Fric et al. critiques: data curation

Elise Larsen & Vaughn Shirey

Updated 1-Dec-2020; Initially assembled 23-Nov-2020

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

This gives a detailed account of some data curation issues we observed in the Fric et al. data and curation.

```
rm(list=ls())
# load libraries
library(tidyverse)
library(readxl)
library(ggplot2)
library(ggExtra)
library(gridExtra)
# install.packages("viridis")
library(viridis)
```

```
## Warning: package 'viridis' was built under R version 4.0.3
```

Data Input

```
#Import formatted occurrence data (alldata; RData file created by LarsenShirey_dataFormatting.Rm
d)
load("data/occurrences.RData")
```

New code 11/25/2020: day of year reconciliation

Until today, we had assumed that the “SuccDay” values were a consistent index for day of year. However, we had not documented our initial spot-checking of altitudes. While identifying GBIF records for documented sopt-checking, we found some inconsistencies in the SuccDay value. Here we identify how “SuccDay” was calculated.

```
#DOES SUCCDAY MATCH DOY?
library(lubridate)
```

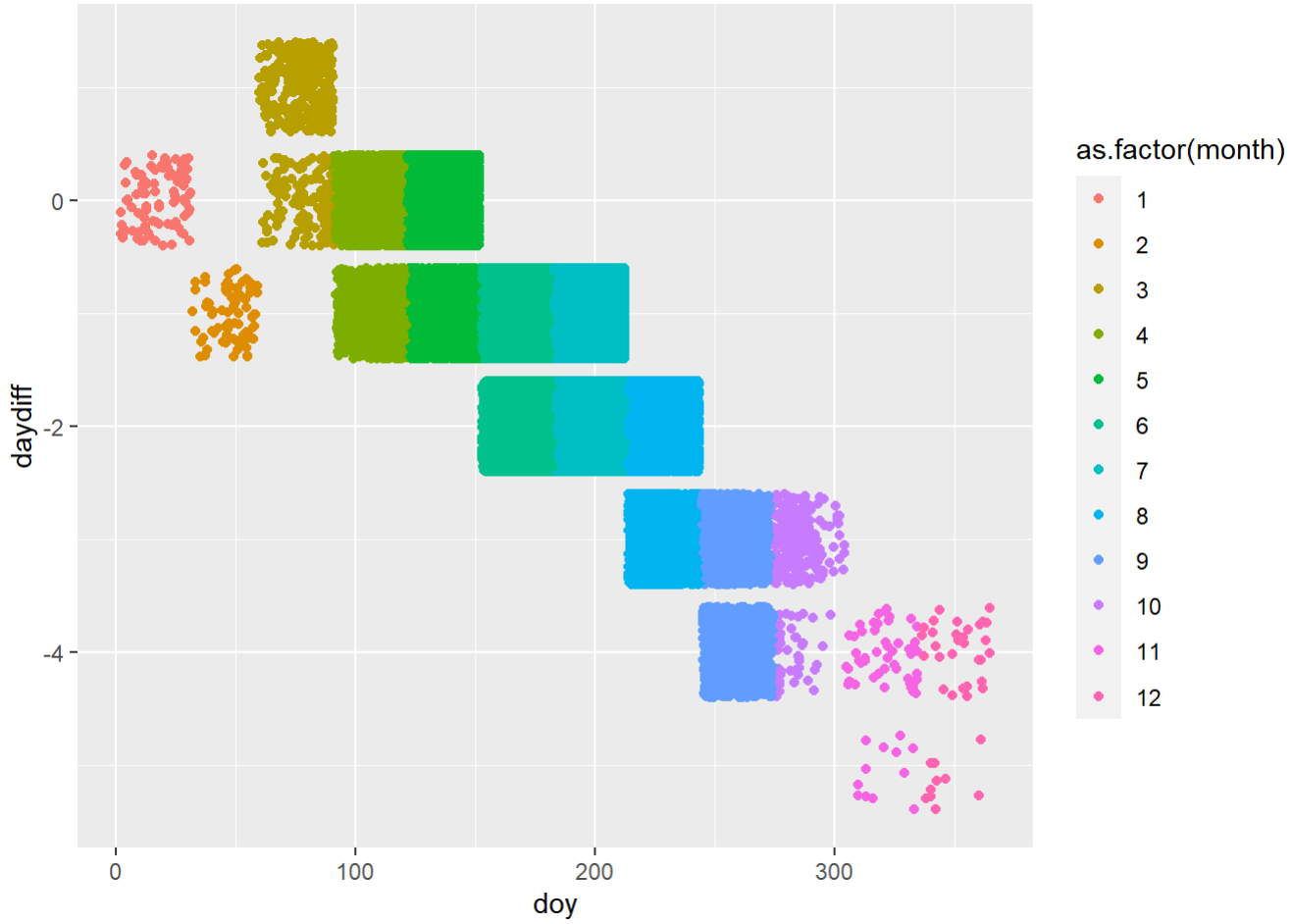
```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
alldata<-na.omit(alldata)
checkdays<-alldata %>%
  mutate(doy=yday(as.Date(paste(year,month,day, sep="-"), "%Y-%m-%d")),
         daydiff=SuccDay-doy, fricday=(month-1)*30+day) %>%
  select(name,day,month,year,SuccDay,doy,daydiff,fricday)
#summary(checkdays)
table(checkdays$fricday-checkdays$SuccDay)
```

```
##
##      0
## 282386
```

```
ggplot(data=checkdays, aes(y=daydiff, x=doy, color=as.factor(month))) + geom_jitter()
```




```
#we'd prefer to use calendar day
alldata<-alldata %>%mutate(doy=yday(as.Date(paste(year,month,day, sep="-"),"%Y-%m-%d")))
```

