Abundance model

Daniela Linero

17/6/2020

1. Explore number of sites

```r
# Import site metrics of plants not dispersed by animals, plants dispersed by endozoochory and
# animals
diversity5 <- readRDS( file = "./output/cleaned_data/02_Statistical_Analysis_Site_metrics_combined_anim
```

Plants not dispersed by animals have enough sites to separate them into:

- Primary minimal, light and intense uses
- ISv minimal and light-intense uses
- Plantation forest minimal and light-intense uses
- Everything else with merge all intensities

I will run in with those classes and then compare if it's better to merge everything.

```r
# Create a vector with the land uses categories that have enough sites to separate
# them into different intensities
LandUse_divide <- c("Primary", "ISV", "Plantation forest")

fourth_model_data <- diversity5 %>%
  # make a level of Primary minimal
  mutate(

    # collapse primary forest and non-forest together into primary vegetation as these
    # aren't well distinguished
    Predominant_land_use = recode_factor(Predominant_land_use,
                                         "Primary forest" = "Primary",
                                         "Primary non-forest" = "Primary"),

    # indeterminate secondary veg and cannot decide are transformed into NA, urban too because
    # it has only 40 sites.
    Predominant_land_use = na_if(Predominant_land_use, "Secondary vegetation (indeterminate age)"),
    Predominant_land_use = na_if(Predominant_land_use, "Cannot decide"),
    Predominant_land_use = na_if(Predominant_land_use, "Urban"),


    # Give a shorter name to some land use categories
    Predominant_land_use = str_replace_all(Predominant_land_use,
                                           pattern = c("Young secondary vegetation" = "YSV",
                                                       "Intermediate secondary vegetation" = "ISV",
                                                       "Mature secondary vegetation" = "MSV")),
```

```r
    # Drop the Cannot decide intensity levels for the land-use categories that have
    # enough sites for the minimal, and light/intense
    Use_intensity = ifelse(Predominant_land_use %in% LandUse_divide & Use_intensity == "Cannot decide",
                           NA,
                           paste(Use_intensity)),

    # Join the intensity levels of light and intense for Plantation forest and ISV
    Use_intensity = ifelse((Predominant_land_use == "Plantation forest" |
                              Predominant_land_use == "ISV") &
                             (Use_intensity == "Intense use" |
                               Use_intensity == "Light use"),
                           str_replace_all(Use_intensity,
                                           pattern = c("Intense use" = "Light-intense use",
                                                       "Light use" = "Light-intense use")
                           paste(Use_intensity)),

    # Merge all the intensity levels for those land-use categories that don't have enough sites
    # in each land-use type/intensity combination
    Use_intensity = ifelse(Predominant_land_use %nin% LandUse_divide,
                           str_replace_all(Use_intensity, pattern = c("Intense use" = "All",
                                                                      "Light use" = "All",
                                                                      "Minimal use" = "All",
                                                                      "Cannot decide" = "All")),
                           paste(Use_intensity)),


    # Paste the land-use classes and intensity levels
    LandUse = ifelse(Predominant_land_use != "NA" & Use_intensity != "NA",
                     paste(Predominant_land_use, Use_intensity),
                     NA),

    # set reference level
    LandUse = factor(LandUse),
    LandUse = relevel(LandUse, ref = "Primary Minimal use")

  )
```

2. Test for collinearity

Since I'm going to explore the collinearity between categorical variables, I'm going to use the Generalized variance Inflation Factors function provided by Zuur et al., (2009)

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

# Calculate the VIF
corvif(fourth_model_data[ , c("LandUse", "Kingdom")])
```

```
##
##
## Variance inflation factors
##
##               GVIF Df GVIF^(1/2Df)
```

```
## LandUse 1.101693 10    1.004854
## Kingdom 1.101693  2    1.024507
```

3. Complete cases

```
# Drop sites that don't have abundance measures or land-use data
fourth_model_data1 <- drop_na(fourth_model_data,
                              Total_abundance, LandUse) %>% droplevels()


# Check the number of sites for animals and plants
addmargins(table(fourth_model_data1$LandUse, fourth_model_data1$Kingdom), 2)
```

```
##
##                                     Animalia Plantae nePlantae  Sum
##    Primary Minimal use                   423     314       303 1040
##    Cropland All                          248      88        88  424
##    ISV Light-intense use                  72      80        83  235
##    ISV Minimal use                        87      51        47  185
##    MSV All                               178      67        21  266
##    Pasture All                           177      57        49  283
##    Plantation forest Light-intense use   357     259       189  805
##    Plantation forest Minimal use          28      81        81  190
##    Primary Intense use                    50      93        93  236
##    Primary Light use                     214      89        79  382
##    YSV All                               133     102       101  336
```

4. Transform rescaled abundance measures

Abundance data usually display a nonnormal error distribution because they have a positive mean-variance relationship and are zero-inflated (Purvis et al., 2018). Given that some abudance measures are not integers (some are relative abudance or densities), I am not going to model the abudance with a Poisson distribution, but I'm going to transform it in order to meet the assumptions of linear mixed models.

