Analysis: insecta diversity data in China and Brazil

Lilei Chen

2021/8/5

This R markdown included every step I have done. I changed the categories to suit a more reasonable analysis, I have also merged sites to deal with the situations such as ten traps for same land use, use intensity, cooridates and start/end times as treated as 10 sites, which is better to treat them as one. The (De Palma, 2019)

## Load the packages

Load the packages we are going to use.

library(dplyr) # for easy data manipulation

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

library(tidyr) # ditto

## Warning: package 'tidyr' was built under R version 4.0.5

library(lme4) # for mixed effects models

## Warning: package 'lme4' was built under R version 4.0.5

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

library(sjPlot) #for visualizing results

## Warning: package 'sjPlot' was built under R version 4.0.5

## Registered S3 methods overwritten by 'parameters':  
## method from   
## as.double.parameters\_kurtosis datawizard  
## as.double.parameters\_skewness datawizard  
## as.double.parameters\_smoothness datawizard  
## as.numeric.parameters\_kurtosis datawizard  
## as.numeric.parameters\_skewness datawizard  
## as.numeric.parameters\_smoothness datawizard  
## print.parameters\_distribution datawizard  
## print.parameters\_kurtosis datawizard  
## print.parameters\_skewness datawizard  
## summary.parameters\_kurtosis datawizard  
## summary.parameters\_skewness datawizard

library(car) # for getting anova tables with significance values

## Warning: package 'car' was built under R version 4.0.5

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(ggplot2) # for plotting

## Warning: package 'ggplot2' was built under R version 4.0.5

library(DHARMa) # for model criticism plots

## Warning: package 'DHARMa' was built under R version 4.0.5

## This is DHARMa 0.4.3. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa') Note: Syntax of plotResiduals has changed in 0.3.0, see ?plotResiduals for details

library(effects) # for extracting model effects

## Warning: package 'effects' was built under R version 4.0.5

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

library(merTools) # using it for extracting estimates for plotting

## Warning: package 'merTools' was built under R version 4.0.5

## Loading required package: arm

## Warning: package 'arm' was built under R version 4.0.5

## Loading required package: MASS

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

##   
## arm (Version 1.11-2, built: 2020-7-27)

## Working directory is F:/Code\_PREDICTS/PREDICTS

##   
## Attaching package: 'arm'

## The following object is masked from 'package:car':  
##   
## logit

## Read in and process the diversity data

Read in the raw data from PREDICTS.

# set working directory  
setwd("F:/Code\_PREDICTS/data")  
  
# read rds file  
diversity <- readRDS("diversity-2021-08-01-02-33-01.rds")

## Insecta data in China and Brazil

Now we have data from PREDICTS, this study focus on insecta in China and Brazil, therefore, the next step is to select such data we want.

Insecta\_diversity <- filter(diversity, Class == "Insecta") %>%  
 filter(Country == "China"| Country == "Brazil")

Let’s have a look about the data to see if filter()worked.

unique(Insecta\_diversity$Country)

## [1] Brazil China   
## 246 Levels: Afghanistan 脜land Islands Albania Algeria ... Zimbabwe

## Land use and use intensity

We are focusing on land use and use intensity, the columns of Predominant\_habitat and Use\_intensity contain information of land use and use intensity. let’s check the details.

##   
## Primary forest Primary non-forest Young secondary vegetation  
## FALSE 35048 34398 41667  
## TRUE 298 98 1137  
##   
## Intermediate secondary vegetation Mature secondary vegetation  
## FALSE 16937 52231  
## TRUE 704 0  
##   
## Secondary vegetation (indeterminate age) Plantation forest Pasture  
## FALSE 2202 20111 46805  
## TRUE 1968 770 492  
##   
## Cropland Urban Cannot decide  
## FALSE 12431 258 0  
## TRUE 3116 6387 1418

It seems that China has 0 mature secondary vegetation and fair less numbers of primary forest and primary non-forest. We need to combine some of these levels for further analysis.

##   
## Minimal use Light use Intense use Cannot decide  
## FALSE 180167 78257 2832 832  
## TRUE 4128 5920 1616 4724

Although the data of use intensity in China is less than data in Brazil, it is able to conduct the analysis. Based on the data, We are going to combine Primary forest and Primary non-forest as Primary vegetation. Young secondary vegetation, Intermediate secondary vegetation, Mature secondary vegetation and Secondary vegetation (indeterminate age) will be combined as Secondary vegetation. The Cannot decide in Land use and use intensity will be put as NA.