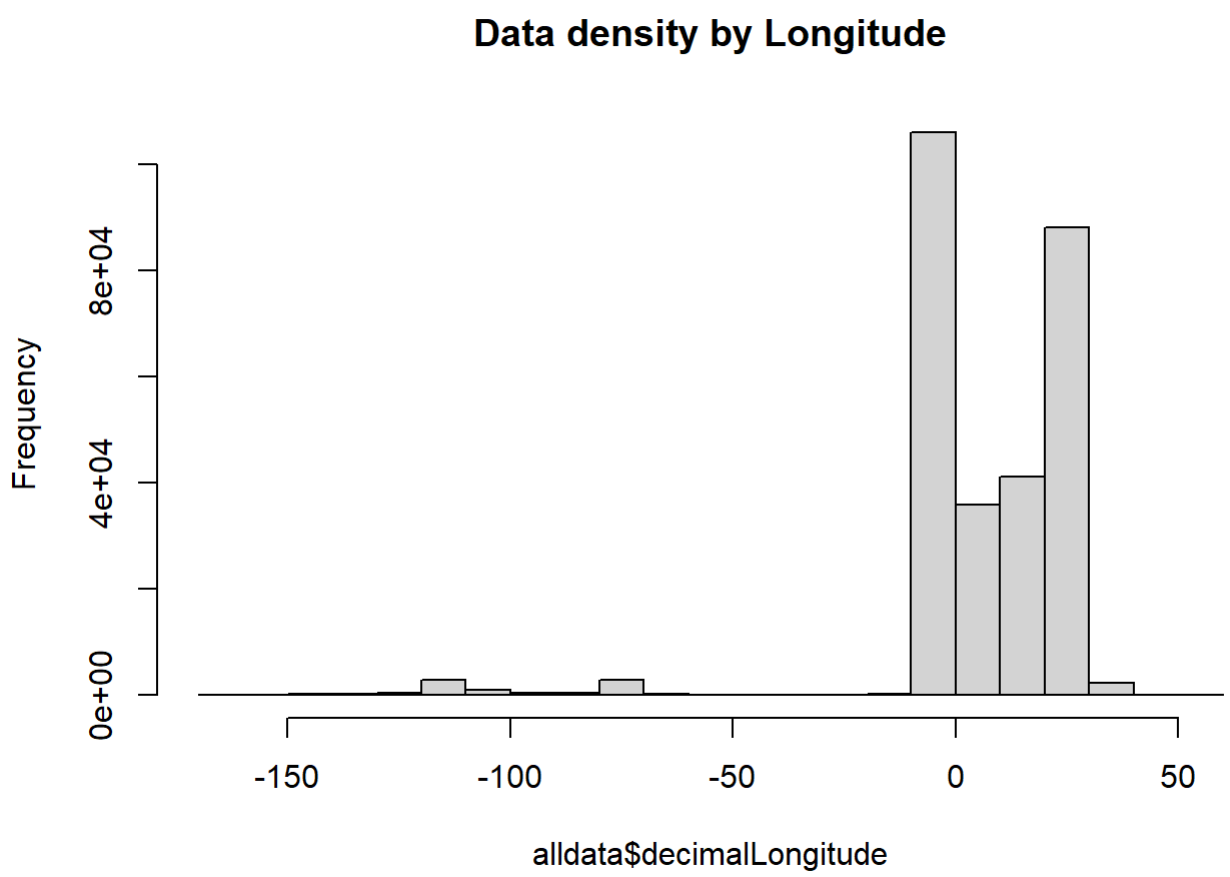
Data exploration 1

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:282386      Min.   :-162.559   Min.   : 5.787
##  1st Qu.: 2369   Class :character   1st Qu.:  -2.782   1st Qu.:52.784
##  Median : 7008   Mode  :character   Median :   9.398   Median :55.628
##  Mean   :14819                      Mean   :   6.317   Mean   :56.271
##  3rd Qu.:20216                      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:282386      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
##     SuccDay          rndLat           alt              region
##  Min.   :  2.0   Min.   : 6.00   Min.   :-2666.74   Length:282386
##  1st Qu.:163.0   1st Qu.:53.00   1st Qu.:   23.25   Class :character
##  Median :186.0   Median :56.00   Median :   64.33   Mode  :character
##  Mean   :181.7   Mean   :56.21   Mean   :  114.22
##  3rd Qu.:202.0   3rd Qu.:61.00   3rd Qu.:  111.09
##  Max.   :361.0   Max.   :71.00   Max.   : 4305.17
##       doy
##  Min.   :  2.0
##  1st Qu.:165.0
##  Median :187.0
##  Mean   :182.9
##  3rd Qu.:203.0
##  Max.   :365.0
```

```
##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"))

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

```
## [1] 105
```

```
#What are the names in the dataset?
data.names<-sort(unique(alldata$name))
#Which of these names shows up in the results?
result.names<-na.omit(read_excel("fric_supplements/ele13419-sup-0003-tables2.xlsx", sheet="~lati
tude", range="A3:A113"))
resultnames<-(strsplit(result.names$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   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##  [19]  19  20  21  22  23  25  26  27  28  29  30  31  32  33  34  35  36  37
##  [37]  38  39  40  41  42  44  45  46  47  48  49  50  51  53  54  55  57  58
##  [55]  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76
##  [73]  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94
##  [91]  95  96  97  98  99 101 102 103 104 105
```

```
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])))
```

```
## Warning in cbind(result.name = c(names_2, sort(unique(result.names))[c(26, :
## number of rows of result is not a multiple of vector length (arg 1)
```

```
print(name_changes)
```

```
##              result.name      data.name
## 1  Callophrys augustinus  Erynnis tages
## 2          Plebejus idas           <NA>
## 3        Phengaris arion           <NA>
## 4  Callophrys augustinus           <NA>
## 5          Plebejus idas           <NA>
## 6        Phengaris arion           <NA>
## 7  Callophrys augustinus Boloria selene
## 8          Plebejus idas           <NA>
## 9        Phengaris arion           <NA>
## 10 Callophrys augustinus           <NA>
## 11         Plebejus idas           <NA>
```

```
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:275081      Min.   :-162.559   Min.   : 5.787
##  1st Qu.: 2354   Class :character   1st Qu.:  -2.676   1st Qu.:52.829
##  Median : 7092   Mode  :character   Median :   9.564   Median :55.775
##  Mean   :15059                      Mean   :   6.778   Mean   :56.359
##  3rd Qu.:20846                      3rd Qu.:  23.763   3rd Qu.:60.677
##  Max.   :85273                      Max.   :  59.333   Max.   :71.216
##       year          month          country               day
##  Min.   :1616   Min.   : 1.000   Length:275081      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   : 6.501                      Mean   :16.19
##  3rd Qu.:2009   3rd Qu.: 7.000                      3rd Qu.:24.00
##  Max.   :2015   Max.   :12.000                      Max.   :31.00
##     SuccDay          rndLat          alt              region
##  Min.   :  2.0   Min.   : 6.0   Min.   :-2666.74   Length:275081
##  1st Qu.:164.0   1st Qu.:53.0   1st Qu.:   23.25   Class :character
##  Median :186.0   Median :56.0   Median :   64.24   Mode  :character
##  Mean   :181.2   Mean   :56.3   Mean   :  113.22
##  3rd Qu.:201.0   3rd Qu.:61.0   3rd Qu.:  110.64
##  Max.   :361.0   Max.   :71.0   Max.   : 4305.17
##       doy
##  Min.   :  2.0
##  1st Qu.:165.0
##  Median :187.0
##  Mean   :182.5
##  3rd Qu.:203.0
##  Max.   :365.0
```

```
#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

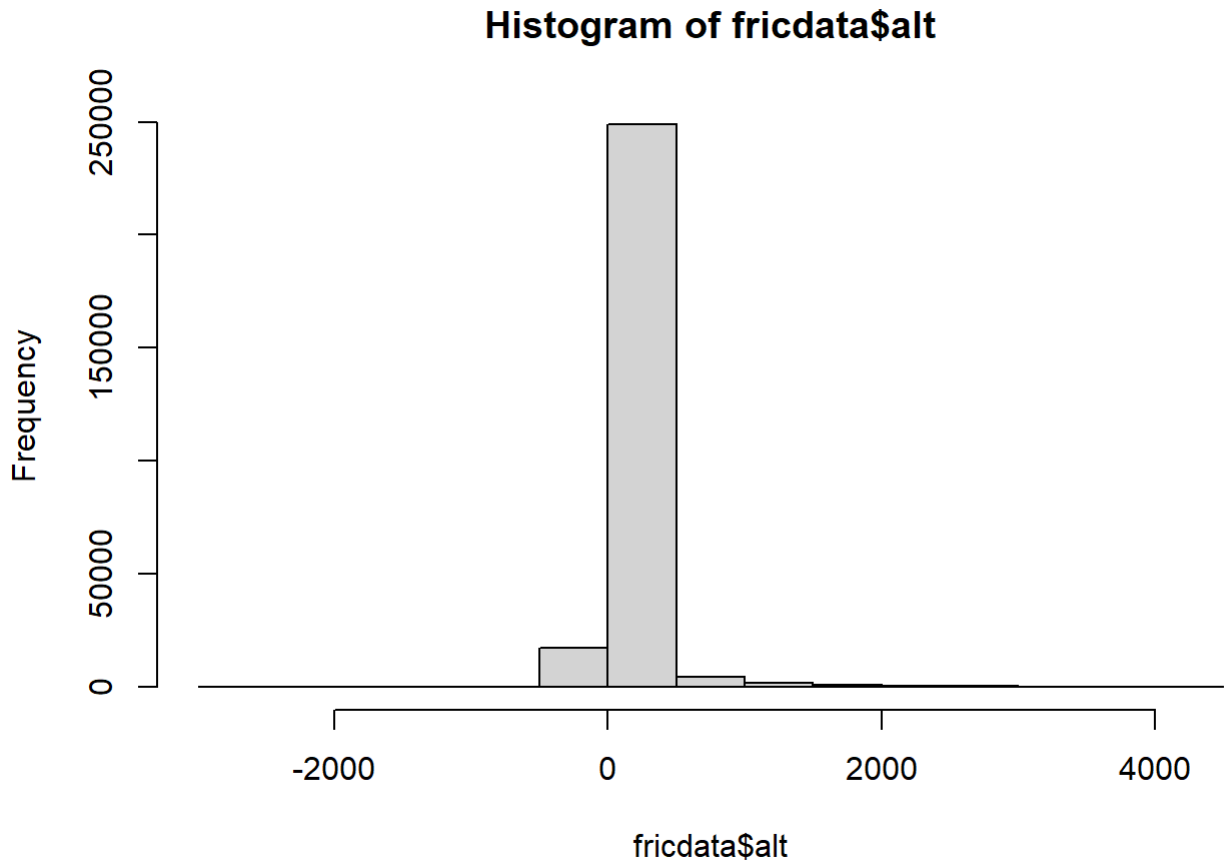
- 1. GIS coordinates had often been assigned by placename, or were otherwise inaccurate, and
- 2. 2. altitudes obtained by using the Google API to extract altitude for coordinates did not provide reliable altitudes for the underlying occurrences.

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   113.22   110.64  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] "10920 records below sea level represent 3.97 percent of all ocurrence records. We examin
ed 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=" "))
```