```
# Transform RescaledAbundance.
fourth_model_data2 <- mutate(fourth_model_data1,
                             logAbundance = log(RescaledAbundance + 1),
                             sqrtAbundance = sqrt(RescaledAbundance)
)
```

5. Select random effects structure

To select the random-effects structure, we use the method recommended by (Zuur et al., 2009) of taking the most complex fixed-effects structure, including all interactions, that will be tested in the second stage of modelling, and use it to compare different random-effects structures.

```
m4_1 <- lmer(logAbundance ~ LandUse + Kingdom + LandUse:Kingdom +
             (1|SS) + (1|SSB), data = fourth_model_data2)
```

```
m4_2 <- lmer(logAbundance ~ LandUse + Kingdom + LandUse:Kingdom +
             (1|SS) + (1|SSB) + (1|Source_ID), data = fourth_model_data2)
```

```r
m4_3 <- lmer(logAbundance ~ LandUse + Kingdom + LandUse:Kingdom +
               (1+LandUse|SS) + (1|SSB), data = fourth_model_data2)
```

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

```r
m4_5 <- lmer(logAbundance ~ LandUse + Kingdom + LandUse:Kingdom +
               (1+Predominant_land_use|SS) + (1|SSB), data = fourth_model_data2)
```

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

```r
# Compare the models that converged
AIC(m4_1, m4_2)
```

```
##      df       AIC
## m4_1 36 -3532.507
## m4_2 37 -3538.588
```

6. Select fixed effect structure

```r
# See the significance of the terms in the model.
Anova(m4_2)
```
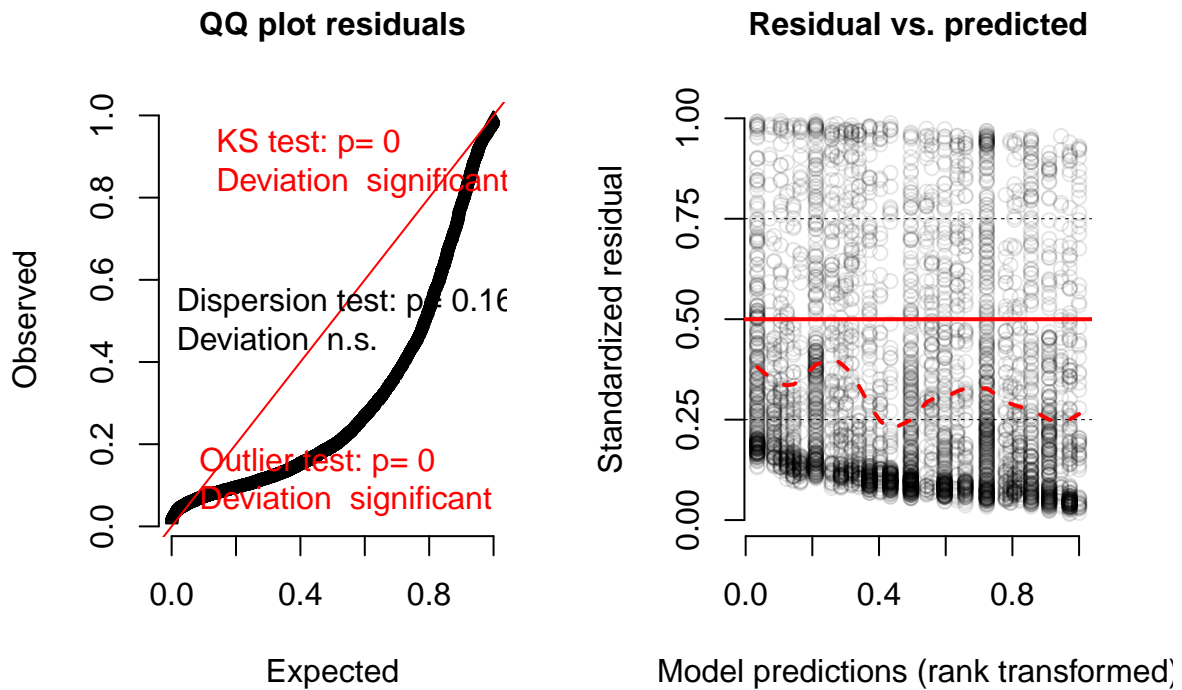
```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: logAbundance
##                  Chisq Df Pr(>Chisq)
## LandUse          52.110 10  1.089e-07 ***
## Kingdom          53.924  2  1.952e-12 ***
## LandUse:Kingdom 113.135 20  5.247e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Land_use:kingdom interaction is significant, so I won't remove this term from the model.