# combine to primary vegetation and secondary vegetation  
relevled\_diversity <- Insecta\_diversity %>%  
 # combine primary forest and primary non-forest into primary vegetation  
 mutate(  
 LandUse = ifelse(Predominant\_habitat == "Primary forest" |  
 Predominant\_habitat == "Primary non-forest",  
 "Primary vegetation",  
 paste(Predominant\_habitat)),  
 # combine all secondary vegetation  
 LandUse = ifelse(Predominant\_habitat == "Young secondary vegetation"|  
 Predominant\_habitat == "Intermediate secondary vegetation"|  
 Predominant\_habitat == "Mature secondary vegetation"|  
 Predominant\_habitat == "Secondary vegetation (indeterminate age)",  
 "Secondary vegetation",  
 paste(LandUse)),  
   
 # change cannot decide into NA  
 LandUse = ifelse(Predominant\_habitat == "Cannot decide",  
 NA,   
 paste(LandUse)),  
 # relevel the factor so that Primary vegetation will be the ref (so that it is the intercept term in models)  
 LandUse = factor(LandUse),  
 LandUse = relevel(LandUse, ref = "Primary vegetation"),  
   
 # change cannot decide into NA in UseIntensity  
 UseIntensity = ifelse(Use\_intensity == "Cannot decide",  
 NA,  
 paste(Use\_intensity)),  
 # relevel the factor so that Minimal use is the first level (so that it is the intercept term in models)  
 UseIntensity = factor(UseIntensity),  
 UseIntensity = relevel(UseIntensity, ref = "Minimal use")  
 )  
  
# check 'LandUse' and 'UseIntensity'  
table(relevled\_diversity$LandUse)

##   
## Primary vegetation Cropland Pasture   
## 69842 15547 47297   
## Plantation forest Secondary vegetation Urban   
## 20881 116846 6645

table(relevled\_diversity$UseIntensity)

##   
## Minimal use Intense use Light use   
## 184295 4448 84177

## Correct sampling effort

Now we have releveled data with suitable categories for further analysis. Next, we need abundance and richness data of insecta. To do that, the rescale of sampling effort is necessary for diversity measurements which are sensitive to sampling effort. We rescale sampling effort for each study to have a maximum value of 1. Let’s have a look if there are studies with missing sampling effort.

studies\_check <- relevled\_diversity %>%  
 # filter the rows where Sampling efforts are NA  
 filter(is.na(Sampling\_effort)) %>%  
 # keep only unique studies  
 distinct(SS) %>%  
 # pull the vector  
 pull(SS)  
  
relevled\_diversity %>%  
 # filter the rows where studies are those that had missing sampling efforts (above)  
 filter(SS %in% studies\_check) %>%  
 # drop missing levels  
 droplevels() %>%  
 # pull out the sampling efforts of these studies  
 pull(Sampling\_effort)%>%  
 # summaries to check that ALL the data are NAs  
 summary()

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
##

There is no NA for sampling effort of all studies so we don’t need to replace any value. Otherwise we should replace NAs as 1 assuming that the sampling efforts don’t vary within a study and correct those that need it.

relevled\_diversity <- relevled\_diversity %>%  
   
 # group by Study  
 group\_by(SS) %>%  
   
 # check how many sampling efforts there are in each study  
 mutate(n\_sample\_effort = n\_distinct(Sampling\_effort),  
 # get the maximum sampling effort for the studies  
   
 max\_sample\_effort = max(Sampling\_effort)  
 ) %>%  
   
 ungroup() %>%  
   
 # if the study has more than one sampling effort, correct the abundance  
   
 # so if there's only one sampling effort, then create a 'dummy sampling effort' of 1 so that we don't change the abundances when we do the divisions. Otherwise, we give it the maximum sampling effort.  
 mutate(DividingEffort = ifelse(n\_sample\_effort == 1, 1, max\_sample\_effort)) %>%  
   
 # if the diversity metric isn't sensitive to the effort, then we'll change the value to 1 too (so we won't end up changing the measurement), otherwise leave it as it is  
 mutate(DividingEffort = ifelse(Diversity\_metric\_is\_effort\_sensitive == FALSE, 1, DividingEffort)) %>%  
   
 # now let's create the effort corrected measurement by dividing the abundances by the sampling efforts  
 mutate(Corrected\_sampling\_effort = Sampling\_effort / max\_sample\_effort,  
 Effort\_corrected\_measurement = Measurement \* Corrected\_sampling\_effort)   
  
summary(relevled\_diversity$Corrected\_sampling\_effort)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.125 1.000 1.000 0.988 1.000 1.000

