Metrics for Recorder behaviour

Tom August

26 September 2016

Metrics

We are going to split metric into three broad groups: *Engagement profile*, *Spatial*, and *Taxanomic*

Temporal Metrics

These metrics measure the recording pattern across time such as the number of days that a recorder produces records. These have been termed engagement profiles by others., The metrics here are from Ponciano and Brasileiro 2014 who used the metrics on participant of zooniverse projects. The metrics were also used by Boakes *et al* 2016.

Activity ratio

"The proportion of days on which the volunteer was active in relation to the total days he/she remained linked to the project" (Ponciano and Brasileiro 2014)

```
# Create a function to calculate activity ratio
activityRatio <- function(recorder_name,
                          data,
                          recorder_col = 'recorders',
                          date_col = 'date_start',
                          format = '%d/%m/%Y'){

  # Get the recorders data
  data <- data[data[,recorder_col] == recorder_name, ]

  # Get unique dates as dates
  dates <- as.Date(as.character(unique(data[,date_col])), format = format)

  # Get the first and last date
  first_last <- range(dates)

  # Total duration of this recorder
  duration <- as.numeric(first_last[2] - first_last[1]) + 1

  # calculate ratio
  activity_ratio <- length(dates)/duration

  # return
  return(data.frame(recorder = recorder_name,
                    activity_ratio = activity_ratio,
                    total_duration = duration,
                    active_days = length(dates)))

}
```

```r
# Test on David and Tom
activityRatio(data = iRB, recorder_name = 'Roy, David')
```

```
##     recorder activity_ratio total_duration active_days
## 1 Roy, David        0.13382           1233         165
```

```r
activityRatio(data = iRB, recorder_name = 'August, Tom')
```

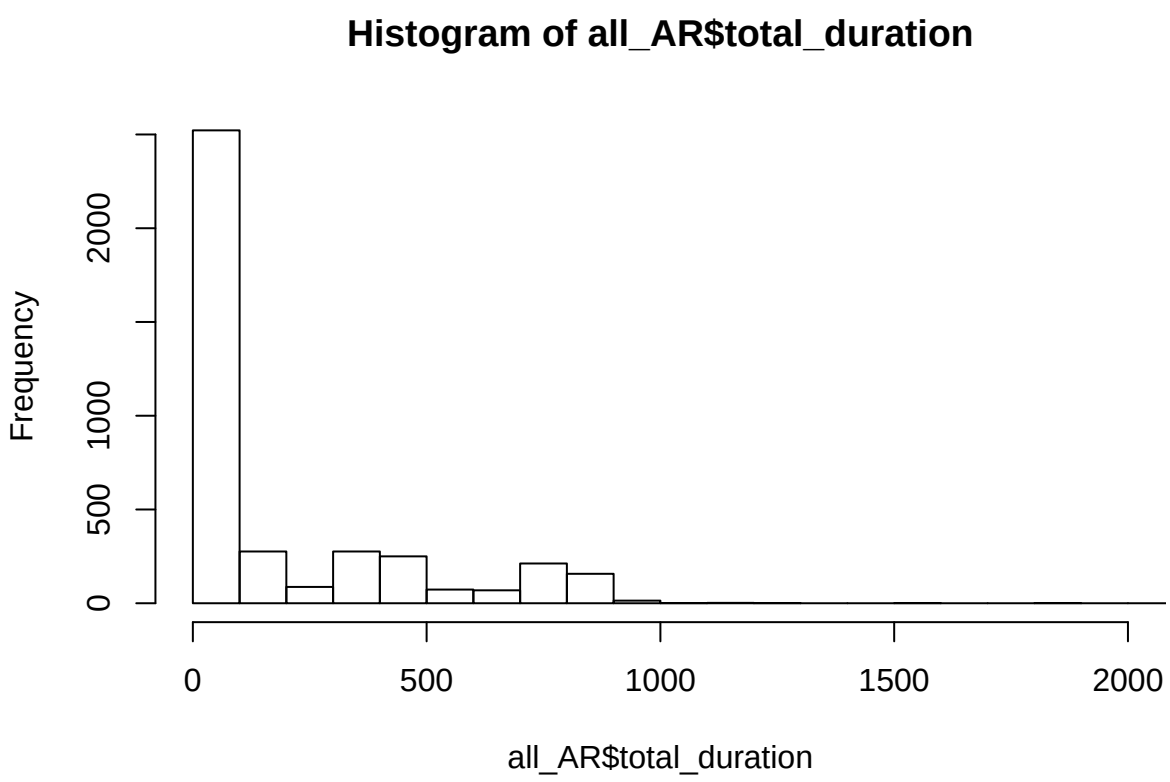
```
##      recorder activity_ratio total_duration active_days
## 1 August, Tom     0.02678571            448          12
```

```r
## David is a more active recorder than Tom ##

# Run for everyone
all_AR <- do.call(rbind, lapply(X = unique(iRB$recorders),
                                FUN = activityRatio,
                                data = iRB))

# Lets have a look at some of these
head(all_AR, 20)
```

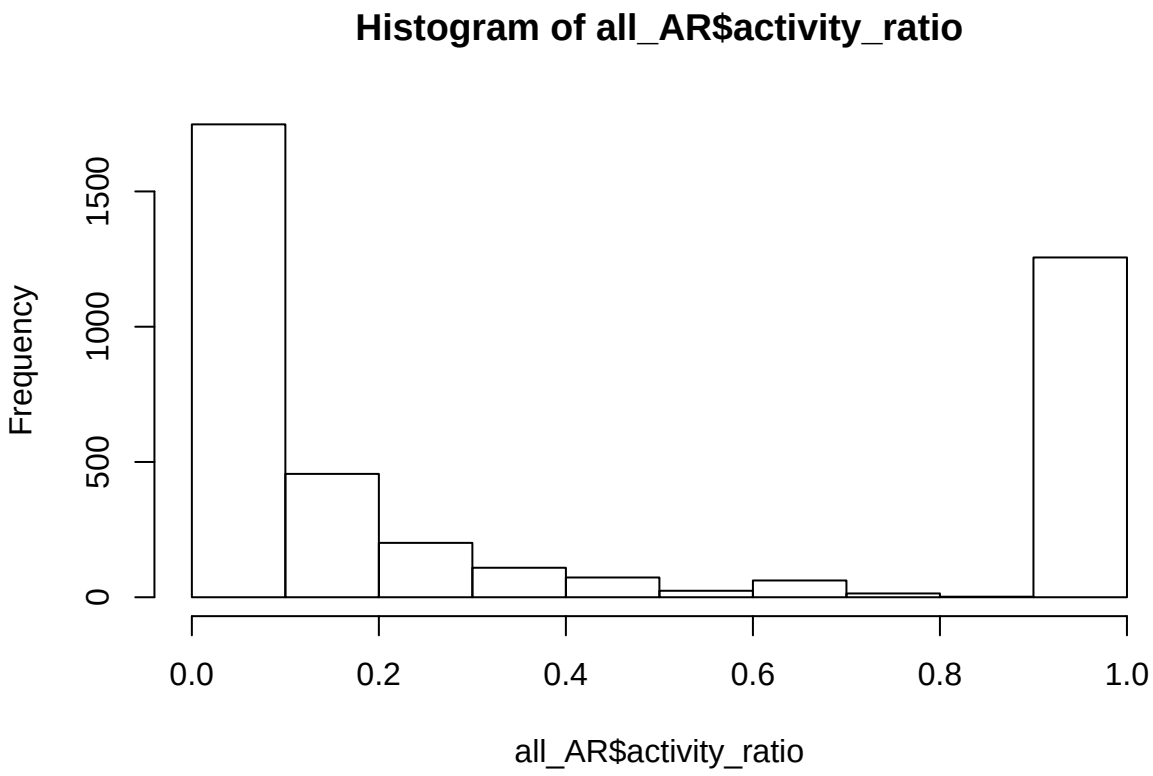
```
##                  recorder activity_ratio total_duration active_days
## 1         Brookes , Anne      0.08383234            835          70
## 2         Burgoyne, Steve     0.32876712            146          48
## 3            Brown, Peter     0.09090909            715          65
## 4      Rutherford, Joanna     0.01716247            874          15
## 5            Allan, David     0.40601504            931         378
## 6        Millward, Martin     0.08000000             50           4
## 7  Foulkes-Arellano, Paul     0.02192448            821          18
## 8            Stewart, Tam     0.30088496            904         272
## 9          Forbes, Andrew     0.11538462             52           6
## 10      Richardson, Rosie     0.28000000             25           7
## 11   Partridge, Francesca     0.22208738            824         183
## 12         Card , Graeme     0.05378973            409          22
## 13            Honey, Hawk     0.09239766            855          79
## 14         Melzack, David     0.31250000             96          30
## 15             Povall, Ed     0.10000000             30           3
## 16          Goodwin, Paul     0.15813953            430          68
## 17           Coulson, Joe     0.18965517            116          22
## 18          Bailey, Peggy     0.29577465            142          42
## 19             Roy, David     0.13381995           1233         165
## 20      Woodley, Caroline     0.14705882            136          20
```