7. plot residuals

```r
simulationOutput4 <- simulateResiduals(fittedModel = m4_2, plot = TRUE)
```

DHARMa residual diagnostics

QQ plot residuals

KS test: p= 0
Deviation significant

Dispersion test: p= 0.16
Deviation n.s.

Outlier test: p= 0
Deviation significant

Observed

Expected

Residual vs. predicted

Standardized residual

Model predictions (rank transformed)

8. Model estimates

Table with confidence intervals

```
CI_m4_2 <- read.csv("./output/summary_m4_2.csv", header = TRUE, sep = ";")
kable(CI_m4_2, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

X	Estimate	Std..Error	t.value	Actual.estimate	CI.lower	CI.upper
(Intercept)	0.323	0.024	13.351	0.323	0.371	0.276
LandUseCropland All	-0.004	0.015	-0.287	0.319	0.349	0.289
LandUseISV Light-intense use	0.003	0.023	0.142	0.326	0.371	0.282
LandUseISV Minimal use	-0.015	0.022	-0.679	0.308	0.351	0.266
LandUseMSV All	0.056	0.021	2.595	0.379	0.421	0.337
LandUsePasture All	-0.032	0.021	-1.512	0.292	0.333	0.251
LandUsePlantation forest Light-intense use	-0.027	0.016	-1.742	0.296	0.327	0.265
LandUsePlantation forest Minimal use	-0.066	0.040	-1.643	0.257	0.336	0.178
LandUsePrimary Intense use	0.006	0.028	0.235	0.330	0.384	0.276
LandUsePrimary Light use	0.029	0.016	1.772	0.352	0.383	0.320
LandUseYSV All	-0.002	0.019	-0.084	0.322	0.359	0.284
KingdomPlantae	-0.077	0.037	-2.099	0.246	0.318	0.174
KingdomnePlantae	-0.120	0.037	-3.272	0.203	0.275	0.131
LandUseCropland All:KingdomPlantae	0.028	0.026	1.105	0.270	0.320	0.220
LandUseISV Light-intense use:KingdomPlantae	0.085	0.031	2.738	0.334	0.395	0.274
LandUseISV Minimal use:KingdomPlantae	0.069	0.033	2.070	0.300	0.366	0.235
LandUseMSV All:KingdomPlantae	0.088	0.033	2.655	0.390	0.455	0.325
LandUsePasture All:KingdomPlantae	0.119	0.034	3.492	0.333	0.400	0.267
LandUsePlantation forest Light-intense use:KingdomPlantae	0.139	0.022	6.204	0.358	0.402	0.314
LandUsePlantation forest Minimal use:KingdomPlantae	0.141	0.046	3.068	0.321	0.412	0.231
LandUsePrimary Intense use:KingdomPlantae	-0.044	0.037	-1.181	0.209	0.281	0.137
LandUsePrimary Light use:KingdomPlantae	0.014	0.027	0.506	0.289	0.342	0.235
LandUseYSV All:KingdomPlantae	0.023	0.027	0.851	0.268	0.321	0.214
LandUseCropland All:KingdomnePlantae	0.043	0.025	1.678	0.241	0.291	0.191
LandUseISV Light-intense use:KingdomnePlantae	0.057	0.031	1.865	0.264	0.324	0.203
LandUseISV Minimal use:KingdomnePlantae	0.077	0.034	2.255	0.265	0.331	0.198
LandUseMSV All:KingdomnePlantae	-0.055	0.043	-1.265	0.204	0.289	0.119
LandUsePasture All:KingdomnePlantae	0.196	0.035	5.639	0.368	0.436	0.300
LandUsePlantation forest Light-intense use:KingdomnePlantae	0.099	0.023	4.354	0.275	0.319	0.230
LandUsePlantation forest Minimal use:KingdomnePlantae	0.090	0.046	1.956	0.227	0.317	0.137
LandUsePrimary Intense use:KingdomnePlantae	0.029	0.037	0.799	0.239	0.311	0.167
LandUsePrimary Light use:KingdomnePlantae	-0.001	0.028	-0.029	0.231	0.285	0.176
LandUseYSV All:KingdomnePlantae	0.052	0.027	1.890	0.253	0.307	0.199

9. Merge all land-use intensities

```r
# make a copy of the database:
WH_merged <- fourth_model_data2

# Same model but with log
m4_2l <- lmer(logAbundance~ LandUse + Kingdom + LandUse:Kingdom +
              (1|SS) + (1|SSB) + (1|Source_ID), data = WH_merged)

# Merge land-use intensities for the fourth_model_data2
WH_merged <- WH_merged %>% mutate(

  # Merge all the intensity levels for all land-use categories
  Use_intensity = ifelse(Use_intensity != "NA", "All", "NA"),

