Appendix to Conspecific density dependence in plant communities: a theory-based toolkit for empirical studies

Authors here–

2023-10-31

Table of contents

# 1. Preface

This document serves as an appendix to the manuscript ***Conspecific density dependence in plant communities: a theory-based toolkit for empirical studies***, with the goal of providing code and guidance on conducting analyses of density dependence following the best practices outlined in the main manuscript.

# 2. Calculating neighborhood densities

## 2.1 Overview

This tutorial follows from the Section 2. “Modeling considerations when estimating CDD” of the main text:

|  |
| --- |
| Section 2. “Modeling considerations when estimating CDD” |
| For survival and growth models, defining the density metric and the size of local neighborhoods also require consideration. The density of conspecific and heterospecific neighbors surrounding a focal individual (or plot) can be measured within different distances (i.e., radii). Within a radius, neighbors should be weighted in the density summation with regard to their size and distance to the focal individuals, with the rationale that neighbors at a greater distance and of smaller size have smaller effects. Typically, weighting involves dividing the basal area or diameter by a linear function or an exponential function of distance to allow competitive effects of an individual of a given size to decline with distance (Uriarte et al. 2010; Canham et al. 2006). Biologically meaningful radii should be tested based on the size/life stage, vital rate of interest, and the likely agents of density dependence in the system. For a system of interest, we suggest comparing models with multiple distances (e.g., using maximum likelihood or information criterion) because forests may differ in how neighbors influence performance. |

Here, we demonstrate how to calculate the density of conspecific, heterospecific, and all neighbors surrounding a focal individual or plot for a given distance when XY coordinates are known. As a result, this section requires that the data set contains the location of mapped stems. If that is not available (as is common in seedling studies or smaller plots), continue to part 2 of this appendix.

We then demonstrate how to weight the calculation of neighborhood density by individual neighbor size (*i.e.,* basal area) and distance using an exponential decay function (one of many possible decay functions), allowing the competitive effects of density effects to saturate (Uriarte et al. 2010; Canham et al. 2006).

To assess which shape parameters of the exponential decay function are most appropriate for the data set, we fit models with multiple combinations of decay function values and compare models using log likelihood.

From a computational perspective, this approach can be relatively resource intensive both in terms of time and object size. It’s possible to make this approach more efficient by subdividing the data (*e.g.,* by plot) or using more efficient data structures such as [data.table](https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html).

We also note that alternative approaches allow the estimation of the effective scale of neighborhood interactions directly from data (see Barber et al. 2022). An excellent case study using Stan is [available here](https://mc-stan.org/users/documentation/case-studies/plantInteractions.html).

|  |
| --- |
| Note |
| The following code is adapted from the [latitudinalCNDD repository](https://github.com/LisaHuelsmann/latitudinalCNDD) by [Lisa Hülsmann](https://demographicecology.com/). |

## 2.2 Load libraries and data

# Load libraries  
library(tidyr) # For data manipulation  
library(dplyr) # For data manipulation  
library(ggplot2) # For data plotting  
library(spatstat.geom) # For spatial analysis  
library(here) # For managing working directories  
library(mgcv) # For fitting gams  
library(lubridate) # For calculating census intervals  
library(broom) # For processing fit models  
library(purrr) # For data manipulation  
library(kableExtra) # For table styling

## 2.3 Data format explanation

For this tutorial, we will be using an example data set from Barro Colorado Island (BCI), Panama ([available here](https://github.com/LisaHuelsmann/latitudinalCNDD/tree/main/data_prep/input/data_tree)) that includes 7,028 observations, of 3,771 individuals of 16 tree species across two censuses from the larger 50 ha BCI 50 ha Forest Dynamics Plot data set ([more information here](https://forestgeo.si.edu/sites/neotropics/barro-colorado-island)). Each stem is individually mapped and its diamter at 1.3m above ground is measured (DBH), which allows us to calculate neighborhood density across different distance thresholds and as a function of neighbor size (*e.g.*, basal area), respectively.