I think this metric tells a story in a combination of the ratio and the total number of days. I think the ratio means more when the recorder has been recording for a long duration

```r
# Have a look at the distribution of these 2 metrics
hist(all_AR$total_duration, xlim = c(0, 2000), breaks = 100)
```

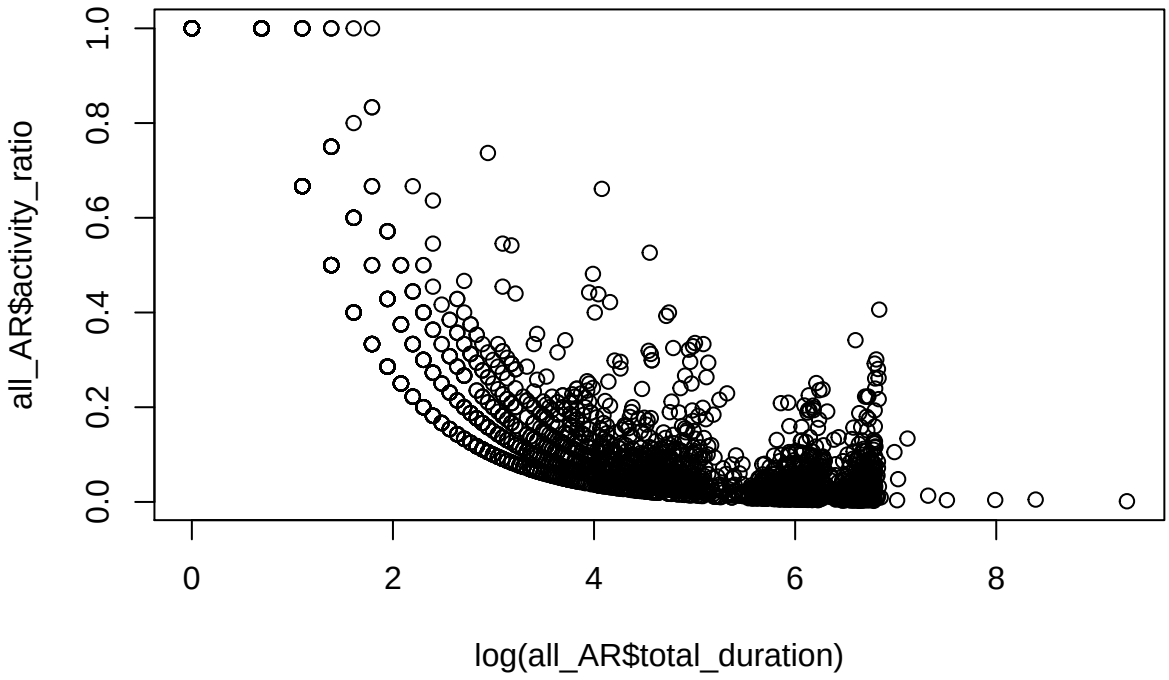



```
hist(all_AR$activity_ratio)
```



Both have nice distributions, though we can see the single record people in the ratio plot

```
# Plot activity_ratio against duration
plot(log(all_AR$total_duration),
     all_AR$activity_ratio)
```



Yearly devoted days

This is an adaptation of the *Daily Devoted Time* in (Ponciano and Brasileiro 2014) which is clearly not applicable to biological recording. Though Boakes *et al* 2016 don't attempt to use this measure I think the idea can be adapted by using days in a year rather than hours in a day. If we want to get a really good number we should probably remove the first and last year of recording as the user probably joined and left part way though the year, but I don't do that here.

```
# Create a function
yearlyDevotedDays <- function(recorder_name,
                              data,
                              recorder_col = 'recorders',
                              date_col = 'date_start',
                              format = '%d/%m/%Y'){

  # Get the recorders data
  data <- data[data[,recorder_col] == recorder_name, ]

  # Get unique dates as dates
  dates <- as.Date(as.character(unique(data[,date_col])), format = format)

  # here are the counts
  year_counts <- table(as.numeric(format(dates, '%Y')))

  # As these are counts taking the median is probably best
```

```
  yearly_devoted_days <- median(year_counts)

  return(data.frame(recorder = recorder_name,
                    yearly_devoted_days = yearly_devoted_days,
                    n_years = length(year_counts),
                    n_recs = sum(year_counts)))
}

# Test on David and Tom
yearlyDevotedDays(data = iRB, recorder_name = 'Roy, David')
```

```
##     recorder yearly_devoted_days n_years n_recs
## 1 Roy, David                48.5       4    165
```

```
yearlyDevotedDays(data = iRB, recorder_name = 'August, Tom')
```

```
##      recorder yearly_devoted_days n_years n_recs
## 1 August, Tom                   6       2     12
```

```
## David contributes more of his time than Tom ##

# Run for everyone
all_YDD <- do.call(rbind, lapply(X = unique(iRB$recorders),
                                 FUN = yearlyDevotedDays,
                                 data = iRB))

# Lets have a look at some of these
head(all_YDD, 20)
```

```
##                     recorder yearly_devoted_days n_years n_recs
## 2016          Brookes , Anne                25.0       3     70
## 20161         Burgoyne, Steve               48.0       1     48
## 2015             Brown, Peter               11.0       3     65
## 20162      Rutherford, Joanna                4.0       3     15
## 2014             Allan, David              121.0       3    378
## 20163        Millward, Martin                4.0       1      4
## 20164  Foulkes-Arellano, Paul                7.0       3     18
## 20151            Stewart, Tam               95.0       3    272
## 20165          Forbes, Andrew                6.0       1      6
## 20166       Richardson, Rosie                7.0       1      7
## 20141    Partridge, Francesca               61.0       3    183
## 1              Card , Graeme                11.0       2     22
## 20167             Honey, Hawk               29.0       3     79
## 20168          Melzack, David               30.0       1     30
## 20169              Povall, Ed                3.0       1      3
## 11               Goodwin, Paul               34.0       2     68
## 201610            Coulson, Joe               22.0       1     22
## 201611           Bailey, Peggy               42.0       1     42
## 12                  Roy, David               48.5       4    165
## 201612      Woodley, Caroline                20.0       1     20
```

Clearly this metric is only really reliable when we have multiple years worth of data for an individual, which by definition takes them out of the 'casual user' type category.