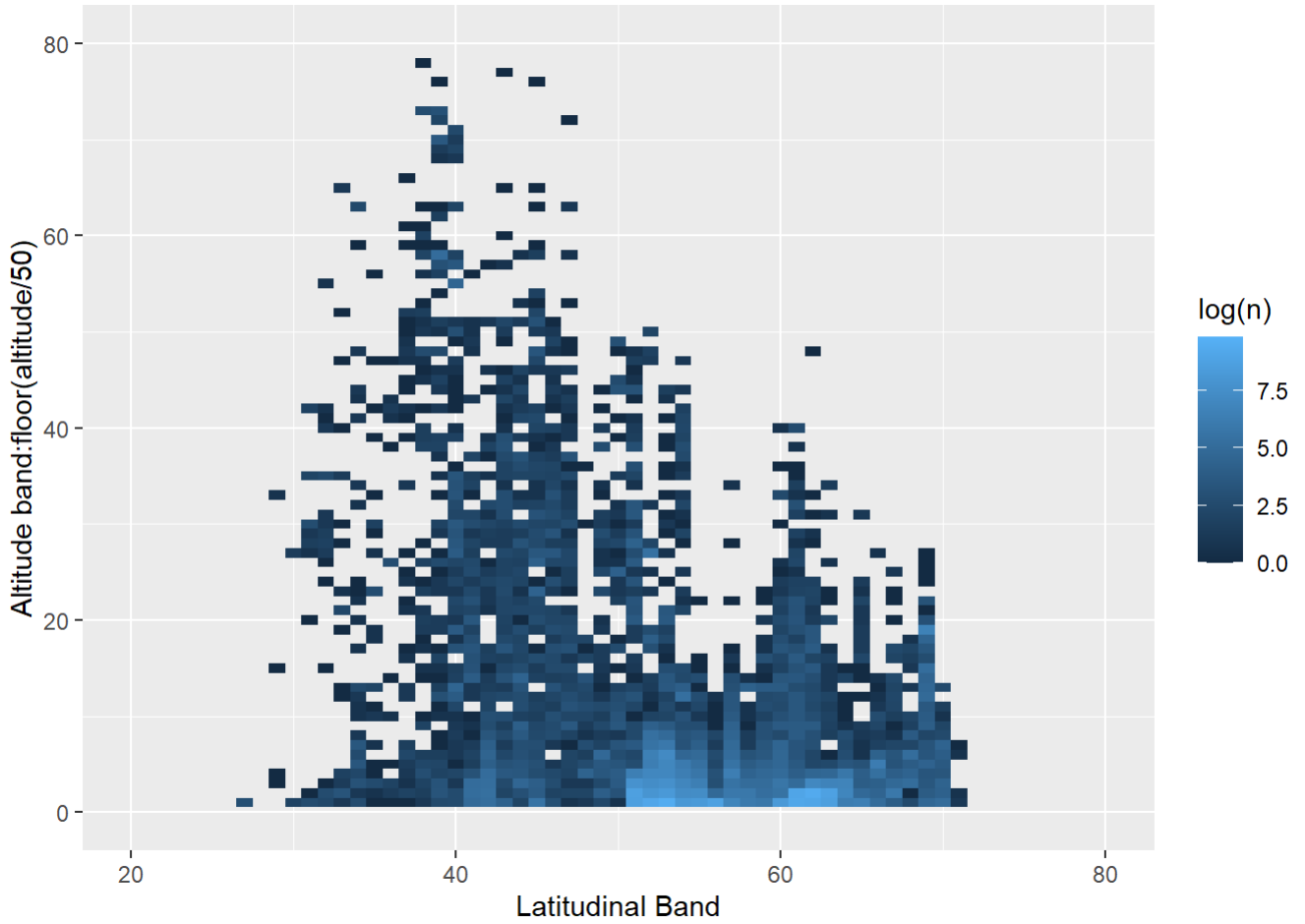
```
## [1] "8864 records above 500m represent 3.22 percent of all ocurrence records. We examined la
t/long and location for a small subset of high altitude records and found vague place names had
been used for geolocation."
```

```
#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] "255297 records  within 0-500m represent 92.81 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?

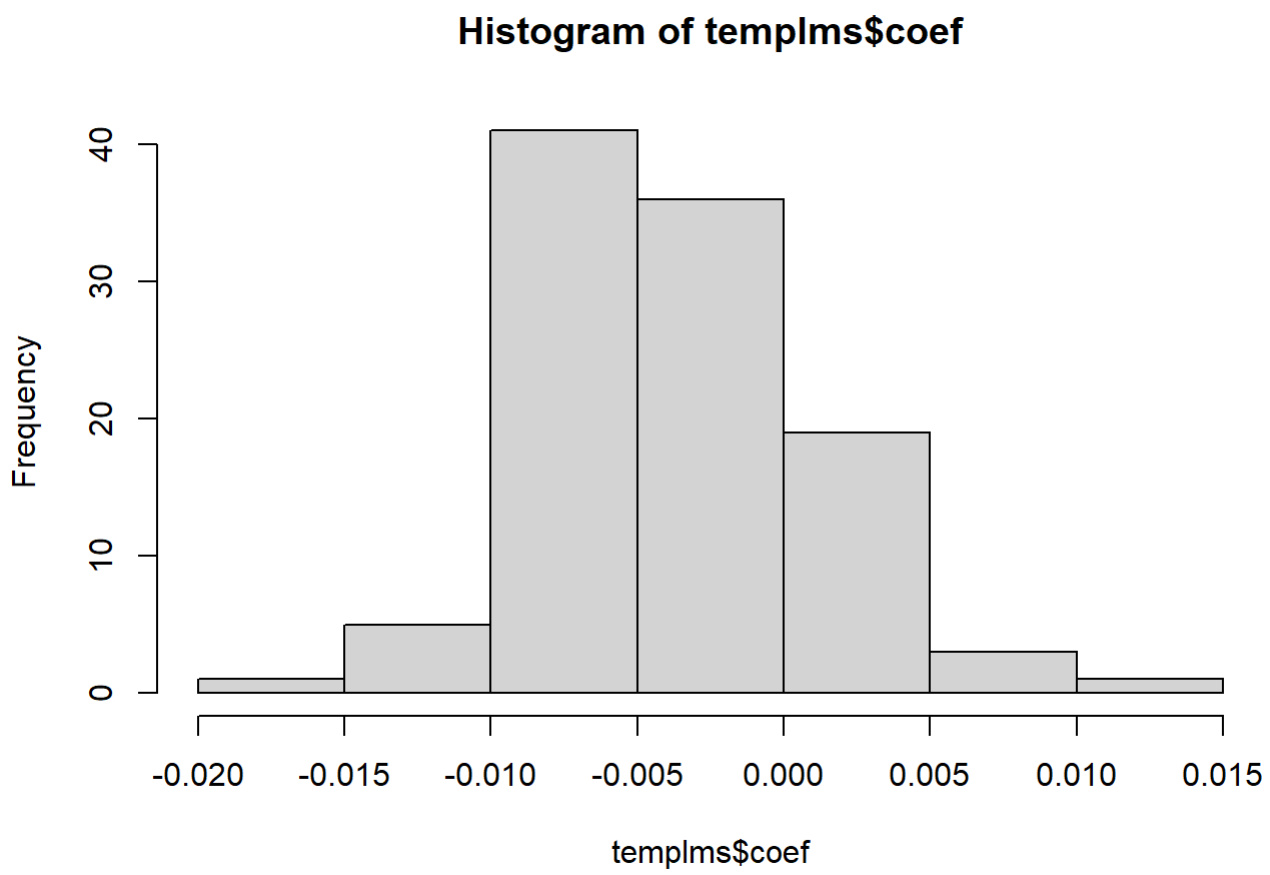
- A Lycaena hippothoe record from 1995 (row.index 2160; altitude 3274m) is from https://www.gbif.org/occurrence/2570253925 (https://www.gbif.org/occurrence/2570253925) which lists an inferred elevation of 2000m.
- A Lycaena virgaureae record from 2002 (row.index 4501; altitude -85.8m) appears to match 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.   :0.000000   Min.   :0.0000222
##  1st Qu.:   79   1st Qu.:-0.006851   1st Qu.:0.000000   1st Qu.:0.0277334
##  Median :  189   Median :-0.004470   Median :0.000000   Median :0.1923171
##  Mean   : 2595   Mean   :-0.003843   Mean   :0.071131   Mean   :0.2799456
##  3rd Qu.: 1108   3rd Qu.:-0.001142   3rd Qu.:0.006875   3rd Qu.:0.5244507
##  Max.   :51819   Max.   : 0.014623   Max.   :0.859071   Max.   :0.8487862
```