The code below assumes the data is in a format where each row is an observation for an individual from a census. For this data set, the column descriptions are as follows:

* **treeID**: unique identifier for each tree
* **sp**: species code
* **gx**: spatial coordinate on x axis
* **gy**: spatial coordinate on y axis
* **dbh:** diameter at breast height (mm)
* **ba:** basal area (m^2)
* **status:** status at census, A = alive, D = dead
* **date:** date of observation
* **census**: census number
* **mort**: mortality status at census, 1 = dead, 0 = alive
* **mort\_next**: mortality status at next census, 1 = dead, 0 = alive
* **interval**: time interval between censuses in years

Let’s take a quick look at the data set we’ll be working with:

head(dat, n = 5)

# A tibble: 5 × 14  
 treeID sp gx gy dbh ba status date surv surv\_next mort  
 <chr> <chr> <dbl> <dbl> <dbl> <dbl> <chr> <date> <dbl> <dbl> <dbl>  
1 3884 ceibpe 296. 24.8 1500 1.77 A 1981-11-10 1 1 0  
2 3998 cordbi 280. 45.4 281 0.0620 A 1981-12-11 1 1 0  
3 4065 ceibpe 289. 266. 2000 3.14 A 1982-01-08 1 1 0  
4 4070 cordbi 283. 287. 356 0.0995 A 1982-01-08 1 1 0  
5 4119 ceibpe 258. 224. 1600 2.01 A 1981-12-13 1 1 0  
# ℹ 3 more variables: mort\_next <dbl>, interval <dbl>, census <dbl>

We can produce a plot [Figure 2.1](#fig-map) of the tree locations, where the size of the point is scaled to basal area and colored by species, with each census displayed as a panel:

ggplot(dat, aes(x = gx, y = gy, size = ba, col = sp)) +   
 geom\_point() +   
 facet\_wrap(~census) +  
 theme\_bw(10) +   
 theme(legend.position = "right") +   
 labs(size = "Basal area", col = "Species")

|  |
| --- |
| Figure 2.1: Map of tree locations by census |

## 2.4 Define exponential decay function

We will demonstrate how to calculate neighborhood densities using an exponential decay function. In principle, it’s possible to use any number of different decay functions and to explore various ranges of values that determine the rate of decay. Here, we present the general framework for how one might explore different scenarios, though the ultimate choice of decay function and parameter values will depend on the study. **Note that this method of calculating neighborhood density is only possible when neighbors’ x-y coordinates are known.**

First, we write a function in R for exponential decay with distance, where mu controls the shape of the curve:

exponential\_decay <- function(mu, distance){  
 return(exp(-(1/mu \* distance)))  
}

Let’s see what the exponential decay function looks like across a range of mu values ([Figure 2.2](#fig-decay)):

# Set range of mu values and distances  
decay\_values <- seq(from = 1, to = 25, by = 2)  
  
# Use sprintf to add leading 0s, will help with sorting later on  
decay\_names = paste("exp", sprintf("%02s", decay\_values), sep = "")   
  
distances <- seq(1, 100, 1)  
  
# Generate a dataframe with each combination of mu and distance  
example\_decay\_values <- expand\_grid(decay\_values, distances) %>%  
 # Rename columns  
 rename(decay\_value = decay\_values,   
 distance = distances)  
  
# Evaluate distance decay function for each combination of   
# mu and distance  
example\_decay\_values$decay <- exponential\_decay(  
 mu = example\_decay\_values$decay\_value,  
 distance = example\_decay\_values$distance)  
  
# Plot results  
ggplot(example\_decay\_values,   
 aes(x = distance, y = decay,   
 color = decay\_value, group = decay\_value)) +   
 ylab("Density") +   
 xlab("Distance (m)") +   
 scale\_color\_continuous(name = "Value of \ndecay constant") +   
 geom\_line() +   
 theme\_bw(12)

|  |
| --- |
| Figure 2.2: Plot of decay function |

### 2.4.1 Determine which trees are at edge of plot

Trees near the edge of plot boundaries have incomplete information about their neighborhood, because neighboring trees outside of the plot boundaries are not mapped. Typically trees within certain distance threshold from the plot edge are excluded from analysis, but still included in calculations of neighborhood densities.

We are going to add a column to our data set called ‘edge’ that is TRUE if within a set distance to the edge of the plot (*e.g.*, 30 m in the example below).

For this example data set, the dimensions of the overall plot are 300 x 300 m, ranging from 0-300 on both the x and y axis, and representing a subset of the overall 50 ha forest dynamics plot at BCI.