We need to test if the Corrected\_sampling\_effort is working, I know that LC1\_2015\_\_Su 1 has multiple sampling efforts during the data inputting, let’s check it to see if it works.

relevled\_diversity %>%   
 # filter out the test study  
 filter(SS == "LC1\_2015\_\_Su 1") %>%   
 # select out the columns to check  
 dplyr::select(Measurement, Effort\_corrected\_measurement, Sampling\_effort, max\_sample\_effort, Corrected\_sampling\_effort)

## # A tibble: 5,490 x 5  
## Measurement Effort\_corrected\_measurement Sampling\_effort max\_sample\_effort  
## <dbl> <dbl> <dbl> <dbl>  
## 1 0 0 5 30  
## 2 0 0 5 30  
## 3 0 0 5 30  
## 4 0 0 5 30  
## 5 0 0 5 30  
## 6 0 0 5 30  
## 7 0 0 5 30  
## 8 0 0 5 30  
## 9 0 0 5 30  
## 10 0 0 5 30  
## # ... with 5,480 more rows, and 1 more variable:  
## # Corrected\_sampling\_effort <dbl>

It is working, now let’s check studies that have same sampling effort, here I test LC1\_2021\_\_Ge 1 since I know it has the same sampling efforts during data inputting.

relevled\_diversity %>%   
 filter(SS == "LC1\_2021\_\_Ge 1") %>%   
 dplyr::select(Measurement, Effort\_corrected\_measurement)

## # A tibble: 350 x 2  
## Measurement Effort\_corrected\_measurement  
## <dbl> <dbl>  
## 1 0 0  
## 2 0 0  
## 3 0 0  
## 4 0 0  
## 5 1 1  
## 6 0 0  
## 7 0 0  
## 8 0 0  
## 9 0 0  
## 10 1 1  
## # ... with 340 more rows

## Merge sites

Next we’ll merge any sites that are within the same land-use type and that have identical coordinates, start and end dates. This is to deal with such situations, for instance, authors might put 10 pit traps seperately as ten sampling points. Then, it looks like we have mores sites but actually they have same land use, same use intensity and same coordinates, which means they are just one site.

relevled\_diversity <- relevled\_diversity %>%  
   
 # group by aspects of the sites that should be identical if we need to merge the abundances  
 # I only want to merge abundances if they are within the same study and block  
 # as I'm assuming that even if the locations and sampling times are the same, if the blocks or studies are different, then there is some good reason for this.  
 group\_by(Source\_ID, Study\_number, Study\_name, Block,  
 #diversity metric type  
 Diversity\_metric, Diversity\_metric\_type, Diversity\_metric\_unit,  
 Diversity\_metric\_is\_effort\_sensitive,  
   
 #details of the sites  
 Predominant\_habitat, Use\_intensity, Years\_since\_fragmentation\_or\_conversion,  
   
 #details of the sampling method  
 Sampling\_method, Sampling\_effort\_unit,  
   
 #species identity  
 Study\_common\_taxon, Rank\_of\_study\_common\_taxon,  
 Taxon\_number, Taxon\_name\_entered,  
 Indication, Parsed\_name,  
 Best\_guess\_binomial, COL\_ID, Taxon, Name\_status,  
 Rank, Kingdom, Phylum, Class, Order, Family, Genus, Species,  
 Higher\_taxon,  
   
 #site location  
 Longitude, Latitude,  
   
 #sampling time  
 Sample\_start\_earliest, Sample\_end\_latest, Sample\_date\_resolution) %>%  
   
 # if the diversity metric is occurrence:  
 # if it is present at all, give it a 1, if it is always absent, give it a 0,  
 # otherwise (if the metric is either abundance or species richness):  
 # calculate the weighted abundance/richness for each taxonomic group, weighted by sampling effort  
   
 mutate(merged\_diversity =   
 ifelse(Diversity\_metric\_type == "Occurrence",  
 # if any of the occurrence values are 1, `any` will return TRUE. If you sum a logical, TRUE becomes 1 and FALSE becomes 0  
 sum(any(Effort\_corrected\_measurement > 0)),  
   
 # note that since we've already corrected the sampling effort, this is essentially a mean rather than a weighted mean for abundance measurements. It's a weighted mean for species richness though where sampling efforts vary.  
 stats::weighted.mean(x = Effort\_corrected\_measurement,  
 w = Corrected\_sampling\_effort))  
 )  
  