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

```
## [1] 0.84
```

```
#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 percent of all datasets. For
datasets with significant collinearity, the mean coefficient is -0.005 (which translates to a sl
ope of -222 meters per degree latitude) and mean r-squared is 0.33  - therefore regression of re
siduals 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)
```

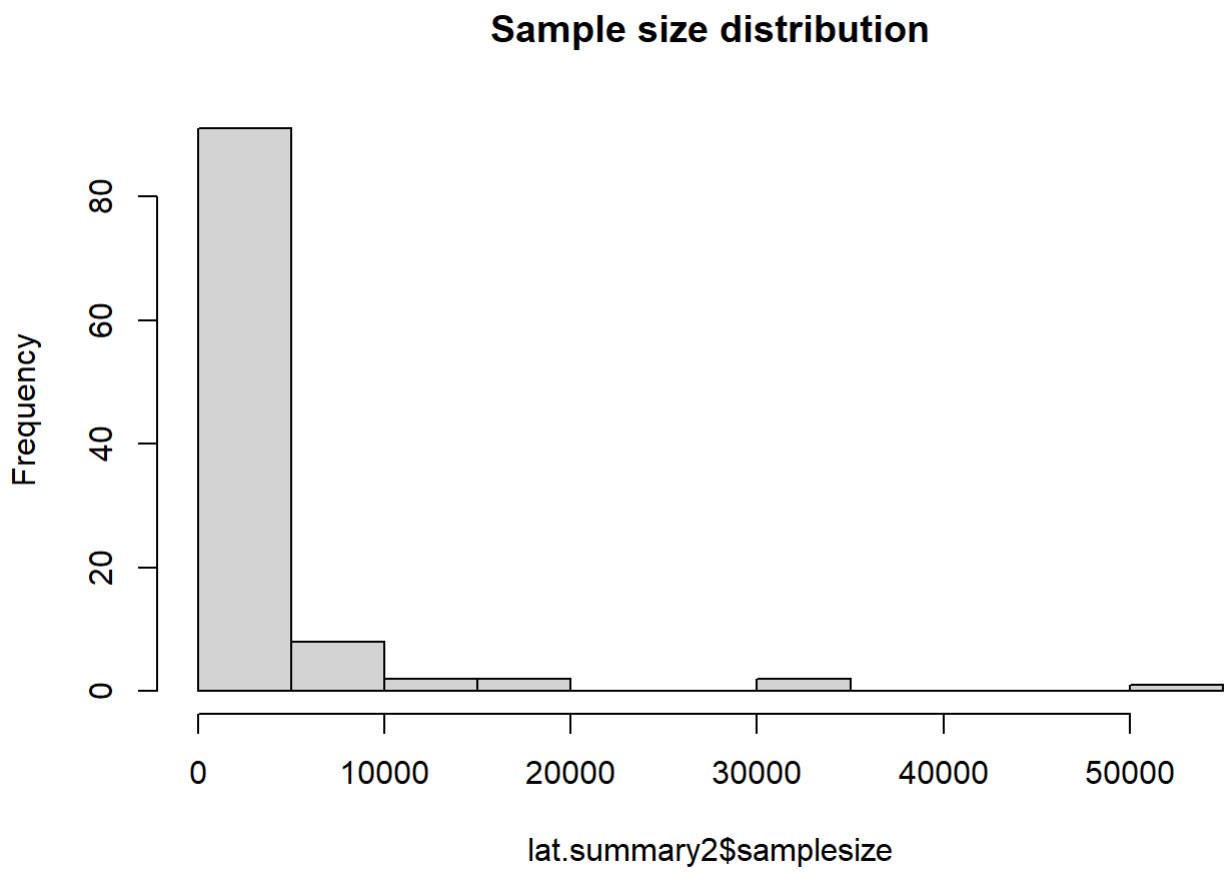
```
lat.summary2<-lat.summary1 %>%
  group_by(name,region) %>%
  summarize(samplesize=sum(lat.samplesize),latspan=max(rndLat)-min(rndLat),nlats=length(unique(r
ndLat)),n.singletons=sum(singleton),prop.singletons=n.singletons/nlats)
```

