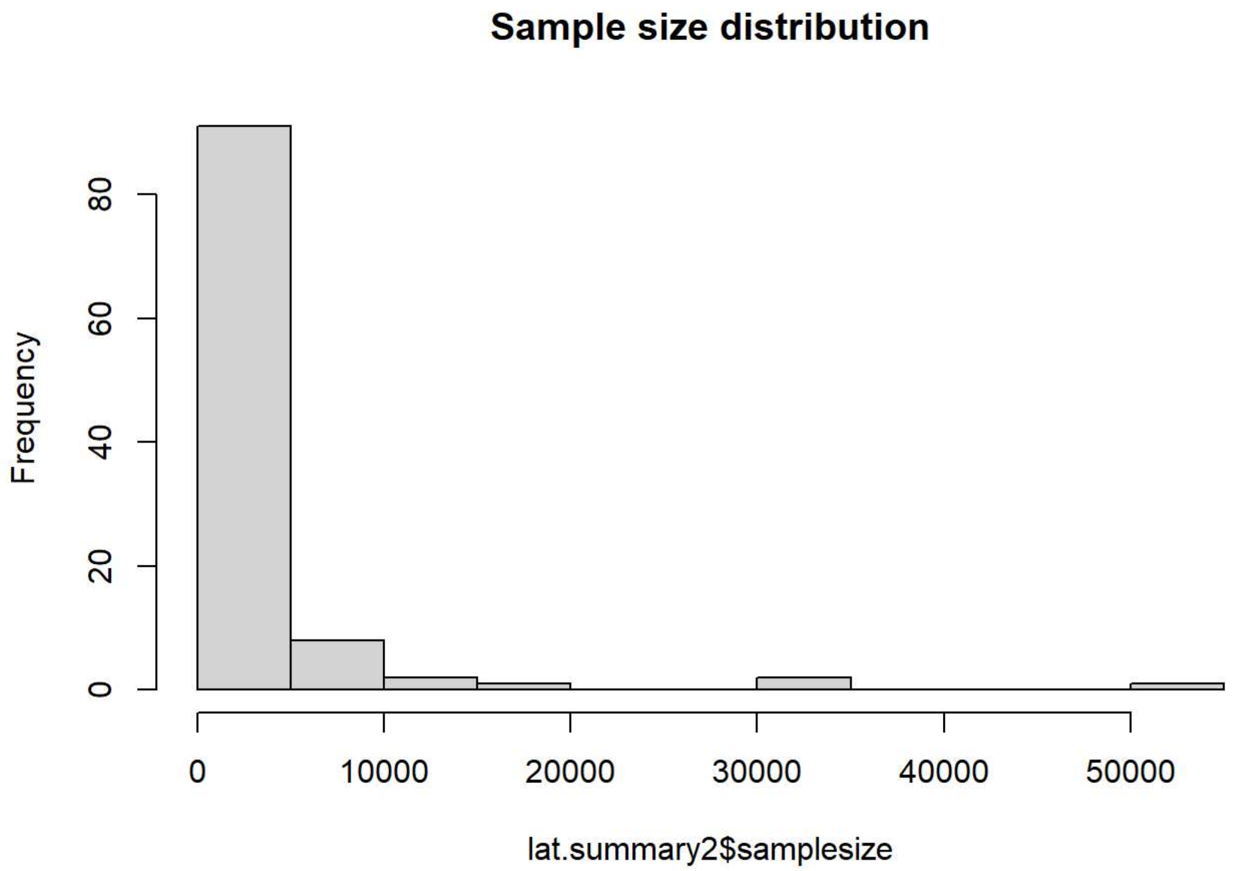
```
lat.summary2<-lat.summary1 %>%
  group_by(name,region) %>%
  summarize(samplesize=sum(lat.samplesize),latspan=max(rndLat)-min(rndLat),nlats=length(unique(r
ndLat)),n.singletons=sum(singleton),prop.singletons=n.singletons/nlats)
```

```
## `summarise()` regrouping output by 'name' (override with `.groups` argument)
```

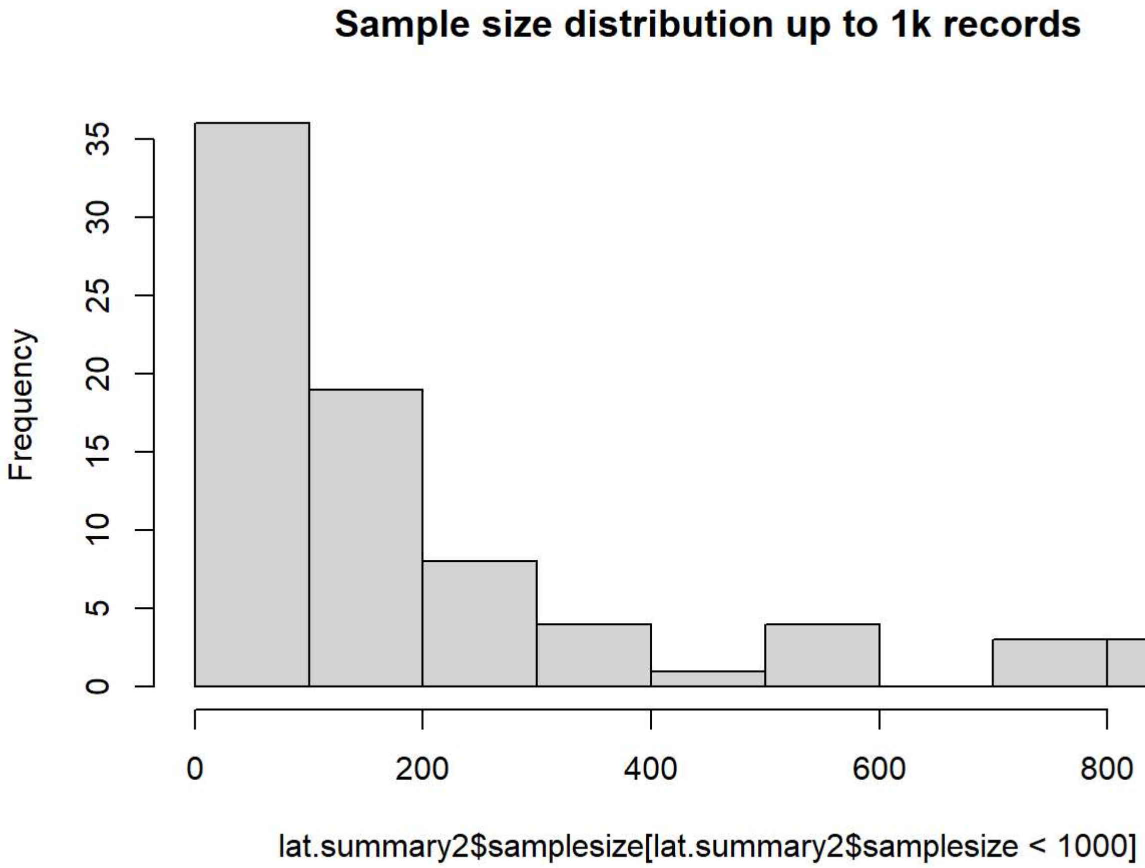
```
summary(lat.summary2)
```

```