# Set threshold for distance to edge  
distance\_threshold\_edge = 30  
  
# Add in min and max values for corners of plot  
min\_x <- 0  
max\_x <- 300  
min\_y <- 0  
max\_y <- 300  
  
dat$edge = dat$gx < min\_x + distance\_threshold\_edge |  
 dat$gx > max\_x - distance\_threshold\_edge |  
 dat$gy < min\_y + distance\_threshold\_edge |  
 dat$gy > max\_y - distance\_threshold\_edge  
  
# How many trees fall within the edge threshold?  
table(dat$edge)

FALSE TRUE   
 4526 2502

Below is a plot [Figure 2.3](#fig-edge) of tree locations colored by whether they fall within the edge threshold or not, separated out for each census.

ggplot(dat, aes(x = gx, y = gy, col = edge)) +   
 geom\_point() +   
 facet\_wrap(~census) +  
 coord\_fixed(ratio = 1) +   
 theme\_bw(12)

|  |
| --- |
| Figure 2.3: Map of tree locations colored by whether they fall within the edge threshold |

## 2.5 Calculate distances among focal individual and neighbors

Next, we will calculate distances among individuals in our plot, using an upper threshold distance of what to consider a neighbor.

We use the [‘spatstat.geom’ package](https://spatstat.org/) to efficiently calculate distances among individuals.

Because this example only includes one census interval (two censuses total), we will subset down to just the first census to calculate neighborhood density.

If the data set were to contain multiple census intervals, it would be necessary to calculate neighborhood density separately for each individual in each census interval, using only the individuals that were alive at the beginning of that census interval.

# Set distance threshold for considering neighbors  
distance\_threshold\_neighborhood <- 30  
  
# Subset to first census - this will be different for different   
# datasets  
dat\_first\_census <- dat %>%  
 filter(census == 1)  
  
# Format into 'ppp' object  
dat\_ppp = spatstat.geom::ppp(dat\_first\_census$gx, dat\_first\_census$gy,   
 window = owin(range(dat$gx),   
 range(dat$gy)),   
 checkdup = F)  
  
# Determine close pairs based on distance threshold  
# Returns a list that we convert to tibble later  
neighbors = spatstat.geom::closepairs(dat\_ppp,   
 rmax = distance\_threshold\_neighborhood, # Max radius  
 what = "ijd", # return indicies i, j, and distance   
 twice = TRUE)  
  
# Convert to dataframe  
neighbors <- as\_tibble(neighbors)  
  
# Take a peek at the data  
# i = index of focal individual  
# j = index of neighbor  
# d = distance to neighbor  
head(neighbors)

# A tibble: 6 × 3  
 i j d  
 <int> <int> <dbl>  
1 3227 3252 27.0   
2 3227 3238 17.1   
3 3227 3240 5.75  
4 3227 3229 18.6   
5 3227 3253 24.4   
6 3227 3222 20.8

Next we add in additional columns for individual characteristic, (*e.g.,* species, size) and whether they are located near the edge of the plot (blue area in [Figure 2.3](#fig-edge)).

# add additional columns  
  
# Add species for individual i  
neighbors$sp\_i = dat\_first\_census$sp[neighbors$i]   
  
# Add whether individual i is near edge  
neighbors$edge\_i = dat\_first\_census$edge[neighbors$i]   
  
# Add species for individual j  
neighbors$sp\_j = dat\_first\_census$sp[neighbors$j]   
  
# Add basal area of individual j  
neighbors$ba\_j = dat\_first\_census$ba[neighbors$j]

We then want to add a column that indicates whether the comparison between the focal individual and the neighbor is conspecific or heterospecific because we are interested separately estimating the densities of conspecifics and heterospecifics.

neighbors$comparison\_type <- ifelse(neighbors$sp\_i == neighbors$sp\_j,  
 yes = "con", # conspecific  
 no = "het") # heterospecific

We then remove focal trees that are too close to the edge of the plot.

# remove focal trees i that are at the edge  
neighbors = neighbors[!neighbors$edge\_i, ]

Next, we add columns to our neighbors data set that indicates the distance decay multiplier and the distance decay multiplier weighted by basal area.