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

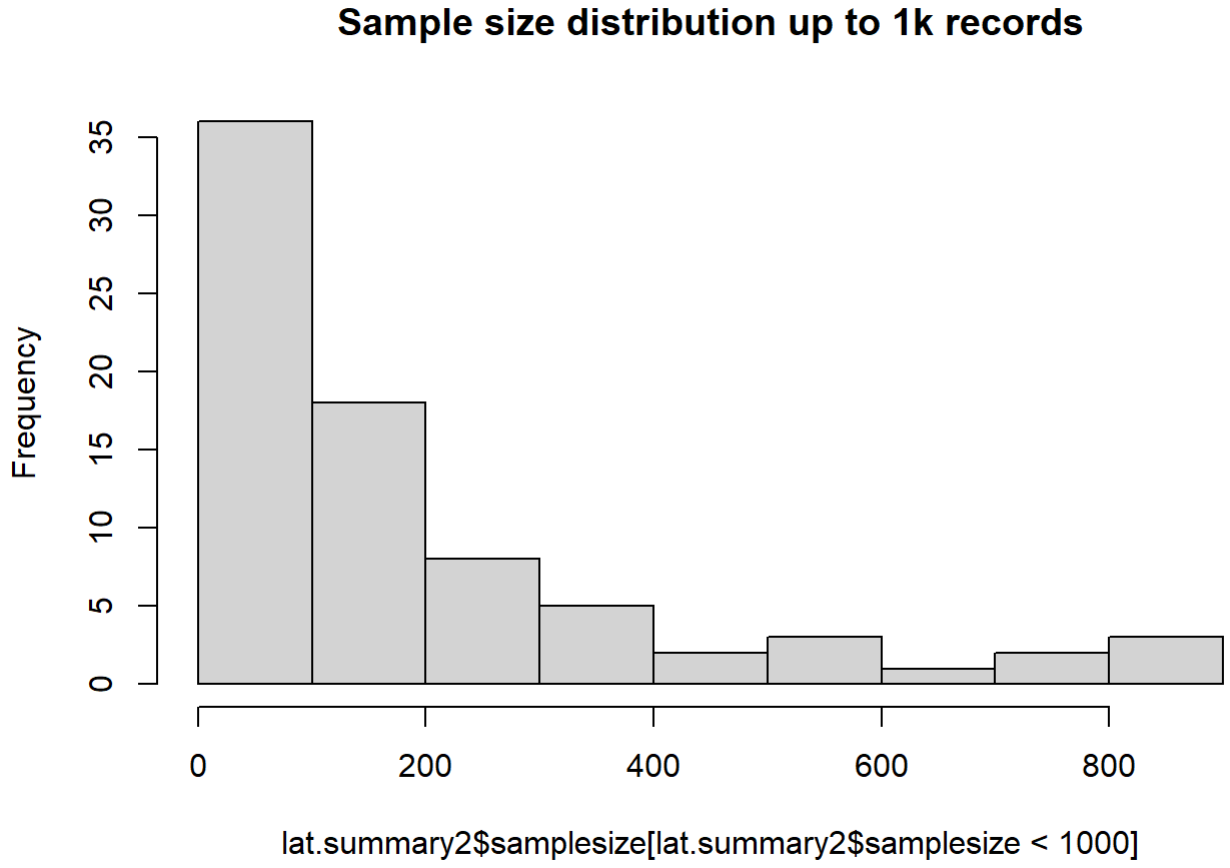
```
summary(lat.summary2)
```

```
##      name              region            samplesize       latspan
##  Length:106         Length:106         Min.   :   15   Min.   :10.00
##  Class :character   Class :character   1st Qu.:   79   1st Qu.:24.00
##  Mode  :character   Mode  :character   Median :  189   Median :27.00
##                                        Mean   : 2595   Mean   :26.35
##                                        3rd Qu.: 1108   3rd Qu.:30.00
##                                        Max.   :51819   Max.   :64.00
##      nlats        n.singletons    prop.singletons
##  Min.   : 5.00   Min.   : 0.000   Min.   :0.00000
##  1st Qu.:13.00   1st Qu.: 2.000   1st Qu.:0.09412
##  Median :18.00   Median : 3.000   Median :0.18634
##  Mean   :19.05   Mean   : 3.415   Mean   :0.20721
##  3rd Qu.:25.75   3rd Qu.: 5.000   3rd Qu.:0.32955
##  Max.   :33.00   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")
```



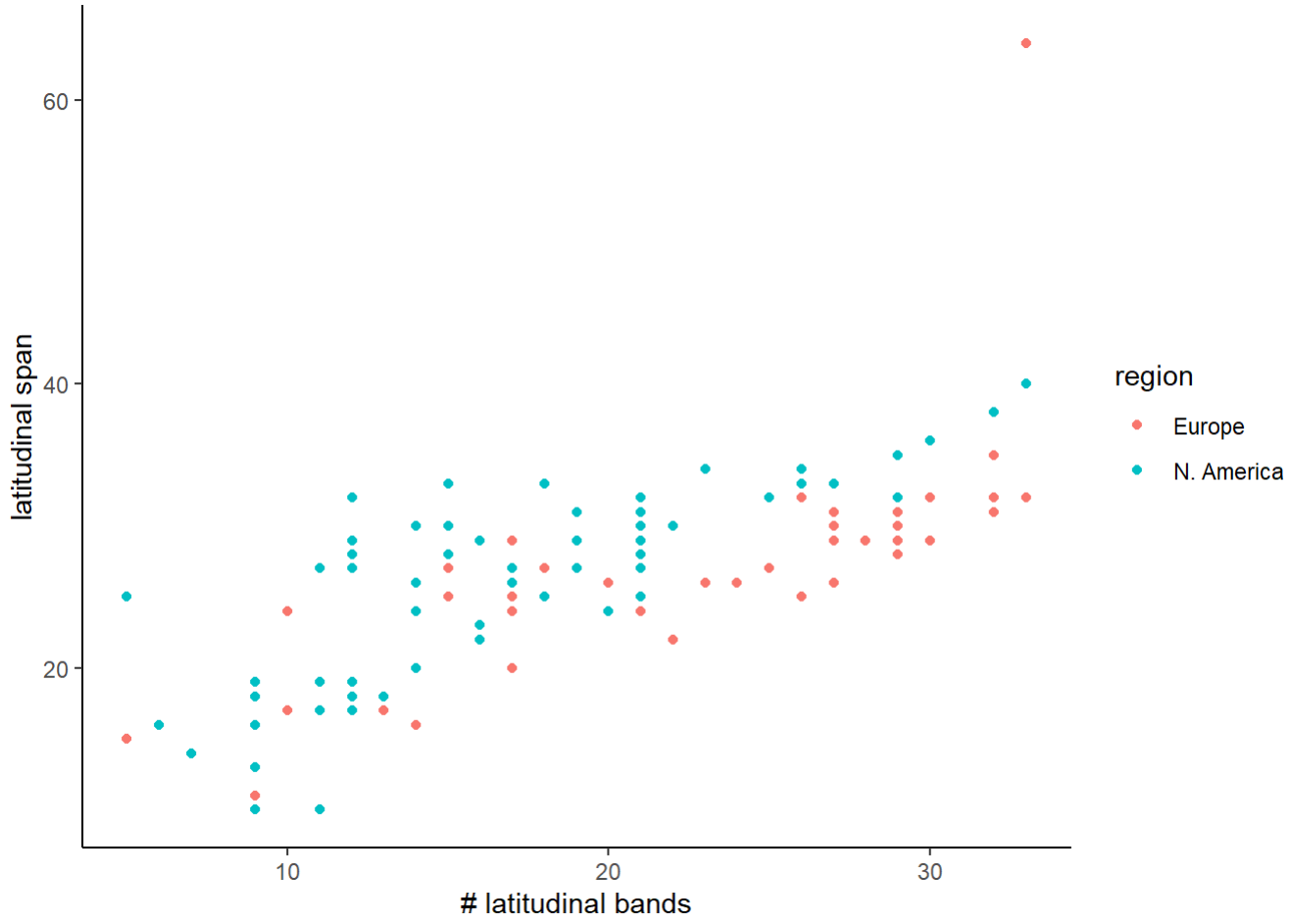

```
nrow(lat.summary2 %>% filter(samplesize<100))
```