##      name              region            samplesize       latspan    
##  Length:105         Length:105         Min.   :   15   Min.   :10.0  
##  Class :character   Class :character   1st Qu.:   78   1st Qu.:24.0  
##  Mode  :character   Mode  :character   Median :  189   Median :27.0  
##                                        Mean   : 2457   Mean   :26.3  
##                                        3rd Qu.: 1067   3rd Qu.:30.0  
##                                        Max.   :51819   Max.   :64.0  
##      nlats       n.singletons    prop.singletons  
##  Min.   : 5.0   Min.   : 0.000   Min.   :0.00000  
##  1st Qu.:13.0   1st Qu.: 2.000   1st Qu.:0.09375  
##  Median :18.0   Median : 3.000   Median :0.19048  
##  Mean   :18.9   Mean   : 3.429   Mean   :0.20831  
##  3rd Qu.:25.0   3rd Qu.: 5.000   3rd Qu.:0.33333  
##  Max.   :33.0   Max.   :10.000   Max.   :0.60000  
```

```
#Visualize range of sample sizes
hist(lat.summary2$samplesize, main="Sample size distribution")
```



```
#look at the lower end of sample sizes, where most datasets are
hist(lat.summary2$samplesize[lat.summary2$samplesize<1000], main="Sample size distribution up to
1k records")
```



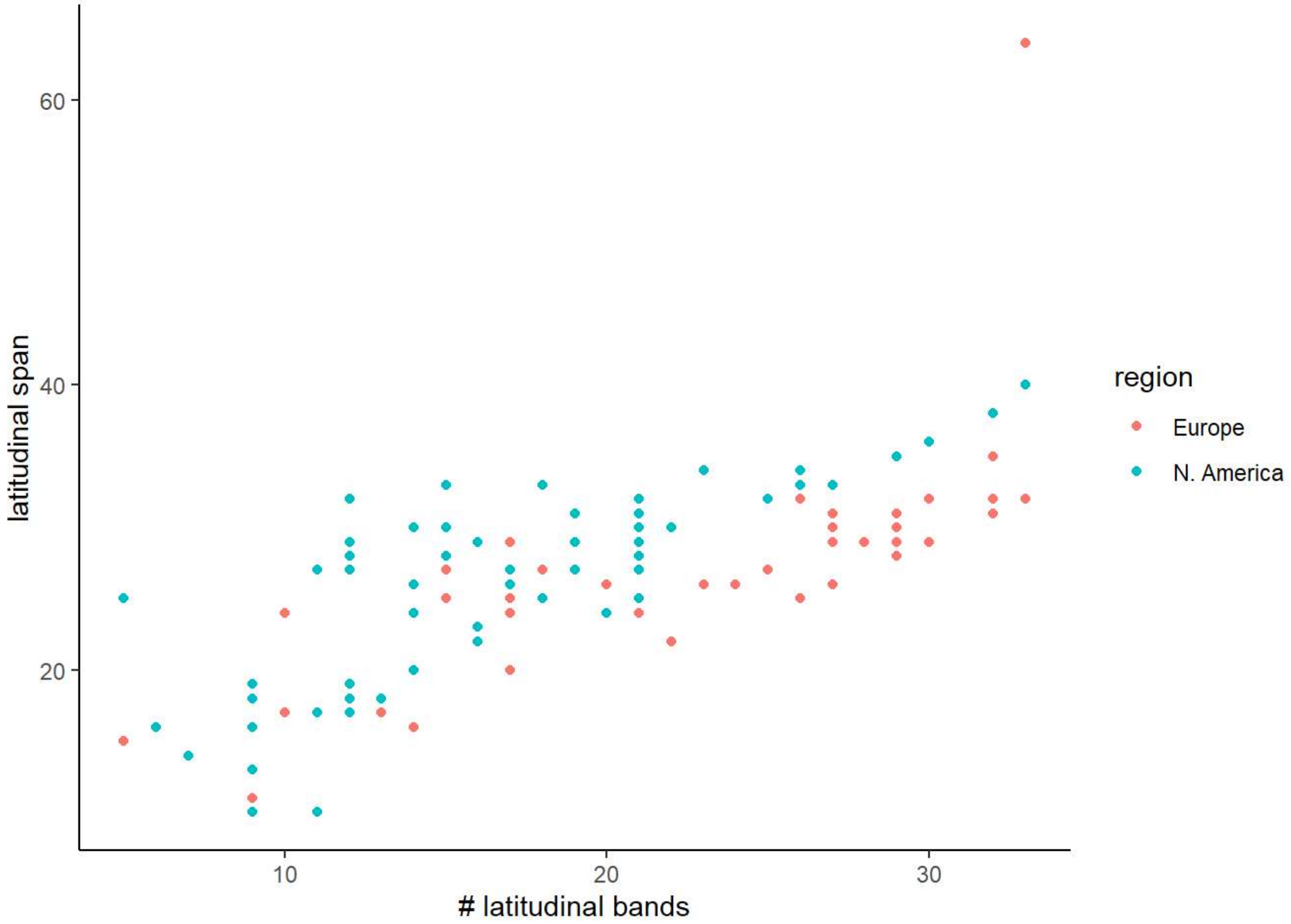
```
nrow(lat.summary2 %>% filter(samplesize<100))
```

```
## [1] 36
```

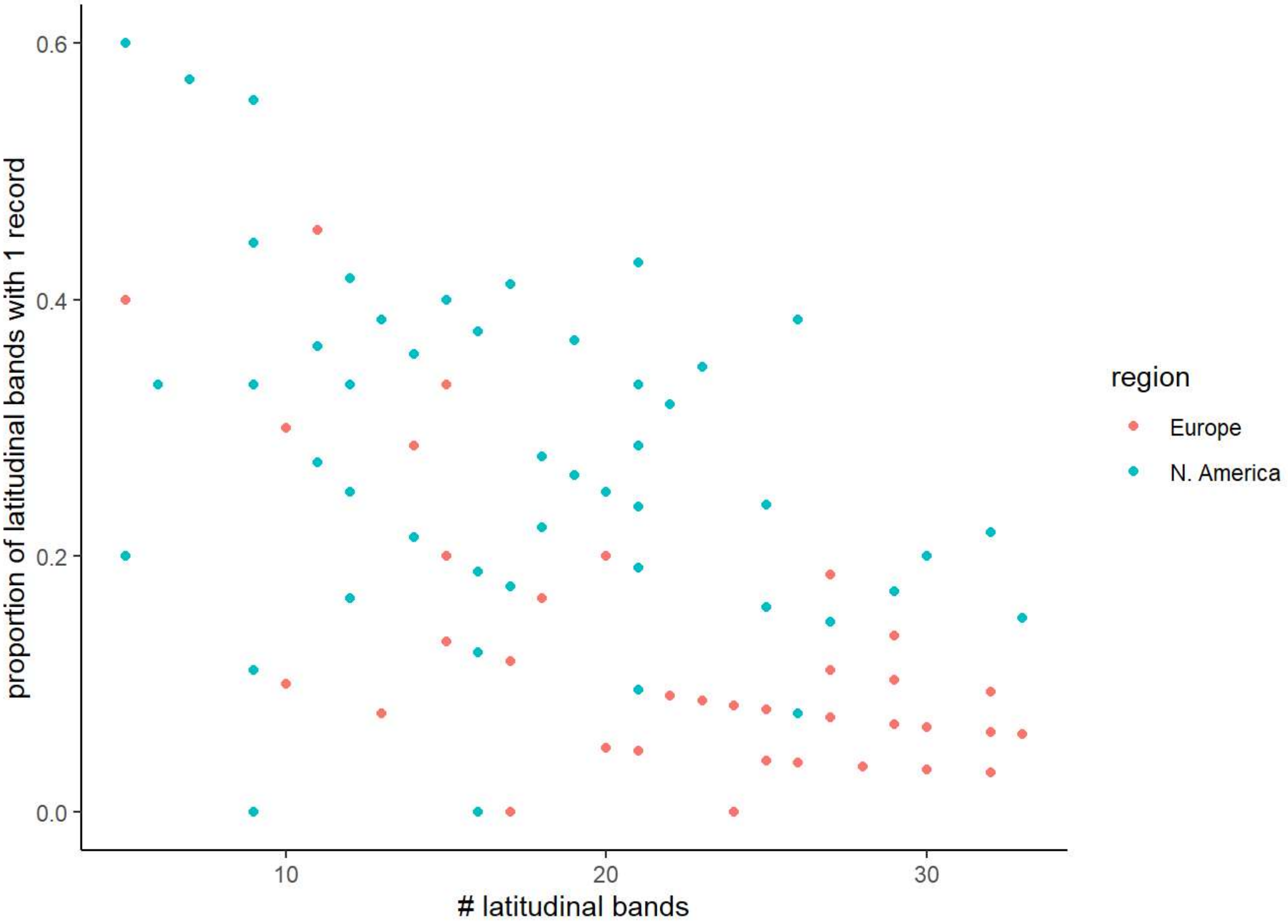
```
print(paste(nrow(lat.summary2 %>% filter(samplesize<100)),"datasets have less than 100 ocurrence
records."))