# Loop through distance decay values  
for(x in 1:length(decay\_values)){  
   
 # Add in column for distance decay multiplier for each decay value  
 # add \_ba suffix to column name - will eventually be summed based on  
 # number of individual neighbors  
 neighbors[, paste0(decay\_names[x], "\_N")] <- exponential\_decay(mu = decay\_values[x],   
 distance = neighbors$d)  
   
 # Weight distance decay multiplier by basal area of neighbor  
 # add \_ba suffix to column name  
 neighbors[, paste0(decay\_names[x], "\_BA")] <- exponential\_decay(  
 mu = decay\_values[x],   
 distance = neighbors$d) \* neighbors$ba\_j  
}

Depending on how many distance decay values are being investigated, there may be many columns in the data frame.

head(neighbors)

# A tibble: 6 × 34  
 i j d sp\_i edge\_i sp\_j ba\_j comparison\_type exp01\_N exp01\_BA  
 <int> <int> <dbl> <chr> <lgl> <chr> <dbl> <chr> <dbl> <dbl>  
1 2973 2993 14.3 cord… FALSE capp… 1.57e-4 het 6.36e- 7 9.98e-11  
2 2973 3122 20.4 cord… FALSE des2… 4.91e-4 het 1.40e- 9 6.86e-13  
3 2973 2974 5.86 cord… FALSE des2… 7.85e-5 het 2.85e- 3 2.24e- 7  
4 2973 3123 27.2 cord… FALSE des2… 7.85e-5 het 1.47e-12 1.16e-16  
5 2973 3121 14.5 cord… FALSE mico… 1.77e-4 het 5.27e- 7 9.32e-11  
6 2973 2936 13.4 cord… FALSE des2… 7.85e-5 het 1.45e- 6 1.14e-10  
# ℹ 24 more variables: exp03\_N <dbl>, exp03\_BA <dbl>, exp05\_N <dbl>,  
# exp05\_BA <dbl>, exp07\_N <dbl>, exp07\_BA <dbl>, exp09\_N <dbl>,  
# exp09\_BA <dbl>, exp11\_N <dbl>, exp11\_BA <dbl>, exp13\_N <dbl>,  
# exp13\_BA <dbl>, exp15\_N <dbl>, exp15\_BA <dbl>, exp17\_N <dbl>,  
# exp17\_BA <dbl>, exp19\_N <dbl>, exp19\_BA <dbl>, exp21\_N <dbl>,  
# exp21\_BA <dbl>, exp23\_N <dbl>, exp23\_BA <dbl>, exp25\_N <dbl>,  
# exp25\_BA <dbl>

## 2.6 Calculate neighborhood density

Then we summarize neighborhood density for each focal tree separately for conspecifics and heterospecifics.

# Simple calculations of number of neighbors and total basal area,   
# ignoring distance decay  
neighbors\_summary <- neighbors %>%  
 group\_by(i, comparison\_type) %>%  
 summarise(nodecay\_N = n(), # count of neighbors  
 nodecay\_BA = sum(ba\_j)) # sum of basal area)  
  
# Add in decay columns  
neighbors\_summary\_decay <- neighbors %>%  
 group\_by(i, comparison\_type) %>%  
 # Select only columns related to distance decay  
 select(starts\_with("exp")) %>%   
 # Summarize them all by summing columns  
 summarise\_all(sum)   
  
# Join both together  
neighbors\_summary <- left\_join(neighbors\_summary,   
 neighbors\_summary\_decay,  
 by = c("i", "comparison\_type"))  
  
# Add treeID column  
neighbors\_summary$treeID<-dat\_first\_census$treeID[neighbors\_summary$i]  
  
  
   
# If there are any focal individuals with no neighbors, add values   
# of 0 for neighborhood densities   
 noNeighbors = dat\_first\_census$treeID[!dat\_first\_census$treeID   
 %in% neighbors\_summary$treeID &  
 !dat\_first\_census$edge]  
   
 # If there are individuals with no neighbors  
 if (length(noNeighbors) > 0) {  
 neighbors\_summary = bind\_rows(neighbors\_summary,   
 expand\_grid(i = NA,   
 treeID = noNeighbors,   
 comparison\_type = c("het", "cons"))) %>%  
 # Add 0s where NA  
 mutate\_all(replace\_na, replace = 0)   
 }  
  