# pull out the grouping data (so we can double check how many records we're merging for each)  
group\_dat <- relevled\_diversity %>%   
 group\_data() %>%  
 mutate(nvals\_merged = lengths(.rows),  
 merge\_ID = row\_number())  
  
# ungroup the relevled\_diversity data for future use  
relevled\_diversity <- ungroup(relevled\_diversity)  
  
# create a dataset where we can extract just the merged data if we want to  
diversity\_merged <- relevled\_diversity %>%  
 left\_join(group\_dat)

## Joining, by = c("Source\_ID", "Study\_number", "Study\_name", "Diversity\_metric", "Diversity\_metric\_unit", "Diversity\_metric\_type", "Diversity\_metric\_is\_effort\_sensitive", "Sampling\_method", "Sampling\_effort\_unit", "Block", "Sample\_start\_earliest", "Sample\_end\_latest", "Sample\_date\_resolution", "Predominant\_habitat", "Use\_intensity", "Years\_since\_fragmentation\_or\_conversion", "Longitude", "Latitude", "Taxon\_number", "Taxon\_name\_entered", "Indication", "Parsed\_name", "COL\_ID", "Taxon", "Name\_status", "Rank", "Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species", "Higher\_taxon", "Study\_common\_taxon", "Rank\_of\_study\_common\_taxon", "Best\_guess\_binomial")

# check that the merging has worked (row numbers should be equal right now)  
nrow(relevled\_diversity) == nrow(diversity\_merged)

## [1] TRUE

Now let’s test that the sites have been merged correctly.

test\_data <- diversity\_merged %>%  
 filter(nvals\_merged > 1) %>%  
 distinct(merge\_ID, .keep\_all = TRUE)  
  
test1 <- diversity\_merged %>%  
 filter(merge\_ID == test\_data$merge\_ID[1]) %>%  
 dplyr::select(SS, SSB,  
 Diversity\_metric, Diversity\_metric\_type, Diversity\_metric\_unit,  
 Predominant\_habitat, Use\_intensity, Years\_since\_fragmentation\_or\_conversion,  
 Sampling\_method, Sampling\_effort\_unit,  
 Study\_common\_taxon, Rank\_of\_study\_common\_taxon,  
 Taxon\_name\_entered,  
 Best\_guess\_binomial,  
 Longitude, Latitude,  
 Sample\_start\_earliest, Sample\_end\_latest, Sample\_date\_resolution,  
 Effort\_corrected\_measurement,  
 Corrected\_sampling\_effort,  
 merged\_diversity,  
 .rows,  
 nvals\_merged,  
 merge\_ID  
 )  
  
test1

## # A tibble: 10 x 25  
## SS SSB Diversity\_metric Diversity\_metri~ Diversity\_metri~  
## <fct> <fct> <fct> <fct> <fct>   
## 1 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 2 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 3 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 4 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 5 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 6 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 7 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 8 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 9 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## 10 AD2\_2013\_\_Schmidt 1 AD2\_2~ occurrence Occurrence presence/absence  
## # ... with 20 more variables: Predominant\_habitat <fct>, Use\_intensity <fct>,  
## # Years\_since\_fragmentation\_or\_conversion <dbl>, Sampling\_method <fct>,  
## # Sampling\_effort\_unit <fct>, Study\_common\_taxon <fct>,  
## # Rank\_of\_study\_common\_taxon <fct>, Taxon\_name\_entered <fct>,  
## # Best\_guess\_binomial <fct>, Longitude <dbl>, Latitude <dbl>,  
## # Sample\_start\_earliest <date>, Sample\_end\_latest <date>,  
## # Sample\_date\_resolution <fct>, Effort\_corrected\_measurement <dbl>, ...

paste(test\_data$SS[8000])

## [1] "AD2\_2017\_\_Novais 1"

test2 <- diversity\_merged %>%  
 filter(merge\_ID == test\_data$merge\_ID[8000]) %>%  
 dplyr::select(SS, SSB,  
 Diversity\_metric, Diversity\_metric\_type, Diversity\_metric\_unit,  
 Predominant\_habitat, Use\_intensity, Years\_since\_fragmentation\_or\_conversion,  
 Sampling\_method, Sampling\_effort\_unit,  
 Study\_common\_taxon, Rank\_of\_study\_common\_taxon,  
 Taxon\_name\_entered,  
 Best\_guess\_binomial,  
 Longitude, Latitude,  
 Sample\_start\_earliest, Sample\_end\_latest, Sample\_date\_resolution,  
 Effort\_corrected\_measurement,  
 Corrected\_sampling\_effort,  
 merged\_diversity,  
 .rows,  
 nvals\_merged,  
 merge\_ID  
 )  
  