```

```
## [1] "36 datasets have less than 100 ocurrence records."
```

```
ggplot(data=lat.summary2, aes(x=nlats, y=latspan, color=region)) + geom_point() + theme_classic
() +
  labs(x="# latitudinal bands", y="latitudinal span")
```



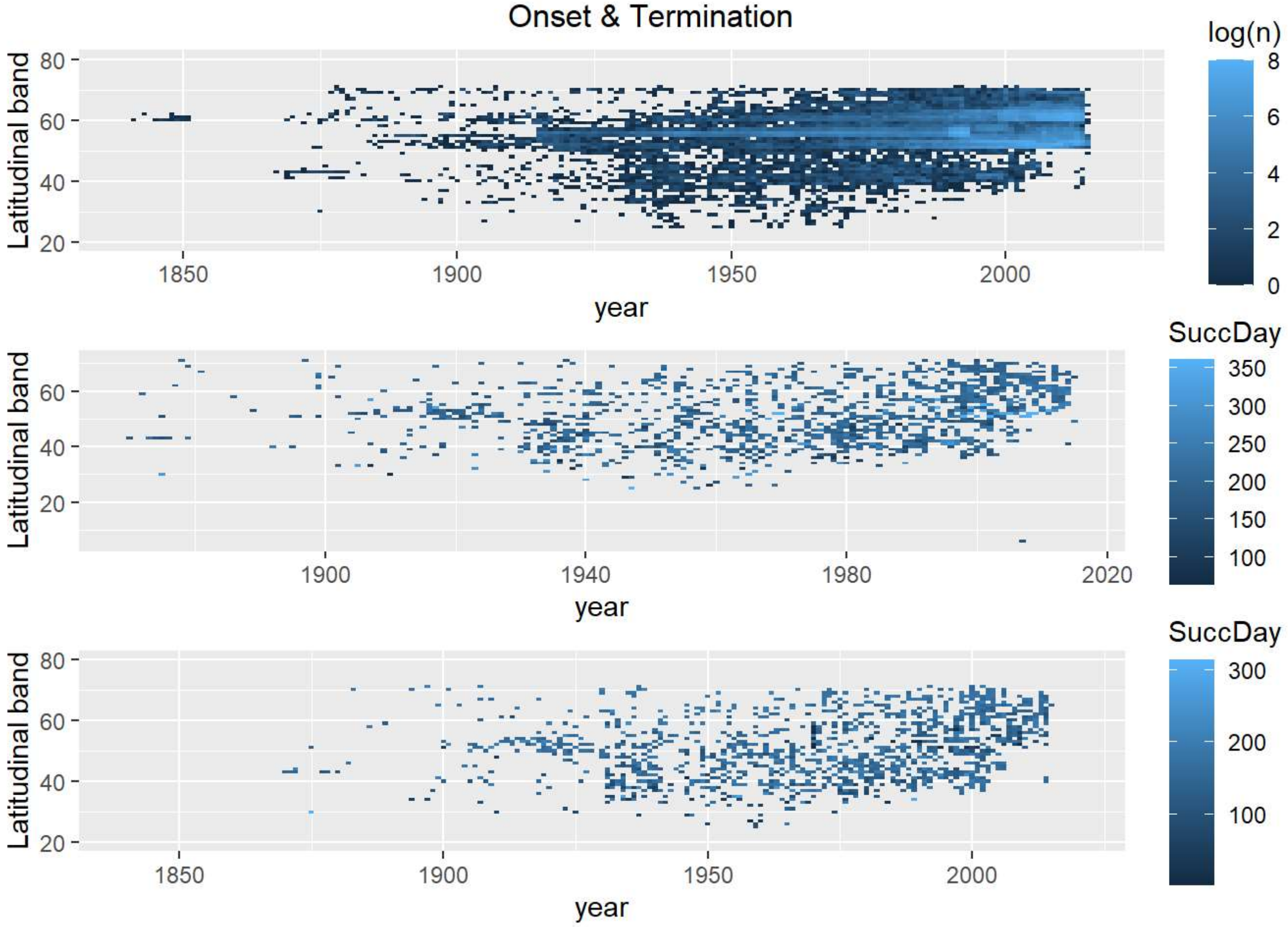
```
ggplot(data=lat.summary2, aes(x=nlats, y=prop.singletons, color=region)) + geom_point() + theme_
classic() +