Relative activity duration

This is a metric from Ponciano and Brasileiro 2014 which is also used in Boakes *et al* 2016 but I don't think can be applied to biological records since there is no official end date for a project: *"The ratio of days during which a volunteer i remains linked to the project in relation to the total number of days elapsed since the volunteer joined the project until the project is over"

Periodicity

There is a cluster of metrics that could be used to look at aspects of periodicity. The measure used in Ponciano and Brasileiro 2014 is 'variation in periodicity'; *"The standard deviation of the times elapsed between each pair of sequential active days"*. At the same time as calculating this I think there are another couple of metrics that might be of use. First, periodicity itself, i.e. "The median time elapsed between each pair of sequential active days". Secondly, streak length, i.e. "The average length of sequential active days"

```
# Create a function to calculate the periodicity metrics
periodicity <- function(recorder_name,
                        data,
                        recorder_col = 'recorders',
                        date_col = 'date_start',
                        format = '%d/%m/%Y',
                        day_limit = 5){

  # Get the recorders data
  data <- data[data[,recorder_col] == recorder_name, ]

  # Get unique dates as dates
  dates <- sort(as.Date(as.character(unique(data[,date_col])), format = format))

  # we cannot calculate these metrics if people have very few
  # dates on which they record
  if(length(unique(dates)) < day_limit){

    # return
    return(data.frame(recorder = recorder_name,
                      periodicity = NA,
                      periodicity_variation = NA,
                      median_streak = NA,
                      sd_streak = NA,
                      max_streak = NA,
                      n_days = length(unique(dates))))

  } else {

    # Calculate the elapsed days between each date in sequence
    elapses <- sapply(1:(length(dates)-1),
         FUN = function(x){
           return(as.numeric(dates[x + 1] - dates[x]))
         })

  # periodicity calculation
  periodicity <- median(elapses)
```

```
  # variation in periodicity
  periodicity_variation <- sd(elapses)

  # average streak length
  # Streaks are IDed by 1's
  non_streak <- length(elapses[elapses > 1])
  streaks <- rle(elapses)
  streaks_1 <- (streaks$lengths[streaks$value == 1]) + 1

  # Combine streaks and non-streaks
  streak_lengths <- c(rep(1, non_streak), streaks_1)

  # calculate ome metrics
  median_streak <- median(streak_lengths)
  sd_streak <- sd(streak_lengths)
  max_streak <- max(streak_lengths)

  # return
  return(data.frame(recorder = recorder_name,
                    periodicity = periodicity,
                    periodicity_variation = periodicity_variation,
                    median_streak = median_streak,
                    sd_streak = sd_streak,
                    max_streak = max_streak,
                    n_days = length(unique(dates))))

  }


}

# Test on David and Tom
periodicity(data = iRB, recorder_name = 'Roy, David')
```

```
##     recorder periodicity periodicity_variation median_streak sd_streak
## 1 Roy, David           3              20.61582             1 0.8909135
##   max_streak n_days
## 1          7    165
```

```
periodicity(data = iRB, recorder_name = 'August, Tom')
```

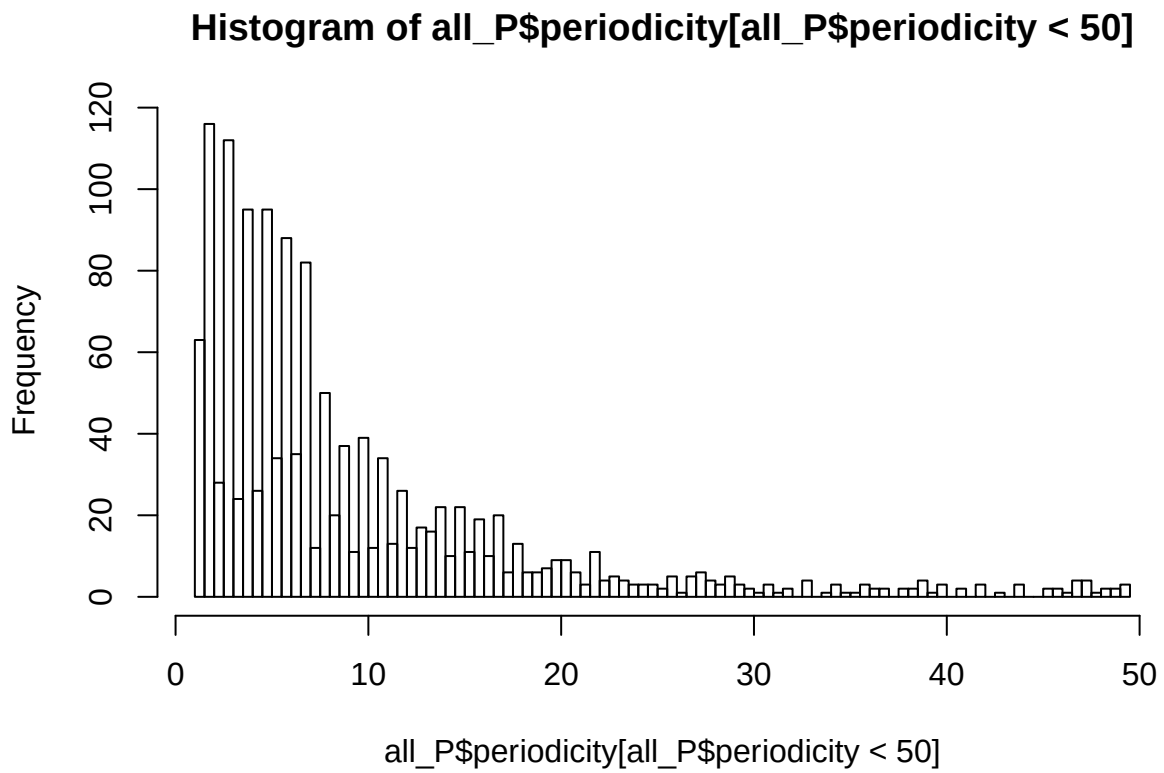
```
##      recorder periodicity periodicity_variation median_streak sd_streak
## 1 August, Tom          30              44.30637             1 0.4045199
##   max_streak n_days
## 1          2     12
```

```
# David is a much more regular recorder than Tom with less
# variation in periodicity and a longer max streak though
# Tom has less days of data to work with

# Run for everyone
all_P <- do.call(rbind, lapply(X = unique(iRB$recorders),
```