test2

## # A tibble: 6 x 25  
## SS SSB Diversity\_metric Diversity\_metric~ Diversity\_metri~  
## <fct> <fct> <fct> <fct> <fct>   
## 1 AD2\_2017\_\_Novais 1 "AD2\_2~ abundance Abundance individuals   
## 2 AD2\_2017\_\_Novais 1 "AD2\_2~ abundance Abundance individuals   
## 3 AD2\_2017\_\_Novais 1 "AD2\_2~ abundance Abundance individuals   
## 4 AD2\_2017\_\_Novais 1 "AD2\_2~ abundance Abundance individuals   
## 5 AD2\_2017\_\_Novais 1 "AD2\_2~ abundance Abundance individuals   
## 6 AD2\_2017\_\_Novais 1 "AD2\_2~ abundance Abundance individuals   
## # ... with 20 more variables: Predominant\_habitat <fct>, Use\_intensity <fct>,  
## # Years\_since\_fragmentation\_or\_conversion <dbl>, Sampling\_method <fct>,  
## # Sampling\_effort\_unit <fct>, Study\_common\_taxon <fct>,  
## # Rank\_of\_study\_common\_taxon <fct>, Taxon\_name\_entered <fct>,  
## # Best\_guess\_binomial <fct>, Longitude <dbl>, Latitude <dbl>,  
## # Sample\_start\_earliest <date>, Sample\_end\_latest <date>,  
## # Sample\_date\_resolution <fct>, Effort\_corrected\_measurement <dbl>, ...

Now let’s check some studies that didn’t need merging.

test\_data <- diversity\_merged %>%  
 filter(nvals\_merged == 1) %>%  
 distinct(merge\_ID, .keep\_all = TRUE)  
  
paste(test\_data$SS[1])

## [1] "AD1\_2002\_\_Tonhasca 1"

test3 <- diversity\_merged %>%  
 filter(merge\_ID == test\_data$merge\_ID[1]) %>%  
 dplyr::select(SS, SSB,  
 Diversity\_metric, Diversity\_metric\_type, Diversity\_metric\_unit,  
 Predominant\_habitat, Use\_intensity, Years\_since\_fragmentation\_or\_conversion,  
 Sampling\_method, Sampling\_effort\_unit,  
 Study\_common\_taxon, Rank\_of\_study\_common\_taxon,  
 Taxon\_name\_entered,  
 Best\_guess\_binomial,  
 Longitude, Latitude,  
 Sample\_start\_earliest, Sample\_end\_latest, Sample\_date\_resolution,  
 Effort\_corrected\_measurement,  
 Corrected\_sampling\_effort,  
 merged\_diversity,  
 .rows,  
 nvals\_merged,  
 merge\_ID  
 )  
  
test3

## # A tibble: 1 x 25  
## SS SSB Diversity\_metric Diversity\_metri~ Diversity\_metri~  
## <fct> <fct> <fct> <fct> <fct>   
## 1 AD1\_2002\_\_Tonhasca 1 "AD1\_~ abundance Abundance individuals   
## # ... with 20 more variables: Predominant\_habitat <fct>, Use\_intensity <fct>,  
## # Years\_since\_fragmentation\_or\_conversion <dbl>, Sampling\_method <fct>,  
## # Sampling\_effort\_unit <fct>, Study\_common\_taxon <fct>,  
## # Rank\_of\_study\_common\_taxon <fct>, Taxon\_name\_entered <fct>,  
## # Best\_guess\_binomial <fct>, Longitude <dbl>, Latitude <dbl>,  
## # Sample\_start\_earliest <date>, Sample\_end\_latest <date>,  
## # Sample\_date\_resolution <fct>, Effort\_corrected\_measurement <dbl>, ...

Everything is working as planed, let’s go for abundance and richness data.

## Calculate abundance and richness data

sites <- diversity\_merged %>%  
   
 # pull out only the merged diversity data  
 distinct(merge\_ID, .keep\_all = TRUE) %>%  
   
 # re-make SSB and SSBS values since we've now dropped a bunch of values  
 mutate(SS = paste(Source\_ID, Study\_number),  
 SSB = paste(SS, Block),  
 SSBS = paste(SSB, Site\_number)) %>%  
   
 # group by SSBS (each unique value corresponds to a unique site)  
 group\_by(SSBS) %>%  
   
 # now add up all the abundance measurements within each site  
 mutate(TotalAbundance = ifelse(Diversity\_metric\_type == "Abundance",  
 sum(merged\_diversity),  
 # if the diversity metric type isn't Abundance, then leave the TotalAbundance measurement as NA  
 NA),  
   
