Statistical_Analysis_Simpson

Daniela Linero

1/7/2020

1. Load data

Load the table that contains the abudance and occurrence measures for frugivores along with the plants dispersed and not dispersed by animals.

```
# Import table
diversityS <- readRDS(file = "./output/cleaned_data/01_Filter_data_frugi_endooPlants_notEndooPlants_rec
```

2. Correct abundance measures using Sampling effort

```
diversityS <- yarg::CorrectSamplingEffort(diversityS)
```

```
## Correcting 0 missing sampling effort values
## Rescaling sampling effort
## Correcting 114923 values for sensitivity to sampling effort
```

3. Select studies that assesed more than one species

```
# Create a dataset with the SS that assessed more than 1 species
list <- diversityS  %>%

  # By grouping by SS we are comparing sites with the same diversity metric
  # type either abundance or occurence
  dplyr::group_by(SS) %>%

  # Create a new column to calculate the number of species sampled per study
  # n_distinct = length(unique())
  dplyr::mutate(N_species_sampled = dplyr::n_distinct(Taxon_name_entered)) %>%

  # ungroup data frame
  dplyr::ungroup() %>%

  # Filter the studies that sampled more than 1 species
  dplyr::filter(N_species_sampled > 1) %>%

  base::droplevels()
```

```
# Create a character vector of unique SS that assessed more than 1 species
list <- as.character(unique(list$SS))
```

```r
diversityS <- diversityS %>%

  # Filter the studies present in the list
  base::subset(SS %in% list) %>%

  droplevels()
```

4. Remove sites that will produce NaN values

```r
diversityS <- diversityS %>%

  # Remove the sites that have a maximum abundance of 0
  base::subset(SSBS %nin% c("MJ1_2009__Lehouck 2 Fururu 10",
                            "MJ1_2009__Lehouck 2 Fururu 11",
                              "MJ1_2009__Lehouck 2 Macha 12",
                                "MJ1_2009__Lehouck 2 Macha 13")) %>%
  base::droplevels()
```

5. Split dataset

I will separate the measures of plants not dispersed by animals, in order to avoid that they are merged together with plants dispersed by animals with the MergeSites function

```r
# I am going to separate the records that belong to plants not dispersed
# by animals
diversityS_notEndo <- diversityS %>%

  base::subset(Kingdom == "nePlantae") %>%

  base::droplevels()

# Get the table without plants dispersed by animals
diversityS_frugi_endo <- diversityS %>%

  base::subset(Kingdom != "nePlantae") %>%

  base::droplevels()
```

6. Merge Sites

```r
diversityS_frugi_endo <- yarg::MergeSites(diversityS_frugi_endo,
                                          silent = TRUE,
                                          merge.extra = "Wilderness_area")

diversityS_notEndo <- yarg::MergeSites(diversityS_notEndo,
                                       silent = TRUE,
                                       merge.extra = "Wilderness_area")
```

7. Rename Predominant Habitat

```r
# Rename the column predominant habitat, as the dataset is actually
# refering to land use
```

```
diversityS_frugi_endo <-  dplyr::rename(diversityS_frugi_endo,
                        Predominant_land_use = Predominant_habitat)

diversityS_notEndo <-  dplyr::rename(diversityS_notEndo,
                     Predominant_land_use = Predominant_habitat)
```

8. Calculate Diversity Metrics

```
# Calculate diversity metrics for animals and endoozoocoric plants
diversity1S_frugi_endo <- diversityS_frugi_endo %>%

  # add Diversity_metric_is_valid column
  dplyr::mutate(Diversity_metric_is_valid = TRUE) %>%

  # The extra.cols parameter is used for columns that we want to
  # transferred to the final site-level data frame and that the function
  # does not add  automatically
  yarg::SiteMetrics(extra.cols = c("SSB", "SSBS", "Predominant_land_use", "Kingdom"))

# Calculate diversity metrics for plants not dispersed by animals
diversity1S_notendo <- diversityS_notEndo %>%

  # add Diversity_metric_is_valid column
  dplyr::mutate(Diversity_metric_is_valid = TRUE) %>%

  # The extra.cols parameter is used for columns that we want to
  # transferred to the final site-level data frame and that the function
  # does not add  automatically
  yarg::SiteMetrics(extra.cols = c("SSB", "SSBS", "Predominant_land_use", "Kingdom"))

# Merge the site metrics for all organisms
diversity_all <- base::rbind.data.frame(diversity1S_frugi_endo,
                                        diversity1S_notendo)
```