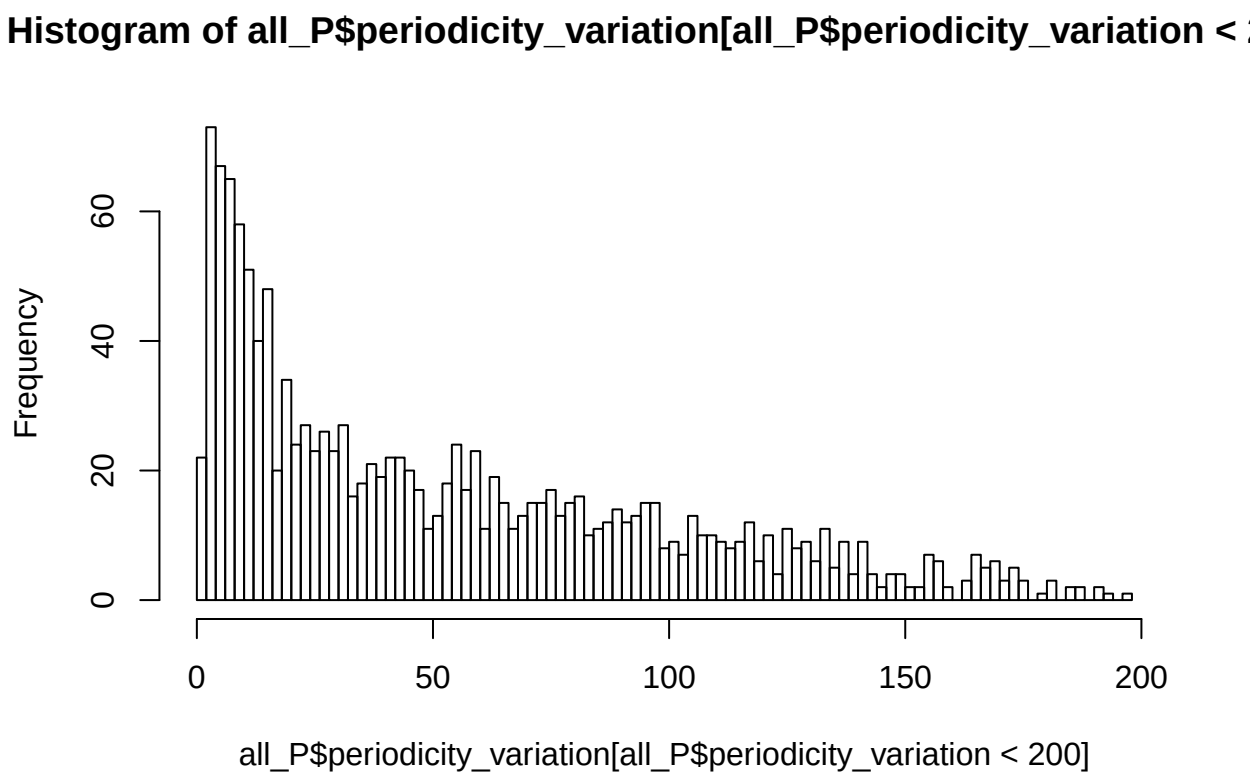
```
                                FUN = periodicity,
                                data = iRB))

# Lets have a look at some of these
head(all_P, 20)[c(5,8,1),]
```

```
##          recorder periodicity periodicity_variation median_streak
## 5    Allan, David           1              5.363846             1
## 8    Stewart, Tam           1             13.823237             1
## 1 Brookes , Anne            5             27.619121             1
##   sd_streak max_streak n_days
## 5 2.0640484         16    378
## 8 1.8028182         15    272
## 1 0.5834249          4     70
```

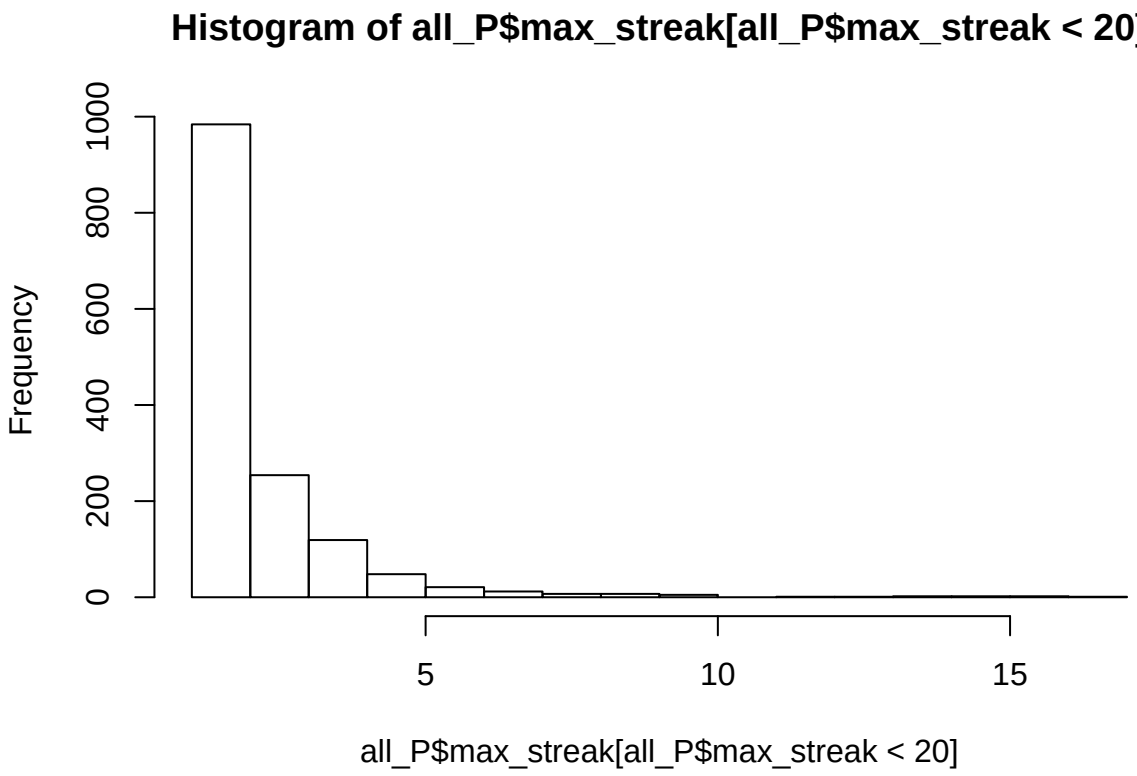
```
# David a Tam are both very studious recorders with
# long max streaks and very low periodicity. David is
# perhaps more studious as he has a lower periodicity variation
# Anne is less studious but still has a low periodicity

# Nice poission dist. for periodicity
hist(all_P$periodicity[all_P$periodicity < 50],
     breaks = 100)
```



```r
# Nice poission dist. for periodicity_variation (long tail)
hist(all_P$periodicity_variation[all_P$periodicity_variation < 200],
     breaks = 100)
```




```r
# Dist. of max_streak
hist(all_P$max_streak[all_P$max_streak < 20],
     breaks = 20)
```



There are some issues with `periodicity_variation` here; I think because most people will have large gaps over the summer and small gaps over the winter, I think this might make the `periodicity` (the median of elapses) a better metric in this case. These metrics cannot be calculate for people who have only made one record. I have included a parameter `day_limit` to allow us to set a limit at which we calculate these metrics.