# Take a peak at the data  
head(neighbors\_summary)

# A tibble: 6 × 31  
# Groups: i [6]  
 i comparison\_type nodecay\_N nodecay\_BA exp01\_N exp01\_BA exp03\_N exp03\_BA  
 <int> <chr> <int> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 5 het 115 0.487 0.0991 4.05e-5 2.03 0.00172   
2 8 het 94 0.0539 0.181 3.63e-5 1.21 0.000703  
3 9 het 151 2.17 0.0464 1.67e-4 1.65 0.00287   
4 10 het 76 0.0509 0.00131 4.72e-7 0.464 0.000175  
5 11 het 59 0.0631 0.00859 1.99e-6 0.493 0.000277  
6 12 het 93 0.0506 0.0441 9.51e-6 1.57 0.000745  
# ℹ 23 more variables: exp05\_N <dbl>, exp05\_BA <dbl>, exp07\_N <dbl>,  
# exp07\_BA <dbl>, exp09\_N <dbl>, exp09\_BA <dbl>, exp11\_N <dbl>,  
# exp11\_BA <dbl>, exp13\_N <dbl>, exp13\_BA <dbl>, exp15\_N <dbl>,  
# exp15\_BA <dbl>, exp17\_N <dbl>, exp17\_BA <dbl>, exp19\_N <dbl>,  
# exp19\_BA <dbl>, exp21\_N <dbl>, exp21\_BA <dbl>, exp23\_N <dbl>,  
# exp23\_BA <dbl>, exp25\_N <dbl>, exp25\_BA <dbl>, treeID <chr>

As described in the main text, it can be advantageous to use total density that includes both conspecific and heterospecific density as a covariate, rather than only heterospecific density.

Here, we calculate overall density by summing heterospecific and conspecific densities.

# First convert to long format which will make it easy to sum across   
# heterospecific and conspecific values  
neighbors\_summary\_long\_format <- neighbors\_summary %>%  
 pivot\_longer(cols = -c("i", "comparison\_type", "treeID"))  
  
# Sum across heterospecific and conspecific values and rename to 'all'  
neighbors\_total\_long\_format <- neighbors\_summary\_long\_format %>%  
 group\_by(i, treeID, name) %>%  
 summarize(value = sum(value)) %>%  
 mutate(comparison\_type = "all")  
  
# Bind together conspecific and 'all' densities  
# remove heterospecific columns  
# fill in 0s where there are no neighbors  
neighbors\_summary\_total = bind\_rows(neighbors\_summary\_long\_format,  
 neighbors\_total\_long\_format) %>%  
 # Can filter out heterospecific neighborhood   
 # values by uncommenting this line of the  
 # objects become too large  
 # filter(comparison\_type != "het") %>%   
 mutate(name = paste0(comparison\_type, "\_", name)) %>%  
 select(-comparison\_type) %>%  
 pivot\_wider(names\_from = name, values\_from = value,   
 values\_fill = 0)

## 2.7 Model mortality as a function of neighborhood density

To determine the ‘best’ decay parameter to use, we fit species-specific mortality models using Generalized Additive Models [GAMs](https://noamross.github.io/gams-in-r-course/) and compare models based on log likelihoods. GAMs are flexible statistical models that combine linear and non-linear components to capture complex relationships in data.

We first create our data set that we will use in the GAMs, subsetting down to just one census interval (because our example dataset only has 2 censuses) and removing trees close to the edge. In other datasets, you may have multiple census intervals, where it would be common practice to include ‘year’ or ‘census’ as a random effect in the model, but otherwise the overall approach is similar.

# Join census data with neighborhood data  
dat\_gam <- left\_join(dat\_first\_census,  
 neighbors\_summary\_total,  
 by = "treeID")  
  
# Remove edge trees  
dat\_gam <- dat\_gam %>%  
 filter(edge == FALSE)

For each species, we summarize data availability to help determine parameters for the GAM smooth terms.

# Summarize data availability at species level to set the degree of   
# smoothness for GAM smooth terms  
sp\_summary <- dat\_gam %>%   
 group\_by(sp) %>%   
 summarise(ndead = sum(mort\_next),  
 nsurv = sum(surv\_next),  
 range\_con\_BA = max(con\_nodecay\_BA) - min(con\_nodecay\_BA),  
 max\_con\_BA = max(con\_nodecay\_BA),  
 unique\_con\_BA = length(unique(con\_nodecay\_BA)),  
 unique\_all\_BA = length(unique(all\_nodecay\_BA)),  
 range\_con\_N = max(con\_nodecay\_N) - min(con\_nodecay\_N),  
 max\_con\_N = max(con\_nodecay\_N),  
 unique\_con\_N = length(unique(con\_nodecay\_N)),  
 unique\_all\_N = length(unique(all\_nodecay\_N)),  
 unique\_dbh = length(unique(dbh))  
 )  
  
# Filter out species that have 0 or 100% mortality  
sp\_summary <- sp\_summary %>%  
 filter(ndead > 0) %>%  
 filter(nsurv > 0)

In this long block of code, we loop over all possible combinations of decay values for neighborhood densities weighted by abundance (N) and size (BA) for each species and fit a separate GAM for each model. For each GAM, we assess whether the model was able to be fit, and if it was able to be fit, whether it converged and for potential warnings. We save the results of successful model fits into a list that we will process later.