 SpeciesRichness = ifelse(Diversity\_metric\_type == "Species richness",  
 merged\_diversity,  
 # for abundance and occurrence measurements, count the number of unique species names that are present at the site   
 n\_distinct(Best\_guess\_binomial[merged\_diversity > 0]))) %>%  
   
 # ungroup  
 ungroup() %>%  
   
 # pull out unique sites  
 distinct(SSBS, .keep\_all = TRUE) %>%  
   
 # now group by Study ID  
 group\_by(SS) %>%  
   
 # pull out the maximum abundance for each study  
 mutate(MaxAbundance = max(TotalAbundance)) %>%  
   
 # ungroup  
 ungroup() %>%  
   
 # now rescale total abundance, so that within each study, abundance varies from 0 to 1.  
 mutate(RescaledAbundance = TotalAbundance/MaxAbundance)

## Start to model

## collinearity

Check if there is collinearity for the variables we are going to use.

source("https://highstat.com/Books/Book2/HighstatLibV10.R")  
  
corvif(sites[ , c("LandUse", "UseIntensity", "Country")])

##   
##   
## Variance inflation factors  
##   
## GVIF Df GVIF^(1/2Df)  
## LandUse 3.186850 5 1.122887  
## UseIntensity 1.970537 2 1.184803  
## Country 1.664981 1 1.290341

From the GVIFs, LandUse, Useintensity and Country all have values smaller than 3 which indicates the collinearity check is ok.

## complete cases

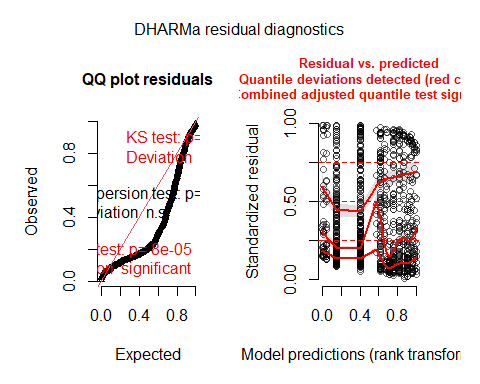
Before the modelling, we need to drop the row that has NA in our explanatory variables which is RescaledAbundance, LandUse, UseIntensity and Country.

model\_data\_ab <- drop\_na(sites,   
 RescaledAbundance, LandUse,  
 UseIntensity, Country)  
  
model\_data\_sr <- drop\_na(sites,   
 SpeciesRichness, LandUse,  
 UseIntensity, Country)

## Modeling

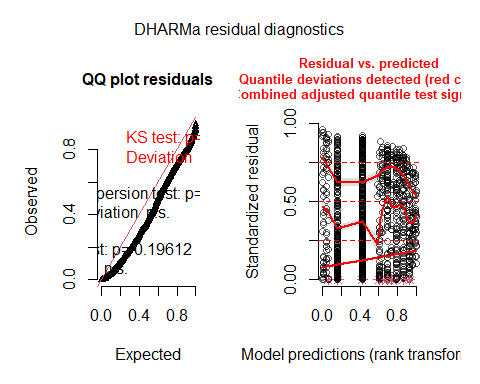
We will analyze the abundance data first. lmer model is used for abundance data, we should check the distribution of residuals first.

mab <- lmer(RescaledAbundance ~ LandUse + UseIntensity + Country + (1|SS), data = model\_data\_ab)  
simulationOutput <-simulateResiduals(fittedModel = mab, plot = F)  
plot(simulationOutput)# check the residuals plot



Based on the above plots, log transformation can help. To avoid log(0), 0.01 is added to ‘RescaledAbundance’.

model\_data\_ab <- mutate(model\_data\_ab,   
 logAbundance = log(0.01 + RescaledAbundance))  
mab <- lmer(logAbundance ~ LandUse + UseIntensity + Country + (1|SS), data = model\_data\_ab)  
simulationOutput <-simulateResiduals(fittedModel = mab, plot = F)  
plot(simulationOutput)# check the residuals plot



This is ok to work now. Next we build a model for richness data. Since it is count data, we will use glmer and poisson errors. For posisson error structure, we need to check for overdispersion first.