9. Check results

```
# Merge the measures of frugivores, endoplants and not endoplants
diversityS_combined <- rbind.data.frame(diversityS_frugi_endo,
                                        diversityS_notEndo)

diversityS_combined  %>%

  # subset one site to check
  base::subset(SSS == "DG1_2013__Zou 1 1") %>% base::droplevels() %>%

  # Calculate each species abundance over total abundance at the site
  dplyr::mutate(Proportion = Measurement/sum(Measurement)) %>%

  # Sum the proportions and squared them
  dplyr::mutate(Simpson_D = sum(Proportion^2)) %>%

  # Calculate 1/D
  dplyr::mutate(one_over_D = 1/Simpson_D) %>%
```

```
   # Select only columns we're interested in
  dplyr::select(SSS, Best_guess_binomial, Measurement, one_over_D)
```

```
##                  SSS  Best_guess_binomial Measurement one_over_D
## 1 DG1_2013__Zou 1 1      Megalaima oorti    20.58333   3.698822
## 2 DG1_2013__Zou 1 1  Pycnonotus sinensis    48.75000   3.698822
## 3 DG1_2013__Zou 1 1  Alophoixus pallidus    34.66667   3.698822
## 4 DG1_2013__Zou 1 1 Hemixos castanonotus    35.75000   3.698822
```

```
# Compare with the SiteMetrics result
diversity_all %>% base::subset(SSS == "DG1_2013__Zou 1 1") %>%
  dplyr::select(SSS, Simpson_diversity)
```

```
##                    SSS Simpson_diversity
## 462 DG1_2013__Zou 1 1          3.698822
```

10. Check the abundance measures units

I am going to check that the measures used that weren't number of individuals or ind/km2, were reasonably concordant to an abundance measure of each species.

```
# Get the unique diversity metric type and unit for each study
Diversity_metric_unit <- diversityS_combined %>%

  # Get only one row for each study
  dplyr::distinct(SS, .keep_all = TRUE) %>%

  # Select the columns we're interested in
  dplyr::select(Source_ID, SS,
                Diversity_metric_unit, Kingdom,
                Study_common_taxon) %>%

  # Filter the studies that need to be checked
  base::subset(Diversity_metric_unit %nin% c("individuals",
                                             "stems/hectare",
                                             "presence/absence",
                                             "effort-corrected individuals",
                                             "individuals/km2",
                                             "groups/colonies per km",
                                             "number of groups",
                                             "proportion of plots",
                                             "individuals/km")) %>%
  droplevels()


kable(Diversity_metric_unit, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

	Source_ID	SS	Diversity_metric_unit	Kingdom	Study_common_taxon
38	HB1_1998___Svenning	HB1_1998___Svenning 1	percentage	Plantae	Arecaceae
41	HP1_2005___Hietz	HP1_2005___Hietz 1	percentage of sites	Plantae	Tracheophyta
92	SH1_2002___Sheil	SH1_2002___Sheil 1	number of plots	Plantae	Plantae
96	SH1_2013___CIFORcameroon	SH1_2013___CIFORcameroon 1	number of plots	Plantae	Plantae
98	SH1_2013___CIFORphilippines	SH1_2013___CIFORphilippines 1	number of plots	Plantae	Plantae
99	SH1_2013___CIFORphilippines	SH1_2013___CIFORphilippines 2	number of plots	Plantae	Tracheophyta
108	YY1_2015___Mandle	YY1_2015___Mandle 1	percentage	Plantae	Tracheophyta
111	YY1_2016___Mohandass	YY1_2016___Mohandass 1	percentage	Plantae	

11. Models

Count number of sites for the SiteMetrics table in order to know the sample size for each land-use type and intensity.