For large datasets where individual GAMs take a long time to run, the code could be modified to run in parallel, either locally on a personal computer or across a computing cluster.

# Initialize list that will save model outputs  
res\_mod <- list()  
  
  
# Model run settings  
run\_settings <- expand\_grid(species = unique(sp\_summary$sp),  
 decay\_con = c("nodecay", decay\_names),  
 decay\_total = c("nodecay", decay\_names),  
 nhood\_data\_type = c("N", "BA"))  
  
  
# Loop through model run settings  
for(run\_settings\_row in 1:nrow(run\_settings)){  
   
 # Extract values from run settings dataframe  
 species <- run\_settings$species[run\_settings\_row]  
 decay\_con <- run\_settings$decay\_con[run\_settings\_row]  
 decay\_total <- run\_settings$decay\_total[run\_settings\_row]  
 nhood\_data\_type <- run\_settings$nhood\_data\_type[run\_settings\_row]  
  
 # Subset down to just focal species  
 dat\_subset <- dat\_gam %>%  
 filter(sp == species)  
  
 # Set run name  
 run\_name <- paste0(species, "\_total", decay\_total,"\_con",   
 decay\_con, "\_", nhood\_data\_type)  
   
 # Print status if desired  
 # cat("Working on run: ", run\_name, " ...\n")  
  
 # Create model formula  
 # Initial DBH included as predictor variable  
 form = paste0("mort\_next ~ s(dbh, k = k1) + s(all\_", decay\_total,   
 "\_", nhood\_data\_type,   
 ", k = k2) + s(con\_", decay\_con,   
 "\_", nhood\_data\_type,   
 ", k = k3)")  
   
 # Convert into formula  
 form <- as.formula(form)  
   
 # Choose penalties for model fitting  
 # set to default 10 (the same as -1)  
 # The higher the value of k, the more flexible the smooth term becomes, allowing for more intricate and wiggly patterns. Conversely, lower values of k result in smoother and simpler representations.  
 k1 = k2 = k3 = 10  
 if (k1 > sp\_summary$unique\_dbh[sp\_summary$sp == species]) {   
 k1 = sp\_summary$unique\_dbh[sp\_summary$sp == species] - 2  
 }  
 if (k2 > sp\_summary$unique\_all\_N[sp\_summary$sp == species]) {  
 k2 = sp\_summary$unique\_all\_N [sp\_summary$sp == species]- 2  
 }  
 if (k3 > sp\_summary$unique\_con\_N[sp\_summary$sp == species]) {   
 k3 = sp\_summary$unique\_con\_N[sp\_summary$sp == species] - 2   
 }  
   
  
 # Fit model  
 # wrap in a try function to catch any errors  
 mod = try(gam(form,  
 family = binomial(link=cloglog),  
 offset = log(interval),  
 data = dat\_subset,  
 method = "REML"),   
 silent = T)  
  
   
 # Check if model was able to fit  
 if (!any(class(mod) == "gam")) {  
 # print(paste("gam failed for:", run\_name))  
 } else {  
  
 # Check if gam converged  
 if (!mod$converged) {  
 # print(paste("no convergence for:", run\_name))  
 } else {  
   
 # check for complete separation  
 # https://stats.stackexchange.com/questions/336424/issue-with-complete-separation-in-logistic-regression-in-r  
 # Explore warning "glm.fit: fitted probabilities numerically 0   
 # or 1 occurred"  
 eps <- 10 \* .Machine$double.eps  
 glm0.resids <- augment(x = mod) %>%  
 mutate(p = 1 / (1 + exp(-.fitted)),  
 warning = p > 1-eps,  
 influence = order(.hat, decreasing = T))  
 infl\_limit = round(nrow(glm0.resids)/10, 0)  
 # check if none of the warnings is among the 10% most   
 # influential observations, than it is okay..  
 num = any(glm0.resids$warning & glm0.resids$influence < infl\_limit)  
   