msr <- glmer(SpeciesRichness ~ LandUse + UseIntensity + Country + (1|SS) + (1|SSB), data = model\_data\_sr,family = "poisson")  
# check overdispersion  
# load in an overdispersion function from here: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#testing-for-overdispersioncomputing-overdispersion-factor  
overdisp\_fun <- function(model) {  
 rdf <- df.residual(model)  
 rp <- residuals(model,type="pearson")  
 Pearson.chisq <- sum(rp^2)  
 prat <- Pearson.chisq/rdf  
 pval <- pchisq(Pearson.chisq, df=rdf, lower.tail=FALSE)  
 c(chisq=Pearson.chisq,ratio=prat,rdf=rdf,p=pval)  
}  
  
# test the species richness model for overdispersion  
overdisp\_fun(msr)

## chisq ratio rdf p   
## 1501.3467993 0.9982359 1504.0000000 0.5144573

P value is 0.514 which indicates the model is fine. My study also want to compare a model in which Country interacts with the other fixed effects (\*), vs a model in which Country is only an additive effect (+).

mab2 <- lmer(logAbundance ~ (LandUse + UseIntensity) \* Country + (1|SS), data = model\_data\_ab)  
  
msr2 <- glmer(SpeciesRichness ~ (LandUse + UseIntensity) \* Country + (1|SS) + (1|SSB), data = model\_data\_sr, family = "poisson", control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 20000))) # the following code: control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 20000), is added since this model cannot converge  
  
  
anova(mab, mab2, refit = FALSE) # refit = FALSE is added due to the warning message: refitting model(s) with ML (instead of REML)

## Data: model\_data\_ab  
## Models:  
## mab: logAbundance ~ LandUse + UseIntensity + Country + (1 | SS)  
## mab2: logAbundance ~ (LandUse + UseIntensity) \* Country + (1 | SS)  
## npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)  
## mab 11 3526.5 3582.8 -1752.2 3504.5   
## mab2 18 3529.6 3621.8 -1746.8 3493.6 10.846 7 0.1455

anova(msr, msr2)

## Data: model\_data\_sr  
## Models:  
## msr: SpeciesRichness ~ LandUse + UseIntensity + Country + (1 | SS) + (1 | SSB)  
## msr2: SpeciesRichness ~ (LandUse + UseIntensity) \* Country + (1 | SS) + (1 | SSB)  
## npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)   
## msr 11 7533.7 7592.2 -3755.8 7511.7   
## msr2 18 7510.9 7606.7 -3737.5 7474.9 36.739 7 5.257e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Reporting the results

For abundance data,

sjPlot::tab\_model(mab2,   
 show.re.var= TRUE)

logAbundance

Predictors

Estimates

CI

p

(Intercept)

-1.43

-1.66 – -1.19

<0.001

LandUse [Cropland]

0.74

0.48 – 0.99

<0.001

LandUse [Pasture]

0.12

-0.13 – 0.37

0.361

LandUse [Plantationforest]

0.37

-0.04 – 0.79

0.080

LandUse [Secondaryvegetation]

0.46

0.20 – 0.72

0.001

LandUse [Urban]

0.15

-0.69 – 0.99

0.725

UseIntensity [Intenseuse]

-0.59

-1.04 – -0.15

0.008

UseIntensity [Light use]

-0.11

-0.32 – 0.10

0.290

Country [China]

0.26

-0.88 – 1.39

0.658

LandUse [Cropland] \*Country [China]

0.35

-0.90 – 1.60

0.585

LandUse [Pasture] \*Country [China]

0.73

-0.53 – 1.99

0.259

LandUse [Plantationforest] \* Country [China]

0.67

-0.63 – 1.97

0.312

LandUse [Secondaryvegetation] \* Country[China]

0.25

-0.92 – 1.42

0.679

LandUse [Urban] \* Country[China]

-0.02

-1.56 – 1.51

0.976

UseIntensity [Intenseuse] \* Country [China]

0.08

-0.56 – 0.73

0.799

UseIntensity [Light use]\* Country [China]

-0.56

-0.96 – -0.16

0.007

Random Effects

σ2

0.87

τ00 SS

0.56

ICC

0.39

N SS

85

Observations

1236

Marginal R2 / Conditional R2

0.077 / 0.440

For richness data,

sjPlot::tab\_model(msr2,   
 show.re.var= TRUE)

SpeciesRichness

Predictors

Incidence Rate Ratios

CI

p

(Intercept)

7.02

5.39 – 9.13

<0.001

LandUse [Cropland]

1.12

1.03 – 1.23

0.011

LandUse [Pasture]

0.78

0.71 – 0.85

<0.001

LandUse [Plantationforest]

0.90

0.80 – 1.00

0.045

LandUse [Secondaryvegetation]