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

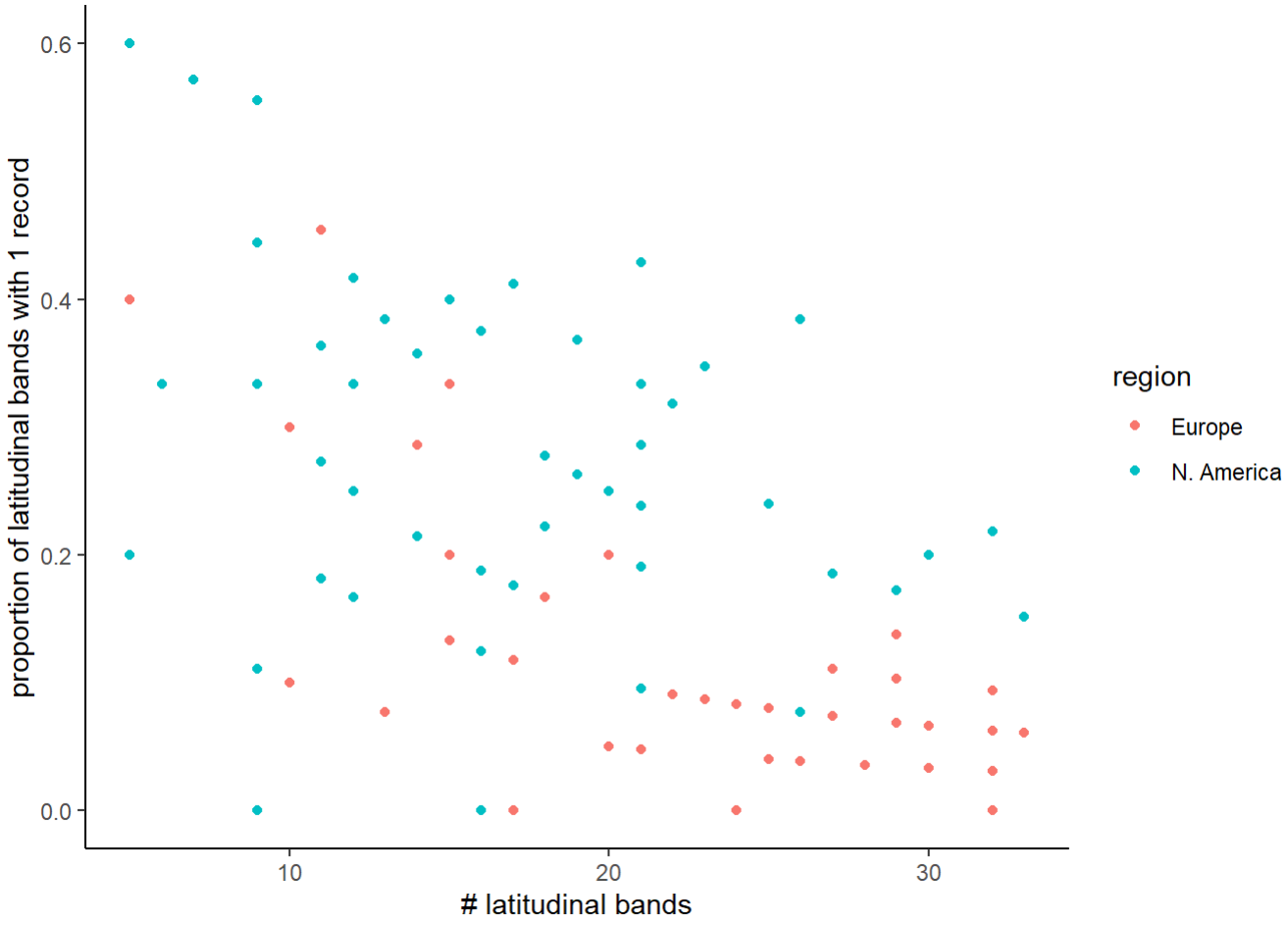
```
print(paste(nrow(lat.summary2 %>% filter(samplesize<100)),"datasets have less than 100 ocurrence records."))
```

```
## [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")
```



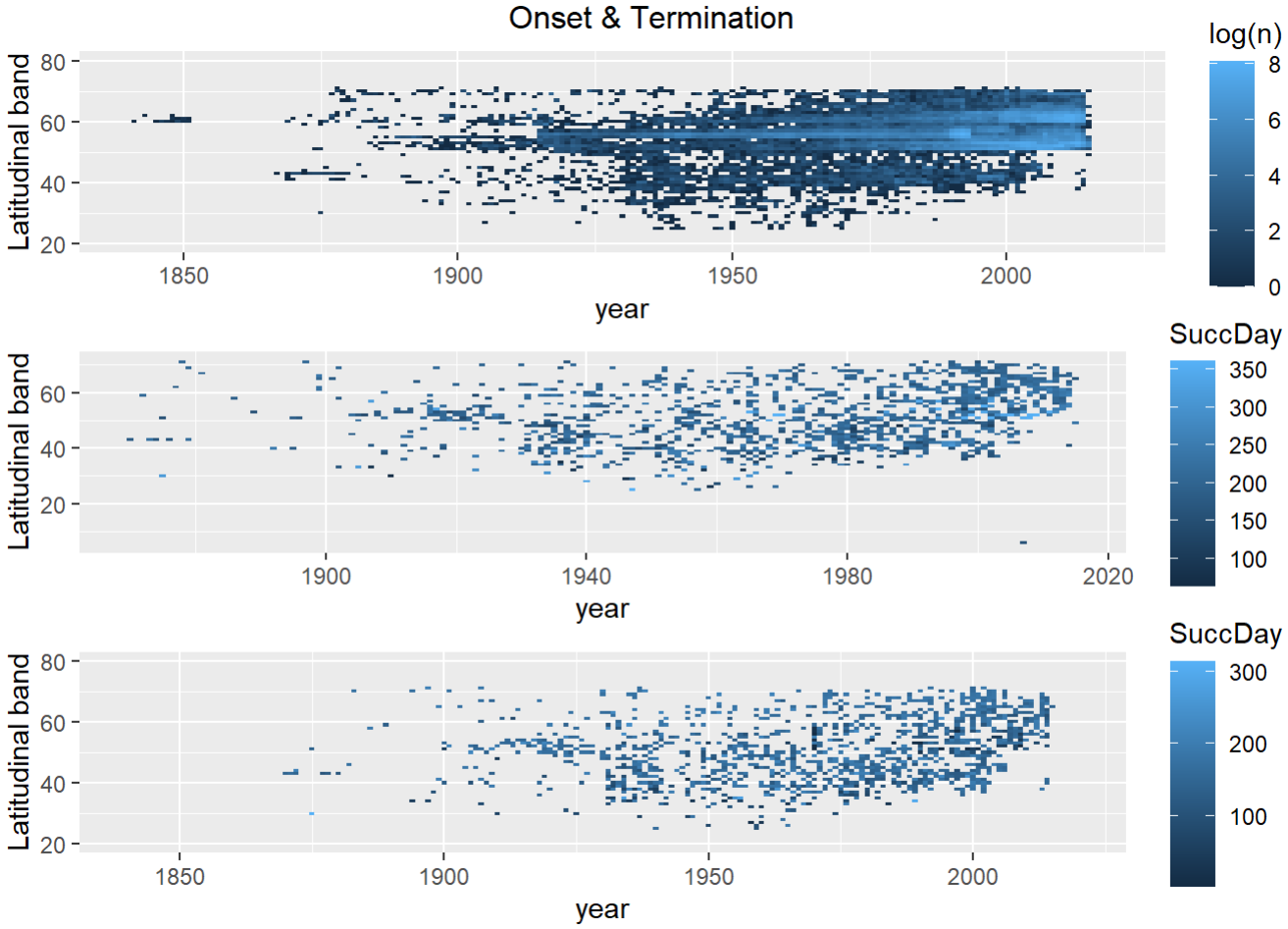

```
ggplot(data=lat.summary2, aes(x=nlats, y=prop.singletons, color=region)) + geom_point() + theme_
classic() +
  labs(x="# latitudinal bands", y="proportion of latitudinal bands with 1 record")
```