Spatial Meterics

These metrics deal with the spatial distribution of records

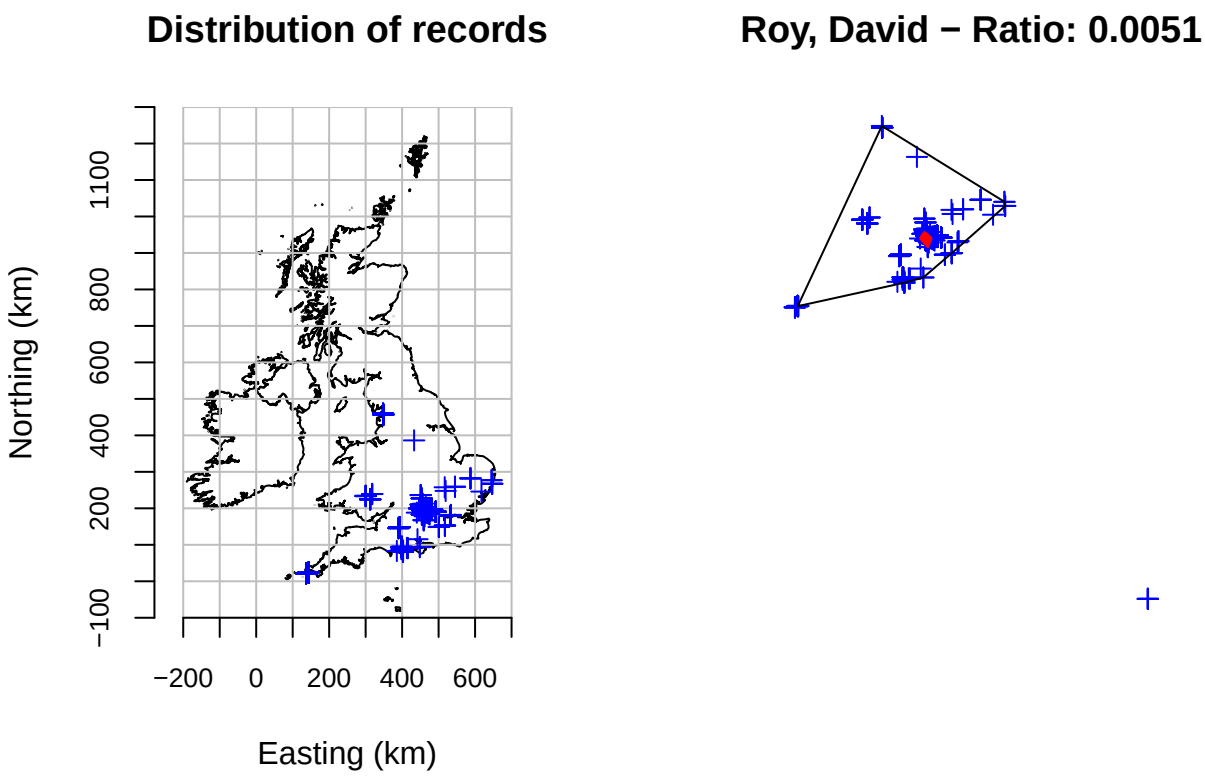
Area and heterogenity of recording

I think the first step for all of these metrics is to turn the points into a SpatialPoints object which will allow us to manipulate then more easily. Once we have done that we can calculate MCP (minimum convex polygons) around the points. We might want to change this method to a method that is less susceptible to outliers such as alpha hull (we can talk to Colin about this). Here I use 95% MCP as the total recording area (hopefully removing outliers), and use the ratio of 95%:50% as a measure of heterogeneity.

```
# Function takes data and username and returns spatial metrics
spatial_behaviour <- function(data, recorder_name,
                              latitude_col, longitude_col,
                              recorder_col = 'recorders',
                              upper_percentile = 95,
                              lower_percentile = 50){

  if(is.factor(recorder_name)){
```

```
    recorder_name <- as.character(recorder_name)
  }

  n_row <- nrow(iRB[iRB[,recorder_col] == recorder_name, ])

  if(n_row >= 5){

    # Convert to SpatialPoints
    spPoints_LL <- SpatialPoints(iRB[iRB[,recorder_col] == recorder_name,
                                     c(longitude_col, latitude_col)])
    # Data is lat long
    proj4string(spPoints_LL) <- CRS("+init=epsg:4326")

    # Convert to Eastings Northings to get meters on X and Y
    spPoint_UK <- spTransform(spPoints_LL, "+init=epsg:27700")

    # Calculate the larger MCP
    mcp_poly_upper <- mcp(spPoint_UK,
                          percent = upper_percentile,
                          unin = 'm',
                          unout = 'km2')

    # Calculate the smaller MCP
    mcp_poly_lower <- mcp(spPoint_UK,
                          percent = lower_percentile,
                          unin = 'm',
                          unout = 'km2')

    return(list(recorder = recorder_name,
                spPoint_UK = spPoint_UK,
                mcp_poly_upper = mcp_poly_upper,
                mcp_poly_lower = mcp_poly_lower,
                upper_area = mcp_poly_upper$area,
                lower_area = mcp_poly_lower$area,
                ratio = mcp_poly_lower$area/mcp_poly_upper$area,
                n = n_row))
  } else {
    return(list(recorder = recorder_name,
                spPoint_UK = NA,
                mcp_poly_upper = NA,
                mcp_poly_lower = NA,
                upper_area = NA,
                lower_area = NA,
                ratio = NA,
                n = n_row))
  }

}

# Test on one recorder
David_spatial <- spatial_behaviour(data = iRB, recorder_name = 'Roy, David',
                                   latitude_col = 'lat', longitude_col = 'st_x')
```

```r
# Function for plotting records
plot_ratio <- function(data){
  par(mfrow = c(1,2))
  data(UK)
  plot_GIS(UK, new.window = FALSE, main = 'Distribution of records')
  points(data$spPoint_UK, pch = 3, col = 'blue')

  # Plot David's heat map
  plot(data$spPoint_UK,
       main = paste(data$recorder, '-', 'Ratio:', round(data$ratio, 4)),
       col = 'blue')
  upper_polygon <- data$mcp_poly_upper@polygons[[1]]@Polygons[[1]]@coords
  polygon(x = upper_polygon[,1],
        y = upper_polygon[,2])
  lower_polygon <- data$mcp_poly_lower@polygons[[1]]@Polygons[[1]]@coords
  polygon(x = lower_polygon[,1],
        y = lower_polygon[,2],
        col = 'red', border = 'red')
  par(mfrow = c(1,1))
}

# Plot
plot_ratio(data = David_spatial)
```



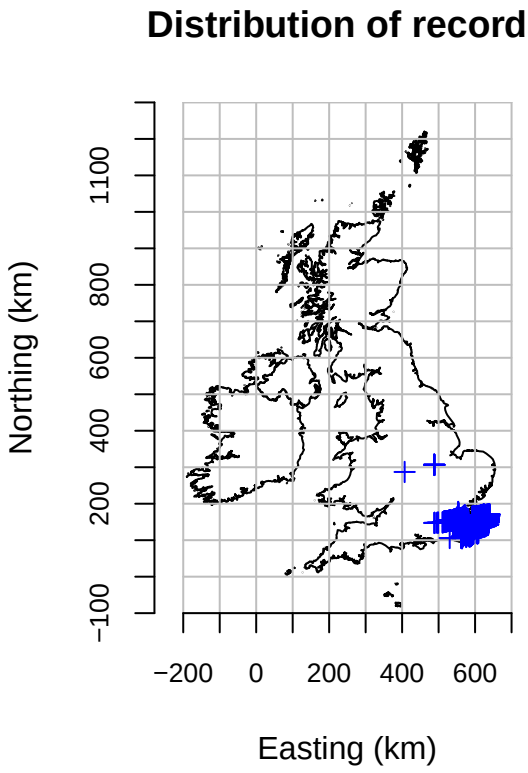
```
## NOTE DAVID HAS A RECORD FROM OUTSIDE THE UK ##

# Apply to all recorders
all_spatial <- lapply(unique(iRB$recorders), FUN = function(x){
  recorder_info <- spatial_behaviour(data = iRB, recorder_name = x,
                                     latitude_col = 'lat', longitude_col = 'st_x')
  return(data.frame(recorder = recorder_info$recorder,
                    upper_area = recorder_info$upper_area,
                    lower_area = recorder_info$lower_area,
                    ratio = recorder_info$ratio,
                    n = recorder_info$n))
})

# combine results
temp <- do.call(rbind, all_spatial)
temp <- temp[temp$n > 400, ]

# Lets have a look at some people who have recorded a lot
temp[order(temp$ratio, decreasing = TRUE),]
```