  labs(x="# latitudinal bands", y="proportion of latitudinal bands with 1 record")
```

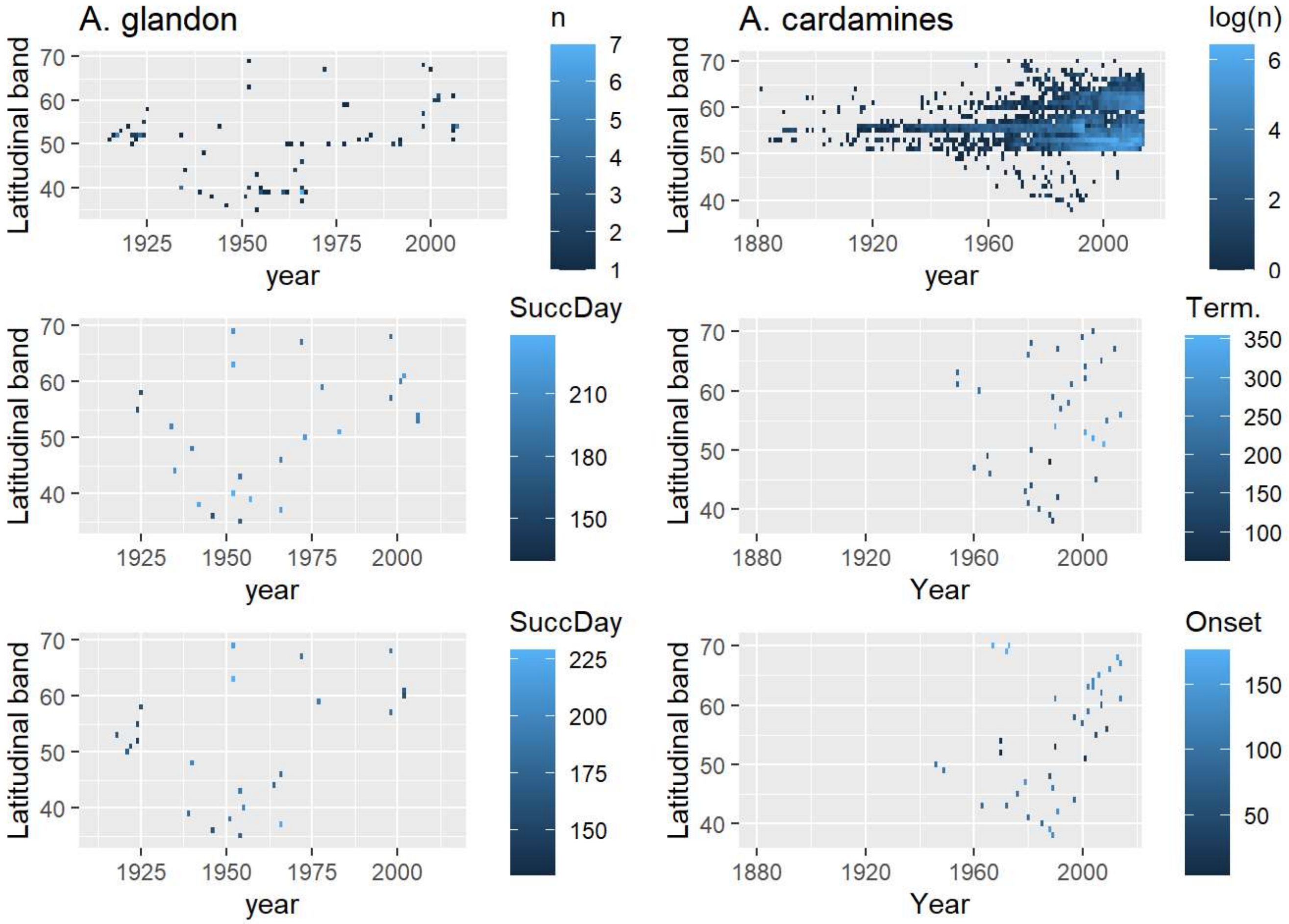


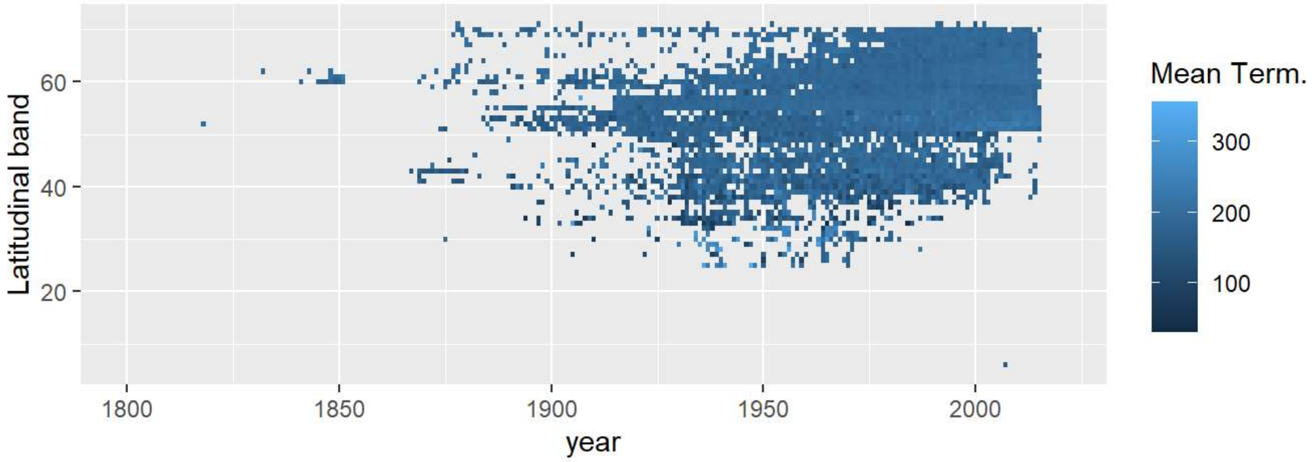
Data exploration: year

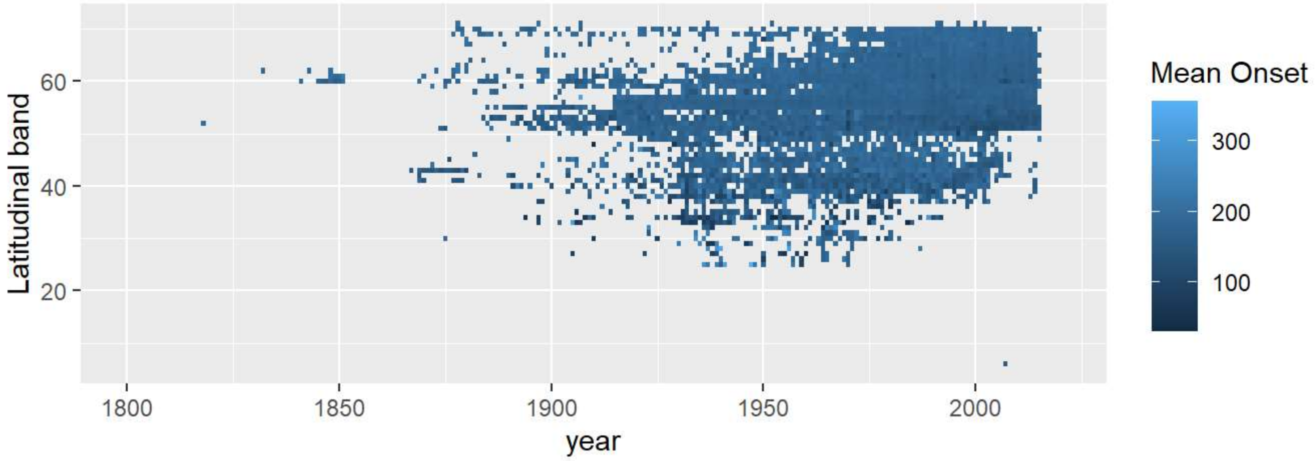
As expected, most data are quite recent. By selecting the min and max day of year per latitudinal band as onset & termination, the authors vastly decrease their sample size and remove most of the variation along the year and altitude axes

We arbitrarily selected two species, one with a low sample size and one with a large sample size, to visualize.









Recreate original results We also wanted to confirm that we understood correctly the Fric et al. analysis. We attemped to recreate the original Fric et al. analysis

```
## NEED TO ADD THIS CODE
```

End of File