```
# Remove sites that don't have a simpson diversity value or land-use type
diversity_simpson <- drop_na(diversity_all,
                             Simpson_diversity,
                             Predominant_land_use) %>%
  droplevels()

# Number of sites
table(diversity_simpson$Predominant_land_use,
                diversity_simpson$Use_intensity,
                diversity_simpson$Kingdom)
```

```
## , ,  = Animalia
##
##
##                                          Minimal use Light use
##   Primary forest                                 325       113
##   Primary non-forest                               0         0
##   Young secondary vegetation                      37         8
##   Intermediate secondary vegetation               81        50
##   Mature secondary vegetation                     11         4
##   Secondary vegetation (indeterminate age)        27         6
##   Plantation forest                               27       207
##   Pasture                                          0         4
##   Cropland                                        59        20
##   Urban                                            0        29
##
##                                          Intense use Cannot decide
##   Primary forest                                  15             9
##   Primary non-forest                               0             0
##   Young secondary vegetation                       0            22
##   Intermediate secondary vegetation                2            26
##   Mature secondary vegetation                      0            21
##   Secondary vegetation (indeterminate age)         0            10
##   Plantation forest                               18             0
##   Pasture                                          1            24
##   Cropland                                        40            18
##   Urban                                            1             0
##
## , ,  = Plantae
##
```

```
##
##                                         Minimal use Light use
##   Primary forest                                247        55
##   Primary non-forest                              2         0
##   Young secondary vegetation                     55        15
##   Intermediate secondary vegetation              36        72
##   Mature secondary vegetation                    34         8
##   Secondary vegetation (indeterminate age)        7        15
##   Plantation forest                              71       166
##   Pasture                                        19        17
##   Cropland                                       18         0
##   Urban                                           0         0
##
##                                         Intense use Cannot decide
##   Primary forest                                 70            59
##   Primary non-forest                              0             0
##   Young secondary vegetation                      0             1
##   Intermediate secondary vegetation               2            11
##   Mature secondary vegetation                     6             0
##   Secondary vegetation (indeterminate age)        0           112
##   Plantation forest                              18             3
##   Pasture                                         2             0
##   Cropland                                       36             8
##   Urban                                           0             0
##
## , ,  = nePlantae
##
##
##                                         Minimal use Light use
##   Primary forest                                196        41
##   Primary non-forest                              3         1
##   Young secondary vegetation                     61        11
##   Intermediate secondary vegetation              38        31
##   Mature secondary vegetation                    10         7
##   Secondary vegetation (indeterminate age)        8        18
##   Plantation forest                              68        65
##   Pasture                                        19        24
##   Cropland                                       19         0
##   Urban                                           0         0
##
##                                         Intense use Cannot decide
##   Primary forest                                 50            51
##   Primary non-forest                              0             0
##   Young secondary vegetation                      0             3
##   Intermediate secondary vegetation               2            18
##   Mature secondary vegetation                     3             0
##   Secondary vegetation (indeterminate age)        0           118
##   Plantation forest                               4             6
##   Pasture                                         4             0
##   Cropland                                        6             8
##   Urban                                           0             0
```

According to the number of sites, I am going to try this initial combination:

- Primary can be divided into the two level intensities in all cases
- Cropland has to be merged
- ISV can be divided in minimal use and light/intense use
- MSV has to be merged
- Pasture has to be merged
- Plantation forest can be divided in minimal use and light/intense
- YSV has to be merged

```
# Call the function that merges lan-uses and intensities
source("./R/02_Statistical_Analysis_merge_LandUses_Intensities.R")

# Create the vectors that hold the land-uses that we want to
# keep with different use intensities
land_uses_separate_1 <- c("Primary", "ISV", "Plantation forest")

# Create a vector with the land-uses where we want to merge the
#light and intense use intensities
land_uses_light_intense_1 <- c("Primary", "ISV", "Plantation forest")

# Merge landuse intensities
diversity_simpson <- Merge_landUses_and_intensities(dataset = diversity_simpson,
                                                    index = 1,
                        land_uses_separate_intensities = land_uses_separate_1,
                        land_uses_merge_light_intense = land_uses_light_intense_1,
                                                    "Primary Minimal use")