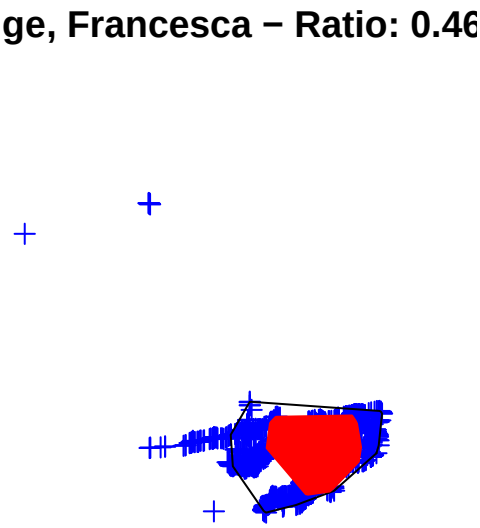
```
##                    recorder   upper_area   lower_area        ratio    n
## 11      Partridge, Francesca 5.176381e+03 2.414166e+03 0.4663809300 1418
## 52          Cornish, Stephen 3.534308e+00 9.106945e-01 0.2576726378  487
## 180               Limb, Ken 3.042189e+04 7.577886e+03 0.2490932324  622
## 395              Atkin, Paul 1.393205e+03 3.223848e+02 0.2313978875  615
## 139           Hunter, Amands 7.531823e+02 1.246409e+02 0.1654856950 1090
## 104              Leaver, Kim 1.394622e+03 2.193449e+02 0.1572790538  537
## 26                fenn, paul 5.583057e+03 8.487535e+02 0.1520230771 2503
## 65              Gillie, Tony 1.750010e+03 2.097561e+02 0.1198599351 1112
## 339             Bowles, Nick 4.155985e+03 3.545317e+02 0.0853062848  590
## 256            Cowton, Keith 2.109076e+04 1.119724e+03 0.0530907308  445
## 113              Hill, Brian 7.170905e+03 3.793400e+02 0.0528998784  851
## 5               Allan, David 1.471503e+04 6.918175e+02 0.0470143525 3180
## 39            Warren, Martin 3.863468e+04 1.337492e+03 0.0346189363 2434
## 72               Jones, Dave 2.767527e+01 9.346723e-01 0.0337728352 2207
## 109           Shanks, Scott  2.523051e+04 8.281931e+02 0.0328250625  513
## 103       Pennington, Robert 6.234135e+03 1.838310e+02 0.0294878026  969
## 1356          Saville, Simon 2.969962e+04 8.676054e+02 0.0292126767  441
## 383           Steele, Andrew 9.131555e+04 2.632030e+03 0.0288234530  563
## 123               Cox, Steve 4.447586e+04 1.265539e+03 0.0284545070  991
## 8               Stewart, Tam 2.886784e+04 8.000475e+02 0.0277141435 1811
## 175              Sims, Clive 2.359346e+04 6.345611e+02 0.0268956338  864
## 41   Lonsdale, Liz and Steve 1.536766e+05 3.975898e+03 0.0258718467  542
## 523           Shersby, Megan 3.790063e+04 9.782456e+02 0.0258108020  478
## 43            Newbould, John 7.404497e+04 1.879715e+03 0.0253861332 1001
## 488             Kilbey, Dave 3.198760e+04 6.217184e+02 0.0194362359  780
## 45              Sell, Claire 7.754820e+02 1.501250e+01 0.0193589276  555
## 197            Lunnon, Marie 6.657662e+01 1.256184e+00 0.0188682491  444
## 96          Checkley, Graham 1.240849e+03 2.166370e+01 0.0174587706 1813
## 143             Fox, Richard 3.168087e+04 4.871148e+02 0.0153756723 1147
## 19                Roy, David 1.065308e+05 5.448197e+02 0.0051142000  615
## 78            shilland, ewan 1.519489e+05 6.898383e+02 0.0045399371 1636
## 87             Dawson, Steve 1.135666e+03 4.548707e+00 0.0040053224  789
```

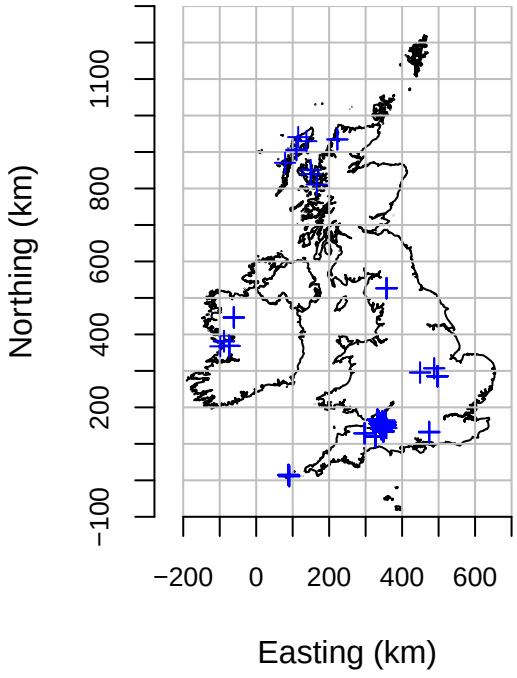
```
## 140            Austin, David 4.573147e+03 1.795774e+01 0.0039267802  441
## 100             Ford, Rachel 7.010182e+01 9.404533e-02 0.0013415532  431
## 158             Harley, Ross 1.873400e+05 7.630491e+01 0.0004073071  682
```

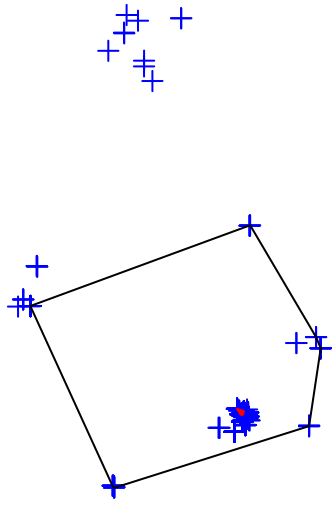
Lets have a look at two people with very different ratios

```
# Get the names of top and bottom
temp <- temp[order(temp$ratio, decreasing = TRUE),]

top <- as.character(head(temp$recorder, 1))
bottom <- as.character(tail(temp$recorder, 1))

# Plot the top and bottom ratio recorder
for(i in c(top, bottom)){
  top_d <- spatial_behaviour(data = iRB,
                             recorder_name = i,
                             latitude_col = 'lat',
                             longitude_col = 'st_x')
  plot_ratio(data = top_d)
}
```


Distribution of records
Northing (km)
−100
200
400
600
800
1100
−200
0
200
400
600
Easting (km)


Partridge, Francesca − Ratio: 0.46


Distribution of records
Northing (km)
−100
200
400
600
800
1100
−200
0
200
400
600
Easting (km)


Harley, Ross − Ratio: 4e−04

Taxanomic Metrics

These metric relate the the species that people record

Taxanomic Breadth

This is simply a measure of the proportion of taxa a person has recorded. Note this is going to be correlated to the number of records.

```
taxa_breadth <- function(data, recorder_name,
                         sp_col = 'preferred_taxon',
                         recorder_col = 'recorders'){

  data_rec <- data[data[,recorder_col] == recorder_name, c(sp_col, recorder_col)]

  return(data.frame(recorder = recorder_name,
                    taxa_breadth = length(unique(data_rec[ ,sp_col])),
                    taxa_prop = length(unique(data_rec[ ,sp_col]))/length(unique(data[ ,sp_col])),
                    n = nrow(data_rec)))
}

taxa_breadth <- do.call(rbind, lapply(unique(iRB$recorders), FUN = taxa_breadth, data = iRB))

temp <- taxa_breadth[taxa_breadth$n > 400, ]

# Lets have a look at some people who have recorded a lot
temp[order(temp$taxa_prop, decreasing = TRUE),]
```

```
##                     recorder taxa_breadth taxa_prop    n
## 39            Warren, Martin           52 0.6265060 2434
## 5               Allan, David           51 0.6144578 3180
## 103       Pennington, Robert           49 0.5903614  969
## 113              Hill, Brian           48 0.5783133  851
## 1356          Saville, Simon           48 0.5783133  441
## 123               Cox, Steve           47 0.5662651  991
## 175              Sims, Clive           47 0.5662651  864
## 143             Fox, Richard           46 0.5542169 1147
## 158             Harley, Ross           45 0.5421687  682
## 383           Steele, Andrew           45 0.5421687  563
## 256            Cowton, Keith           42 0.5060241  445
## 395              Atkin, Paul           42 0.5060241  615
## 26                fenn, paul           41 0.4939759 2503
## 180                Limb, Ken           41 0.4939759  622
## 488             Kilbey, Dave           41 0.4939759  780
## 87             Dawson, Steve           40 0.4819277  789
## 65              Gillie, Tony           39 0.4698795 1112
## 523           Shersby, Megan           38 0.4578313  478
## 19                Roy, David           37 0.4457831  615
## 41   Lonsdale, Liz and Steve           37 0.4457831  542
## 78            shilland, ewan           36 0.4337349 1636
## 339             Bowles, Nick           36 0.4337349  590
## 43            Newbould, John           33 0.3975904 1001
## 11      Partridge, Francesca           32 0.3855422 1418
```

```
## 45              Sell, Claire          32 0.3855422  555
## 139           Hunter, Amands          31 0.3734940 1090
## 8               Stewart, Tam          29 0.3493976 1811
## 197             Lunnon, Marie         28 0.3373494  444
## 109           Shanks, Scott           26 0.3132530  513
## 104              Leaver, Kim          24 0.2891566  537
## 72               Jones, Dave          23 0.2771084 2207
## 96          Checkley, Graham          22 0.2650602 1813
## 140            Austin, David          22 0.2650602  441
## 52          Cornish, Stephen          19 0.2289157  487
## 100              Ford, Rachel          15 0.1807229  431
```