  # Paste the land-use classes and intensity levels
  LandUse = ifelse(Predominant_land_use != "NA" & Use_intensity != "NA",
            paste(Predominant_land_use, Use_intensity),
            NA),

  # set reference level
  LandUse = factor(LandUse),
  LandUse = relevel(LandUse, ref = "Primary All")
```

```
)

# same model but with log
m4_2_ml <- lmer(logAbundance ~ LandUse + Kingdom + LandUse:Kingdom +
                (1|SS) + (1|SSB) + (1|Source_ID), data = WH_merged)
```

Next, I will check if I've lost a significant amount of explanatory power by merging some land-use intensities.If I have, I want to keep the more complex model. If I haven't lost a significant amount of explanatory power, then I can keep this simpler model.

```
anova(m4_2_ml, m4_2l, test = "F")
```

```
## refitting model(s) with ML (instead of REML)


## Data: WH_merged
## Models:
## m4_2_ml: logAbundance ~ LandUse + Kingdom + LandUse:Kingdom + (1 | SS) +
## m4_2_ml:     (1 | SSB) + (1 | Source_ID)
## m4_2l: logAbundance ~ LandUse + Kingdom + LandUse:Kingdom + (1 | SS) +
## m4_2l:     (1 | SSB) + (1 | Source_ID)
##         npar    AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## m4_2_ml   25 -3736.8 -3577.1 1893.4  -3786.8
## m4_2l     37 -3737.0 -3500.8 1905.5  -3811.0 24.291 12    0.01856 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The LRT is significant, so by removing the interaction from the model, I did lose a significant amount of explanatory power. This means that I should keep the more complex model.

9. Plant contrast

What I'm going to do is a Plant contrast, that means I'm firts going to model the complex model (the one that assummes the reponse between endo and not-endo plants is significantly different), in this model I will use the Kingdom interaction that has 3 levels: animals and endo plants and not endo plants. Then I am going to model the null model that assumes there is no difference between endo and not-endo plants. In this model the Kingdom will have 2 levels: animals, plants. Then I will compare both models using an ANOVA

```
# make a copy of the database:
WI_merged <- fourth_model_data2
```

```
# According to the last steps, our complex model is:
m4_2 <- lmer(logAbundance ~ LandUse + Kingdom + LandUse:Kingdom +
             (1|SS) + (1|SSB) + (1|Source_ID), data = WI_merged)
```

```
# For the null model I am going to replace the nePlantae for Plantae
WI_merged <- WI_merged %>%

  # Replace nePlantae for Plantae
  mutate(
    Kingdom = recode_factor(Kingdom, "nePlantae" = "Plantae"),

# set reference level
```

7

```r
LandUse = factor(LandUse),
LandUse = relevel(LandUse, ref = "Primary Minimal use"),
Kingdom = factor(Kingdom),
Kingdom = relevel(Kingdom, ref = "Animalia"))


m4_2_m <- lmer(logAbundance ~ LandUse + Kingdom + LandUse:Kingdom +
                 (1|SS) + (1|SSB) + (1|Source_ID), data = WI_merged)
```

Compare complex and null model:

```r
anova(m4_2_m, m4_2, test = "F")


## refitting model(s) with ML (instead of REML)


## Data: WI_merged
## Models:
## m4_2_m: logAbundance ~ LandUse + Kingdom + LandUse:Kingdom + (1 | SS) +
## m4_2_m:     (1 | SSB) + (1 | Source_ID)
## m4_2: logAbundance ~ LandUse + Kingdom + LandUse:Kingdom + (1 | SS) +
## m4_2:      (1 | SSB) + (1 | Source_ID)
##        npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## m4_2_m   26 -3663.5 -3497.5 1857.8  -3715.5
## m4_2     37 -3737.0 -3500.8 1905.5  -3811.0 95.548 11  1.354e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10. Plot the results of the selected abundance model

```r
# Read r code from a file which contains the function to make the plot
source("./R/PlotErrBar_interactions.R")
source("./R/PlotErrBar_interactions_modified.R")

# Plot the differences between estimates
PlotErrBar_interactions_modi(model = m4_2, resp = "Abundance", Effect1 = "LandUse", Effect2 = "Kingdom"
                             ylims = c(-0.25,0.25), pointtype = c(16,17,18),blackwhite = FALSE)

# Plot the x label
text(x = c(0.8:10.8), y = -0.18, labels = c("Primary minimal use",
                                            "Cropland all uses",
                                            "ISV light-intense use",
                                            "ISV minimal use",
                                            "MSV all uses",
                                            "Pasture all uses",
                                            "Plantation forest light-intense use",
                                            "Plantation forest minimal use",
                                            "Primary intense use",
                                            "Primary light use",
                                            "YSV all uses"), srt = 18, cex= 0.9)
```