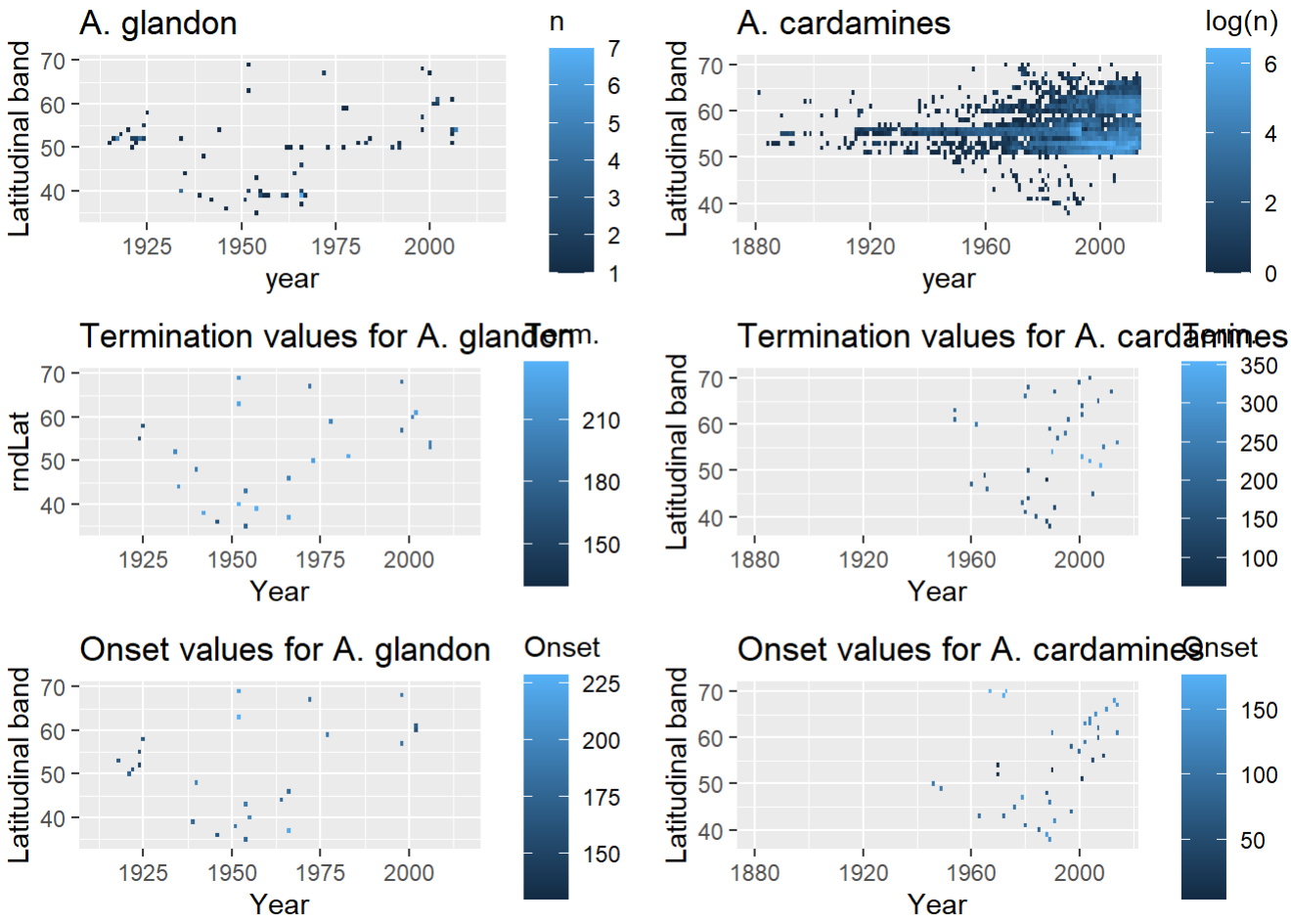


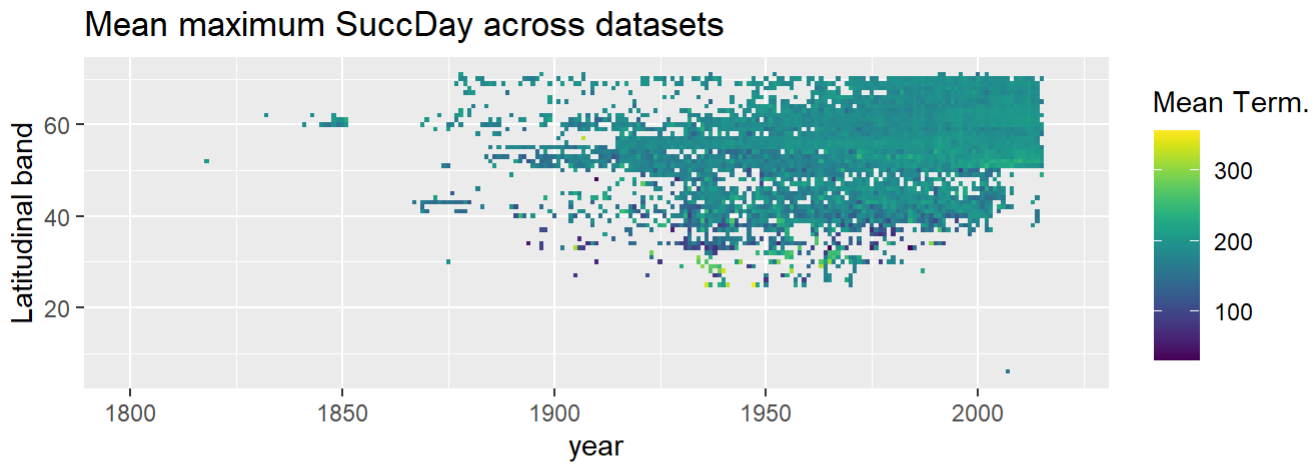
Data exploration: year

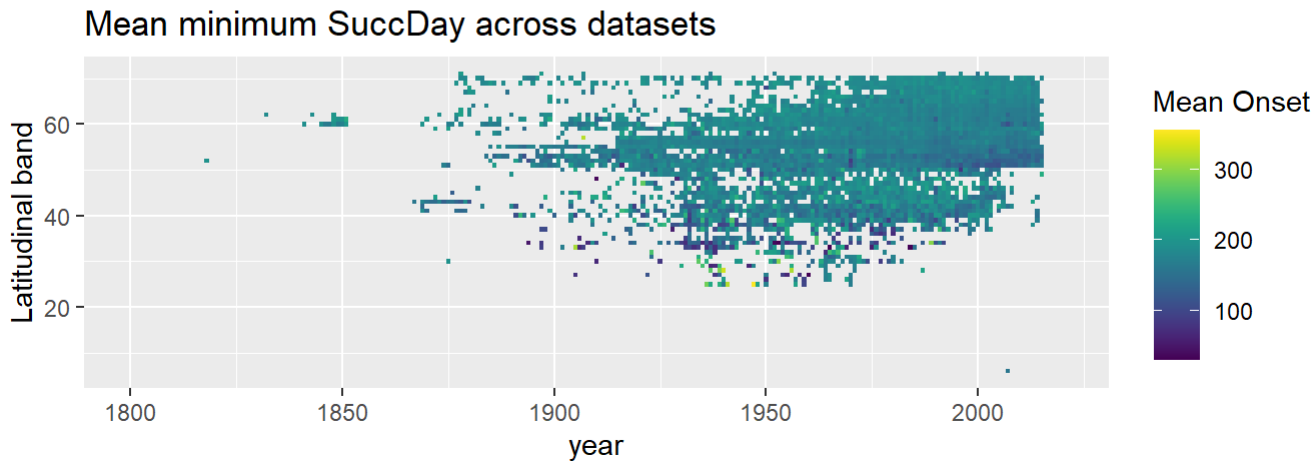
As expected, most data are quite recent. By selecting the min and max day of year per latitudinal band as onset & termination, the authors vastly decrease their sample size and remove most of the variation along the year and altitude axes