Species Rarity

We want to capture the rarity of the species that people record. For example are they just recording the common species or are they only recording the rare ones, or perhaps they are recording everything. Since we don't know the real frequency distribution we can only compare people to the global average in the dataset. We can look to see what the distribution of species rank for each recorder is and how this compares to all records. A recorder only interested in rare species will have a median rank higher than the average. A recorder only recording common species will have a value lower than the average.

```
# Lets look at a recorder
species_rank <- function(data, recorder_name,
                         sp_col = 'preferred_taxon',
                         recorder_col = 'recorders'){

  data <- data[,c(sp_col, recorder_col)]
  rank_species <- rank(abs(table(data[,sp_col])-max(table(data[,sp_col]))))
  sp_counts <- table(data[,sp_col])

  rank_reps <- rep(rank_species, sp_counts)
  grand_median <- median(rank_reps)
  grand_sd <- sd(rank_reps)

  recorder_data <- data[data[,recorder_col] == recorder_name,]
  recorder_data$rank <- rank_species[recorder_data[ ,sp_col]]

  return(data.frame(recorder = as.character(recorder_name),
                    median = median(recorder_data$rank),
                    median_diff = median(recorder_data$rank) - grand_median,
                    stdev = sd(recorder_data$rank),
                    n = nrow(recorder_data)))
}

rarity_preference <- do.call(rbind,
                             lapply(unique(iRB$recorders),
                                    FUN = species_rank,
                                    data = iRB))

temp <- rarity_preference[rarity_preference$n > 400, ]

# Lets have a look at some people who have recorded a lot
temp[order(temp$median_diff, decreasing = TRUE),]
```

```
##                     recorder median median_diff     stdev    n
## 1356          Saville, Simon     13           5 12.191833  441
## 256            Cowton, Keith     12           4 10.283900  445
## 39            Warren, Martin     11           3 10.754206 2434
## 175              Sims, Clive     11           3 10.132960  864
## 339             Bowles, Nick     10           2  8.557264  590
## 395              Atkin, Paul     10           2  9.738285  615
## 523           Shersby, Megan     10           2  8.613459  478
## 8               Stewart, Tam      9           1 10.764394 1811
## 19                Roy, David      9           1  9.647095  615
## 26                fenn, paul      9           1  8.779256 2503
## 43            Newbould, John      9           1  8.245020 1001
## 45              Sell, Claire      9           1  8.912894  555
## 65              Gillie, Tony      9           1  8.645367 1112
## 103       Pennington, Robert      9           1  9.100094  969
## 109           Shanks, Scott       9           1  9.482688  513
## 113              Hill, Brian      9           1 10.226885  851
## 139           Hunter, Amands      9           1  7.199181 1090
## 158             Harley, Ross      9           1  9.410956  682
## 180                Limb, Ken      9           1  9.165788  622
## 197            Lunnon, Marie      9           1  7.004225  444
## 41   Lonsdale, Liz and Steve      8           0  8.646054  542
## 78            shilland, ewan      8           0  8.303214 1636
## 96          Checkley, Graham      8           0  6.931797 1813
## 104              Leaver, Kim      8           0  6.082150  537
## 143             Fox, Richard      8           0  9.681677 1147
## 383           Steele, Andrew      8           0  9.108308  563
## 488             Kilbey, Dave      8           0  9.170174  780
## 87             Dawson, Steve      7          -1  7.926813  789
## 100             Ford, Rachel      7          -1  5.281118  431
## 123               Cox, Steve      7          -1  9.048282  991
## 5               Allan, David      6          -2  8.643921 3180
## 11      Partridge, Francesca      6          -2  6.888191 1418
## 72                Jones, Dave     6          -2  4.862982 2207
## 52          Cornish, Stephen      5          -3  5.081520  487
## 140             Austin, David     5          -3  5.474312  441
```

Here `median_diff` gives the difference between the grand median for all records and the recorders median. This suggests `Saville, Simon` prefers to record rare species and `Cornish, Stephen` prefers to record common species.

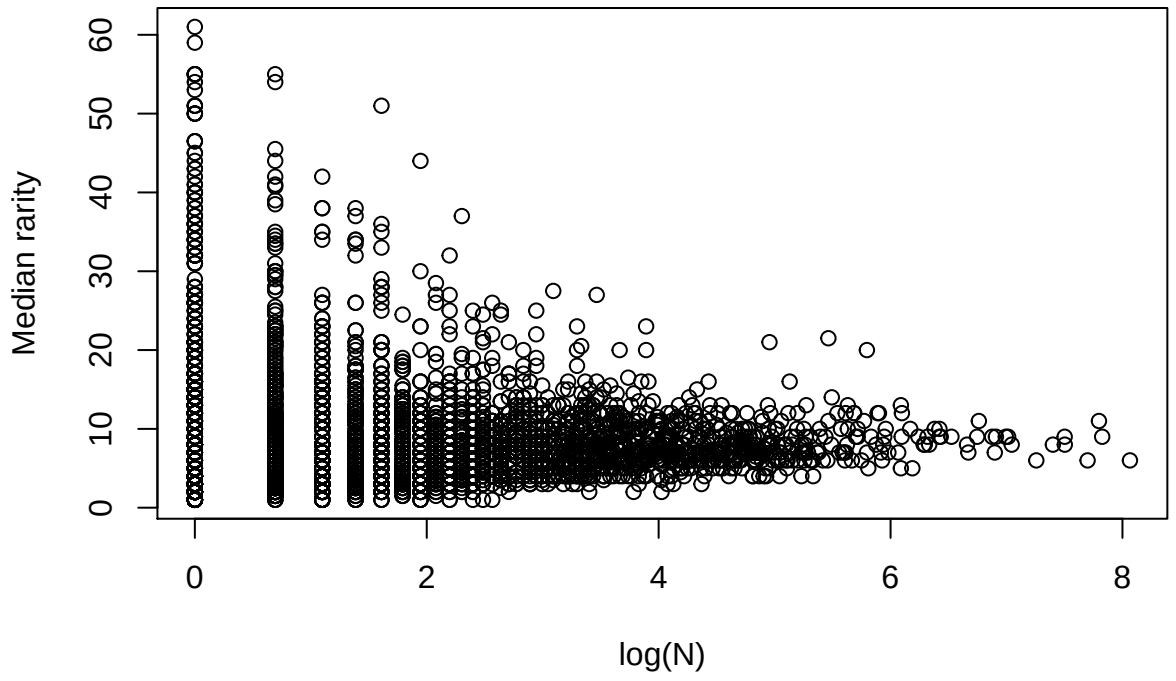
This could be correlated to the number of records.

```
mod <- glm(median ~ log(n), data = rarity_preference, family = 'quasipoisson')
summary(mod)
```

```
##
## Call:
## glm(formula = median ~ log(n), family = "quasipoisson", data = rarity_preference)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7397  -1.6376  -0.4224   0.7604  10.7394
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.334807   0.018526 126.030   <2e-16 ***
## log(n)      -0.070761   0.008457  -8.367   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 5.26387)
##
##     Null deviance: 17611  on 3944  degrees of freedom
## Residual deviance: 17232  on 3943  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
plot(log(rarity_preference$n),
     rarity_preference$median,
     xlab = 'log(N)',
     ylab = 'Median rarity')
```