 # complete separation  
 if (num) {  
 # print(paste("complete separation is likely for:", run\_name))  
 } else {  
   
 # missing Vc  
 if (is.null(mod$Vc)) {  
 # print(paste("Vc not available for:", run\_name))  
 } else {  
   
 # Add resulting model to list if it passes all checks  
 res\_mod[[run\_name]] <- mod  
   
 } # Vc ifelse  
 } # complete separation ifelse  
 } # convergence ifelse  
 } # model available ifelse  
} # end run settings loop

## 2.8 Summarize model fits

Next, we will extract regression coefficients into a dataframe using broom::tidy()

# Extract coefficients for each model into a list  
coefs = lapply(res\_mod, broom::tidy)   
  
# Add a column for model run to each object in the list  
coefs = Map(cbind, coefs, run\_name = names(coefs))   
coefs = do.call(rbind, coefs) # Bind elements of list together by rows  
rownames(coefs) <- NULL # Remove row names  
coefs <- coefs %>%  
 select(run\_name, everything()) # Rearrange columns  
  
# Take a look at the data  
knitr::kable(head(coefs), digits = 2, booktabs = T) %>%  
 kable\_styling(latex\_options = c("striped", "scale\_down"))

| run\_name | term | edf | ref.df | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| cappfr\_totalnodecay\_connodecay\_N | s(dbh) | 1.00 | 1.00 | 2.46 | 0.12 |
| cappfr\_totalnodecay\_connodecay\_N | s(all\_nodecay\_N) | 1.23 | 1.43 | 1.25 | 0.31 |
| cappfr\_totalnodecay\_connodecay\_N | s(con\_nodecay\_N) | 1.00 | 1.00 | 1.84 | 0.17 |
| cappfr\_totalnodecay\_connodecay\_BA | s(dbh) | 4.36 | 5.31 | 9.83 | 0.09 |
| cappfr\_totalnodecay\_connodecay\_BA | s(all\_nodecay\_BA) | 1.00 | 1.00 | 0.36 | 0.55 |
| cappfr\_totalnodecay\_connodecay\_BA | s(con\_nodecay\_BA) | 1.00 | 1.00 | 3.04 | 0.08 |

Next we will extract model summaries for each model with broom::glance() that provides key information like degrees of freedom, log likelihood, AIC, etc.

# Extract summaries for each model into a list  
sums = lapply(res\_mod, broom::glance)   
  
# Add a column for model run to each object in the list  
sums = Map(cbind, sums, run\_name = names(sums))   
sums = do.call(rbind, sums) # Bind elements of list together by rows  
rownames(sums) <- NULL # Remove row names  
  
# Separate run name into columns for species, decay, and density type  
sums <- sums %>%  
 separate(run\_name, into = c("sp", "decay\_total",   
 "decay\_con", "density\_type"),   
 remove = FALSE)  
  
# Remove 'total' and 'con' from decay columns  
sums$decay\_total <- gsub("total", "", sums$decay\_total)  
sums$decay\_con <- gsub("con", "", sums$decay\_con)  
  
# Rearrange columns   
sums <- sums %>%  
 select(run\_name, sp, decay\_total, decay\_con, density\_type,   
 everything())   
  
# Take a look at the model summaries  
knitr::kable(head(sums), digits = 2) %>%  
 kable\_styling(latex\_options = c("striped", "scale\_down"))

| run\_name | sp | decay\_total | decay\_con | density\_type | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| cappfr\_totalnodecay\_connodecay\_N | cappfr | nodecay | nodecay | N | 4.23 | -22.62 | 54.10 | 70.37 | 45.23 | 285.77 | 290 |
| cappfr\_totalnodecay\_connodecay\_BA | cappfr | nodecay | nodecay | BA | 7.36 | -17.21 | 51.03 | 81.52 | 34.41 | 282.64 | 290 |
| cappfr\_totalexp01\_connodecay\_N | cappfr | exp01 | nodecay | N | 5.06 | -22.64 | 56.52 | 77.14 | 45.29 | 284.94 | 290 |
| cappfr\_totalexp01\_connodecay\_BA | cappfr | exp01 | nodecay | BA | 8.08 | -17.54 | 53.90 | 88.44 | 35.07 | 281.92 | 290 |
| cappfr\_totalexp03\_connodecay\_N | cappfr | exp03 | nodecay | N | 8.07 | -17.87 | 54.98 | 90.30 | 35.73 | 281.93 | 290 |
| cappfr\_totalexp03\_connodecay\_BA | cappfr | exp03 | nodecay | BA | 7.93 | -17.62 | 53.85 | 87.98 | 35.24 | 282.07 | 290 |