1.13

1.04 – 1.22

0.004

LandUse [Urban]

1.14

0.90 – 1.43

0.275

UseIntensity [Intenseuse]

0.85

0.73 – 0.98

0.023

UseIntensity [Light use]

1.13

1.05 – 1.21

0.001

Country [China]

0.61

0.33 – 1.15

0.129

LandUse [Cropland] \*Country [China]

1.13

0.74 – 1.73

0.576

LandUse [Pasture] \*Country [China]

1.21

0.75 – 1.97

0.438

LandUse [Plantationforest] \* Country [China]

1.55

1.00 – 2.40

0.052

LandUse [Secondaryvegetation] \* Country[China]

0.74

0.54 – 1.03

0.075

LandUse [Urban] \* Country[China]

0.55

0.23 – 1.32

0.181

UseIntensity [Intenseuse] \* Country [China]

0.77

0.58 – 1.01

0.062

UseIntensity [Light use]\* Country [China]

0.63

0.52 – 0.77

<0.001

Random Effects

σ2

0.15

τ00 SSB

0.03

τ00 SS

1.31

ICC

0.90

N SS

103

N SSB

302

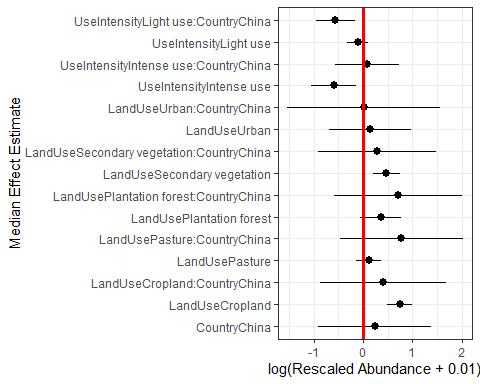
Observations

1515

Marginal R2 / Conditional R2

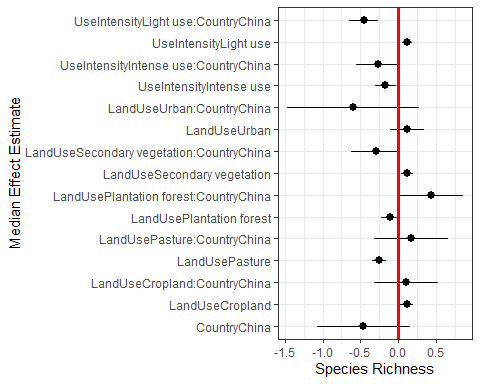
0.075 / 0.906

# gather the effects and confidence intervals using simulation  
sim\_eff\_ab <- FEsim(mab2, 1000)  
  
sim\_eff\_ab %>%  
 # drop the first row of the dataframe (the intercept value), since we're plotting the effects \_compared\_ to this value  
 slice(-c(1)) %>%  
 ggplot() +  
 # plot the coefficients plus/minus 95% confidence intervals (1.96 x standard deviation)  
 aes(x = term, ymin = median - 1.96 \* sd,  
 ymax = median + 1.96 \* sd, y = median) +  
 geom\_pointrange() +  
 # add the line for the Intercept (Minimal Primary vegetation) at 0 (everything is compared to this)  
 geom\_hline(yintercept = 0, size = I(1.1), color = I("red")) +  
 coord\_flip() +  
 theme\_bw() + labs(x = "Median Effect Estimate", y = "log(Rescaled Abundance + 0.01)")



Same plot for richness data

# gather the effects and confidence intervals using simulation  
sim\_eff\_sr <- FEsim(msr2, 1000)  
  
sim\_eff\_sr %>%  
 # drop the first row of the dataframe (the intercept value), since we're plotting the effecs \_compared\_ to this value  
 slice(-c(1)) %>%  
 ggplot() +  
 # plot the coefficiets plus/minus 95% confidence intervals (1.96 x standard deviation)  
 aes(x = term, ymin = median - 1.96 \* sd,  
 ymax = median + 1.96 \* sd, y = median) +  
 geom\_pointrange() +  
 # add the line for the Intercept (Minimal Primary vegetation) at 0 (everything is compared to this)  
 geom\_hline(yintercept = 0, size = I(1.1), color = I("red")) +  
 coord\_flip() +  
 theme\_bw() + labs(x = "Median Effect Estimate", y = "Species Richness")



## Reference

De Palma, A., Sanchez-Ortiz, K. and Purvis, A. (2019) Calculating the Biodiversity Intactness Index: the PREDICTS implementation. Zenodo. doi: 10.5281/ZENODO.3518067.