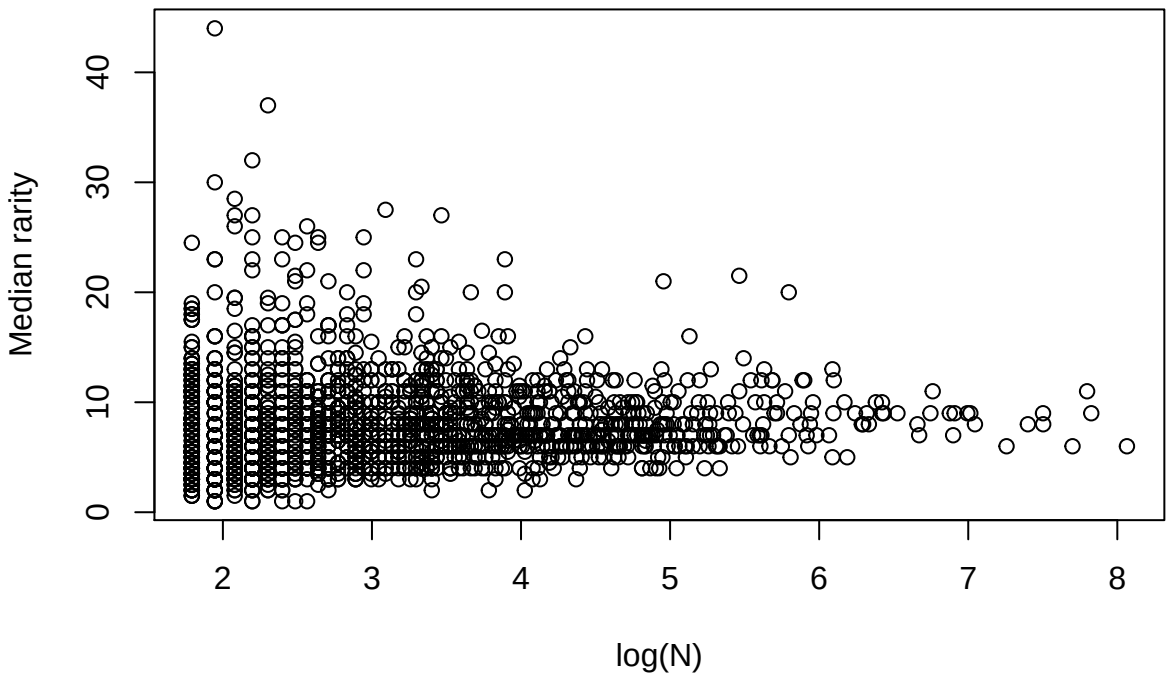


There is a significant negative relationship. The more records you make the lower your median value. This could be a result of the fact that people who make only a few records record rare stuff?

```
rarity_preference_above <- rarity_preference[rarity_preference$n > 5, ]
mod <- glm(median ~ log(n), data = rarity_preference_above, family = 'quasipoisson')
summary(mod)
```

```
##
## Call:
## glm(formula = median ~ log(n), family = "quasipoisson", data = rarity_preference_above)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1729  -0.9646  -0.2134   0.6260   8.7863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.076269   0.034427  60.310   <2e-16 ***
## log(n)      0.007479   0.010513   0.711    0.477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.087498)
##
##     Null deviance: 3474.5  on 1877  degrees of freedom
## Residual deviance: 3473.5  on 1876  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
plot(log(rarity_preference_above$n),
     rarity_preference_above$median,
     xlab = 'log(N)',
     ylab = 'Median rarity')
```



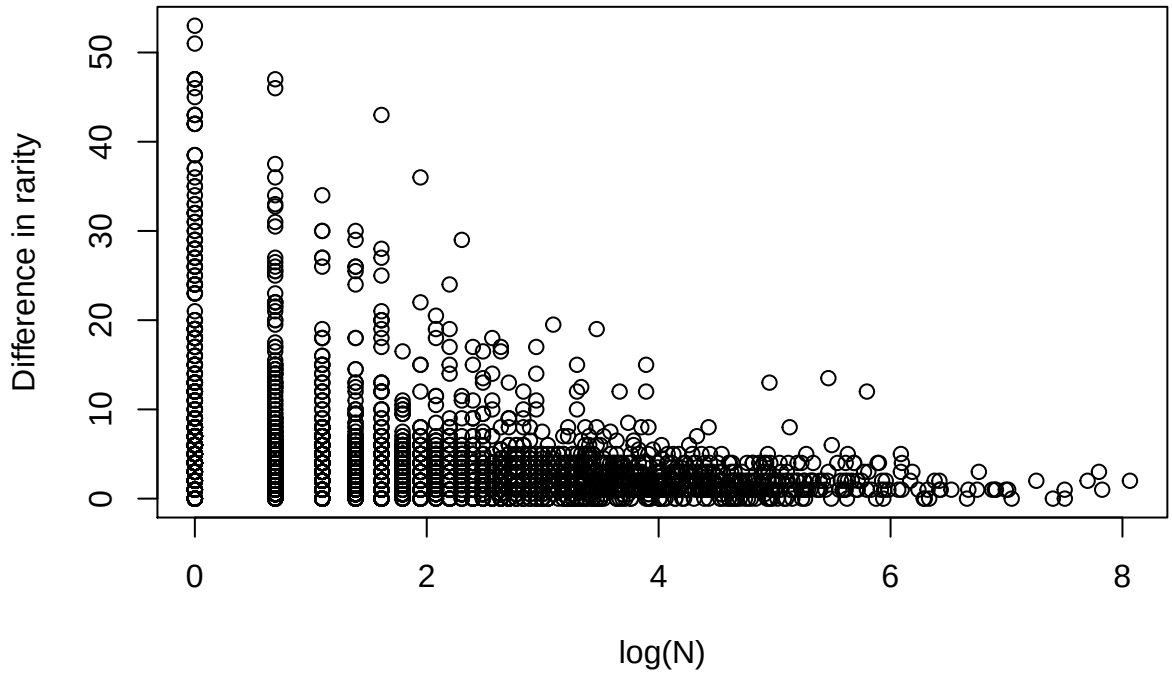
Okay, the relationship falls down once we get rid of the people who only record a few species. I suggest this metric not be estimates for people who contribute only a few records. The relationship might actually be between deviation from the median and `n`.

```
rarity_preference$median_diff_abs <- abs(rarity_preference$median_diff)
mod <- glm(median_diff_abs ~ log(n), data = rarity_preference, family = 'quasipoisson')
summary(mod)
```

```
##
## Call:
## glm(formula = median_diff_abs ~ log(n), family = "quasipoisson",
##     data = rarity_preference)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8262  -1.3717  -0.5038   0.3857  10.8928
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.99062    0.02414   82.47   <2e-16 ***
## log(n)      -0.31472    0.01388  -22.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 5.259472)
##
```

```
##     Null deviance: 18546  on 3944  degrees of freedom
## Residual deviance: 15454  on 3943  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
plot(log(rarity_preference$n),
     rarity_preference$median_diff_abs,
     xlab = 'log(N)',
     ylab = 'Difference in rarity')
```



The more records you record the less you deviate from the median. This is probably because you only get extreme values where the sample size is small.