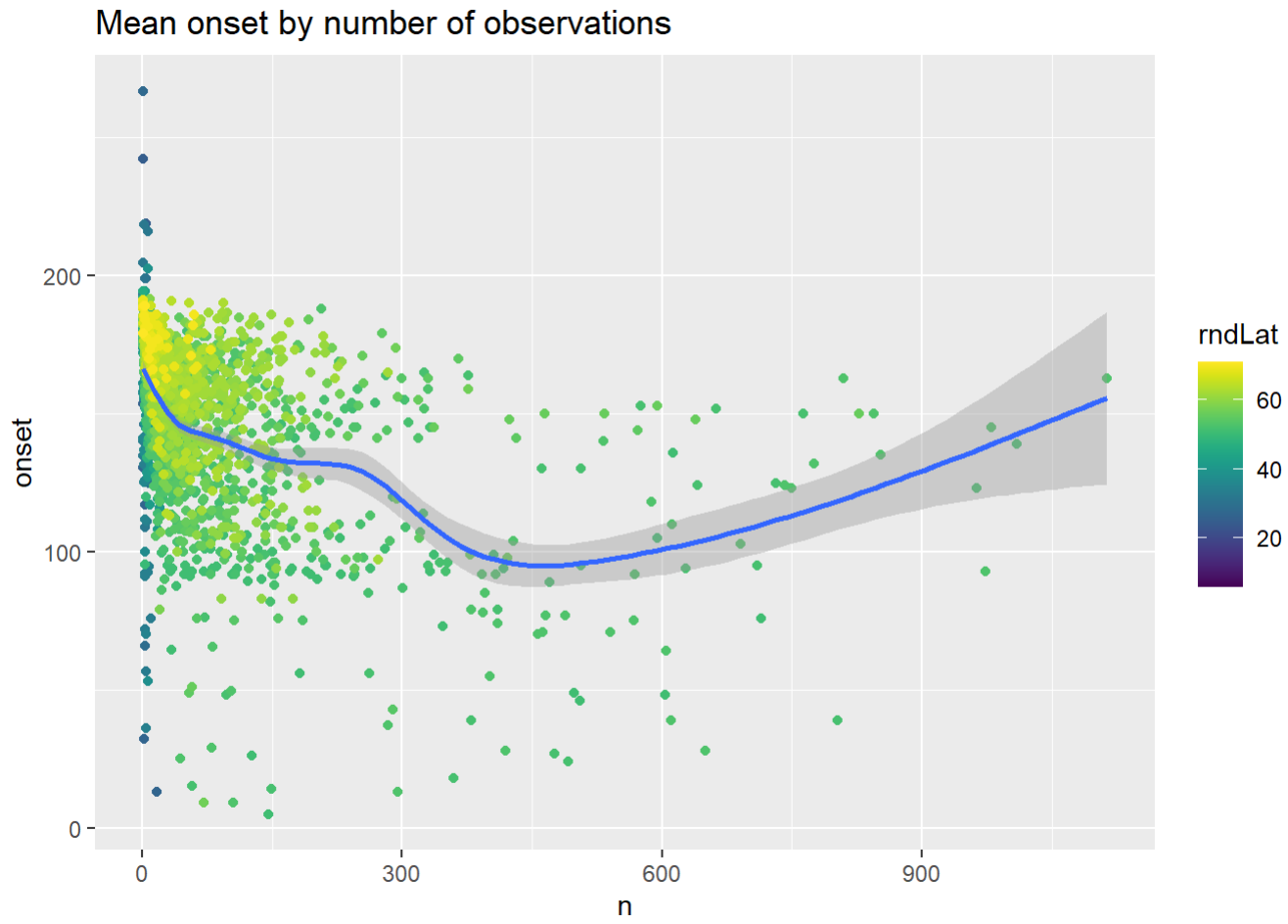
We arbitrarily selected two species, one with a low sample size and one with a large sample size, to visualize.

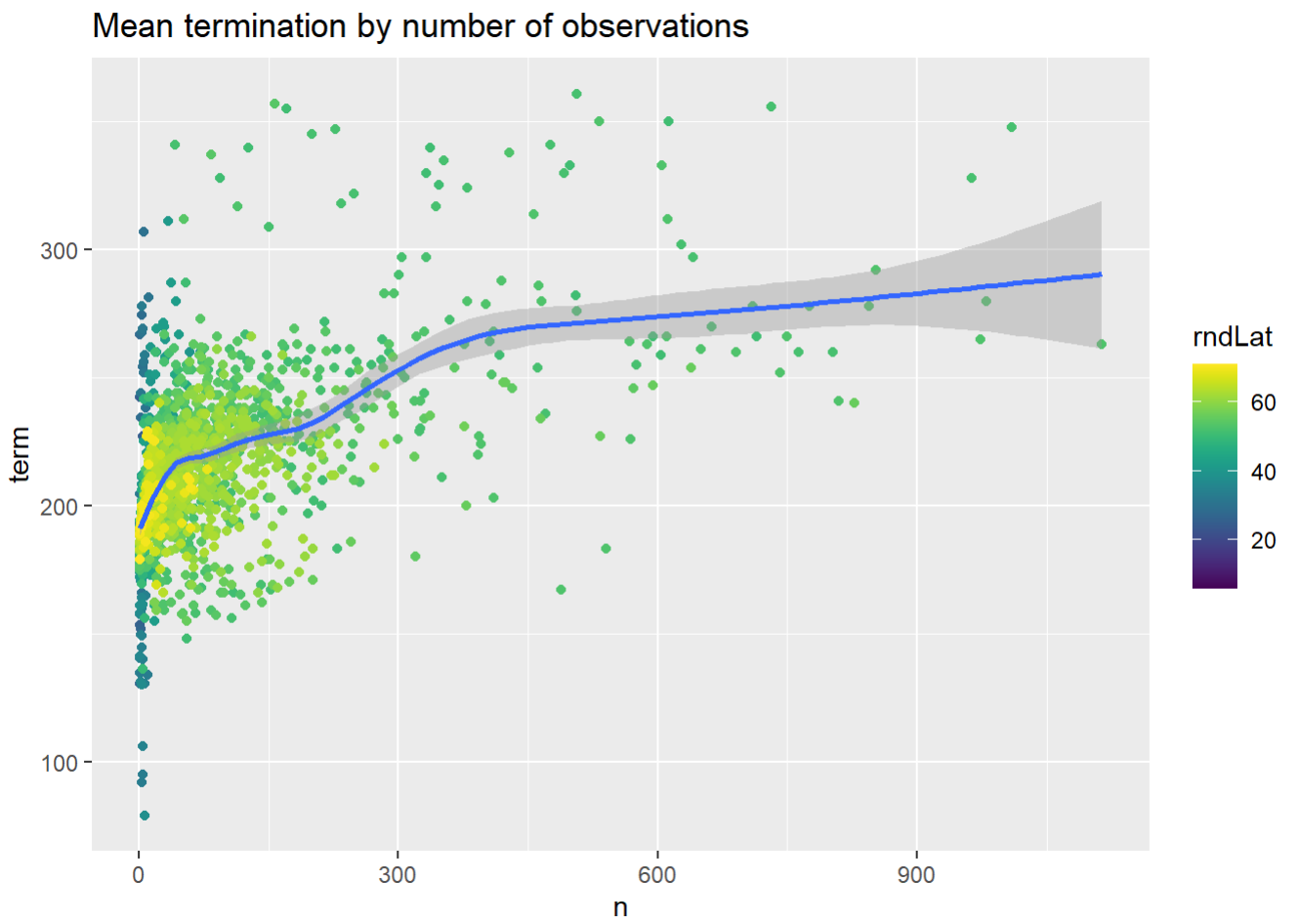





Mean maximum SuccDay across datasets
Latitudinal band
60
40
20
1800
1850
1900
1950
2000
year
Mean Term.
300
200
100


Mean minimum SuccDay across datasets
Latitudinal band
60
40
20
1800
1850
1900
1950
2000
year
Mean Onset
300
200
100


Mean onset by number of observations
onset
200
100
0
0
300
600
900
n
rndLat
60
40
20


Mean termination by number of observations
term
300
200
100
0
300
600
900
n
rndLat
60
40
20

End of File.