# Check number of sites
addmargins(table(diversity_simpson$LandUse.1, diversity_simpson$Kingdom), 2)
```

```
##
##                                     Animalia Plantae nePlantae Sum
##   Primary Minimal use                    325     249       199 773
##   Cropland All                           137      62        33 232
##   ISV Light-intense use                   52      74        33 159
##   ISV Minimal use                         81      36        38 155
##   MSV All                                 36      48        20 104
##   Pasture All                             29      38        47 114
##   Plantation forest Light-intense use    225     184        69 478
##   Plantation forest Minimal use           27      71        68 166
##   Primary Light-intense use              128     125        92 345
##   YSV All                                 67      71        75 213
```

Test for collinearity

```
# Get the function
source("https://highstat.com/Books/Book2/HighstatLibV10.R")

# Calculate the VIF
corvif(diversity_simpson[ , c("LandUse.1", "Kingdom")])
```

```
##
##
## Variance inflation factors
```

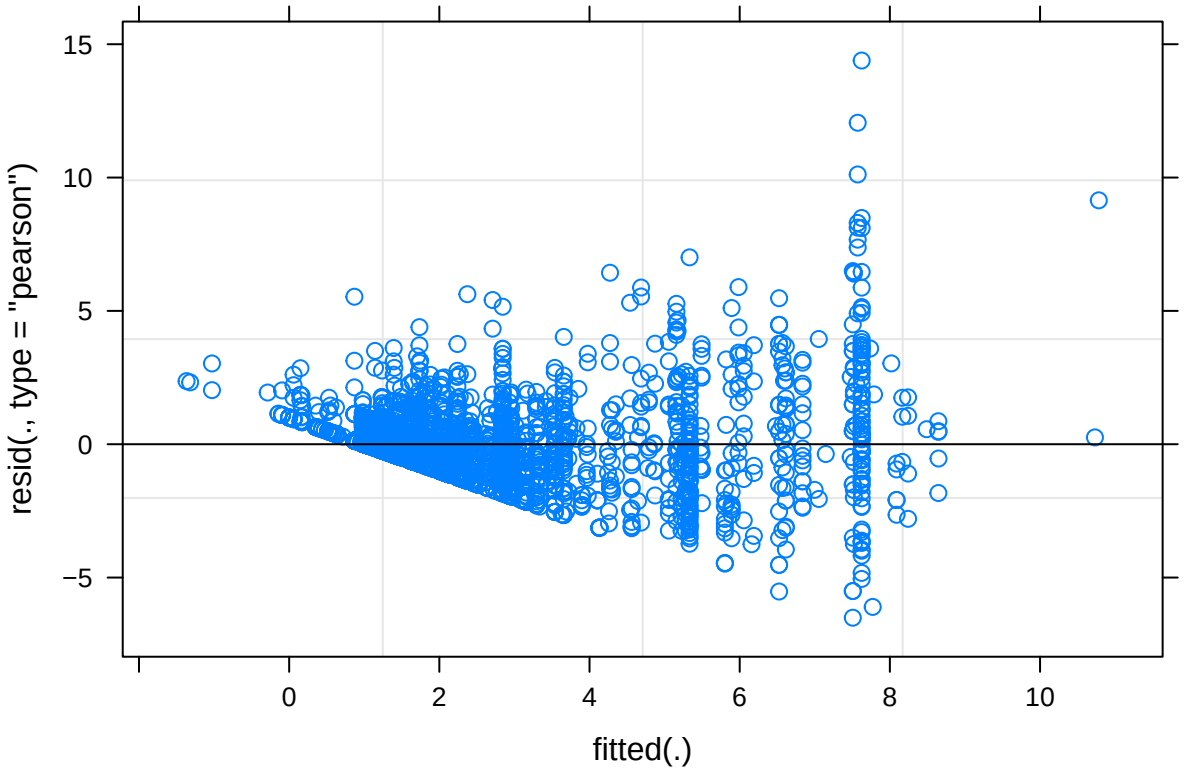
```
##
##                GVIF Df GVIF^(1/2Df)
## LandUse.1 1.065772  9     1.003545
## Kingdom   1.065772  2     1.016052
```

Choose random effects structure

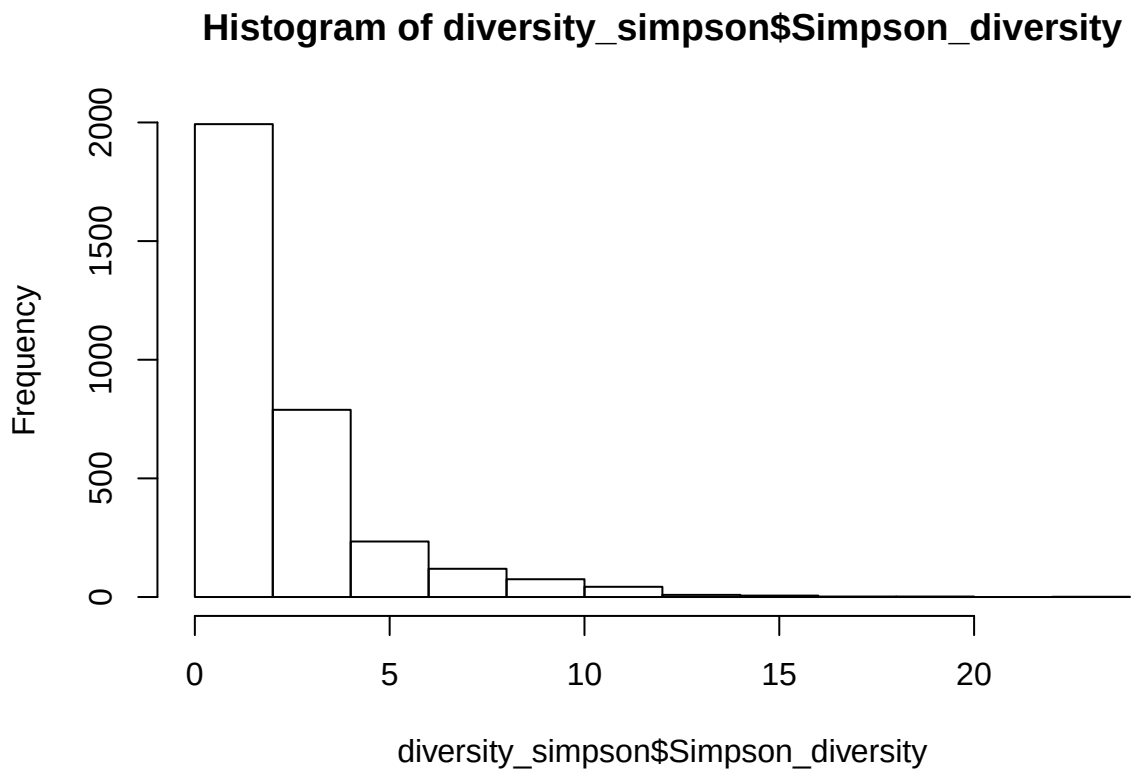
```
# Simplest random effects structure
m1 <- lmer(Simpson_diversity ~ LandUse.1 + Kingdom + LandUse.1:Kingdom +
             (1|SS) + (1|SSB), data = diversity_simpson)
```

```
## boundary (singular) fit: see ?isSingular
```

```
plot(m1)
```



```r
hist(diversity_simpson$Simpson_diversity)
```




```r
# The simplest model does not converge
```

12. Second attempt

```r
# I am going to merge all categories

# Create the vectors that hold the land-uses that we want to keep
#with different use intensities
land_uses_separate_2 <- "NA"
# Create a vector with the land-uses where we want to merge the
# light and intense use intensities
land_uses_light_intense_2 <- "NA"

diversity_simpson <- Merge_landUses_and_intensities(dataset = diversity_simpson,
                                                    index = 2,
               land_uses_separate_intensities = land_uses_separate_2,
               land_uses_merge_light_intense = land_uses_light_intense_2,
                                                    "Primary All")

addmargins(table(diversity_simpson$LandUse.2, diversity_simpson$Kingdom), 2)
```

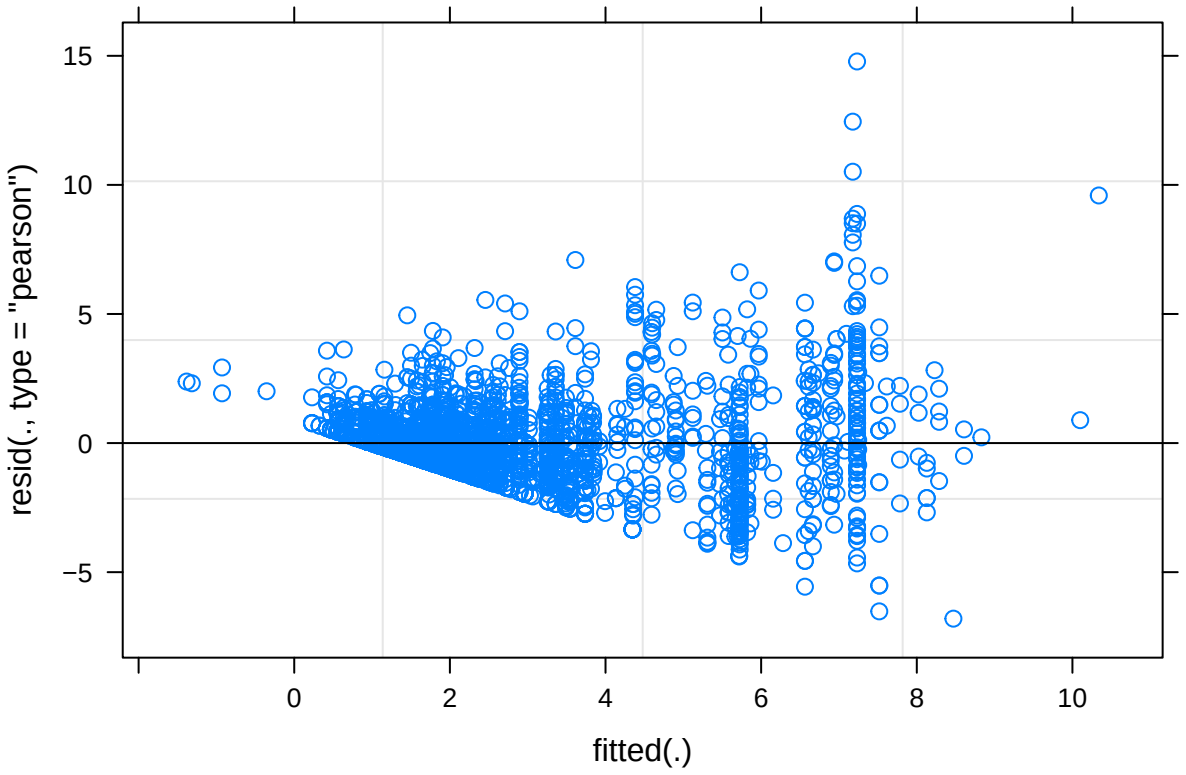
```
##
```

```
##                          Animalia Plantae nePlantae  Sum
##   Primary All                 462     433       342 1237
##   Cropland All                137      62        33  232
##   ISV All                     159     121        89  369
##   MSV All                      36      48        20  104
##   Pasture All                  29      38        47  114
##   Plantation forest All       252     258       143  653
##   YSV All                      67      71        75  213
```

```
m2 <- lmer(Simpson_diversity ~ LandUse.2 + Kingdom + LandUse.2:Kingdom +
             (1|SS) + (1|SSB), data = diversity_simpson)
```

```
## boundary (singular) fit: see ?isSingular
```

```
plot(m2)
```



I am going to try to remove the MSV, since it only has 20 sites in total for nePlants

```
# Remove MSV records
diversity_simpson1 <- diversity_simpson %>%
  subset(LandUse.2 != "MSV All") %>% droplevels()

# Run the model
m3 <- lmer(Simpson_diversity ~ LandUse.2 + Kingdom + LandUse.2:Kingdom +
             (1|SS) + (1|SSB), data = diversity_simpson1) #Is Singular
```

```
## boundary (singular) fit: see ?isSingular
```

```
plot(m3)
```