Due to limited sample sizes, it is likely that GAMs will not fit for each species. For example, in the table below of summary data of species where models did not successfully fit/converge, there are several instances of where all individuals of the species survived during the census, leading to no mortality events to use in the model.  
  
We need to exclude species without complete model runs from our overall calculations when looking for optimal decay parameters across the entire data set.

# Tally up number of model runs by species and total decay values  
table(sums$sp, sums$decay\_total)

exp01 exp03 exp05 exp07 exp09 exp11 exp13 exp15 exp17 exp19 exp21  
 cappfr 28 28 28 28 28 28 28 28 28 28 28  
 cordbi 27 28 28 28 28 28 28 28 28 27 27  
 des2pa 28 28 28 28 28 28 28 28 28 28 28  
 micone 28 28 28 28 28 28 28 28 28 28 28  
 pentma 28 28 28 28 28 23 27 28 28 28 28  
 stylst 28 28 28 28 28 28 28 28 28 28 28  
   
 exp23 exp25 nodecay  
 cappfr 28 28 28  
 cordbi 28 28 28  
 des2pa 28 28 28  
 micone 28 28 28  
 pentma 28 28 28  
 stylst 28 28 28

# get incomplete run-site-species combinations  
run\_counts\_by\_sp <- sums %>%   
 group\_by(sp) %>%  
 tally() %>%  
 # Join with overall species list  
 left\_join(sp\_summary %>% select(sp), ., by = "sp")   
  
# Get expected number of runs if all models work  
expected\_total\_runs <- run\_settings %>%  
 group\_by(species) %>%  
 tally() %>%  
 pull(n) %>%  
 max()  
  
# Save species names where they didn't have all expected combinations   
# of model runs  
incomplete = run\_counts\_by\_sp$sp[run\_counts\_by\_sp$n <   
 expected\_total\_runs |   
 is.na(run\_counts\_by\_sp$n)]  
  
# Species with successful runs  
knitr::kable(sp\_summary[!sp\_summary$sp %in% incomplete, ],   
 digits = 2) %>%  
 kable\_styling(latex\_options = c("striped", "scale\_down"))

| sp | ndead | nsurv | range\_con\_BA | max\_con\_BA | unique\_con\_BA | unique\_all\_BA | range\_con\_N | max\_con\_N | unique\_con\_N | unique\_all\_N | unique\_dbh |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| cappfr | 5 | 285 | 0.02 | 0.02 | 273 | 290 | 30 | 32 | 31 | 111 | 31 |
| des2pa | 174 | 1172 | 0.14 | 0.15 | 1335 | 1343 | 123 | 140 | 123 | 157 | 77 |
| micone | 38 | 28 | 0.00 | 0.00 | 48 | 66 | 12 | 12 | 13 | 55 | 11 |
| stylst | 3 | 101 | 0.02 | 0.02 | 91 | 104 | 13 | 13 | 14 | 68 | 30 |

# Species without successful runs  
knitr::kable(sp\_summary[sp\_summary$sp %in% incomplete, ],   
 digits = 2) %>%  
 kable\_styling(latex\_options = c("striped", "scale\_down"))

| sp | ndead | nsurv | range\_con\_BA | max\_con\_BA | unique\_con\_BA | unique\_all\_BA | range\_con\_N | max\_con\_N | unique\_con\_N | unique\_all\_N | unique\_dbh |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| acalma | 2 | 5 | 0.00 | 0.00 | 4 | 7 | 2 | 2 | 2 | 7 | 6 |
| cordbi | 7 | 36 | 0.25 | 0.25 | 36 | 43 | 10 | 10 | 11 | 38 | 28 |
| pentma | 12 | 21 | 0.00 | 0.00 | 20 | 33 | 6 | 6 | 6 | 31 | 11 |
| sponra | 3 | 15 | 0.08 | 0.08 | 12 | 18 | 5 | 5 | 4 | 17 | 16 |

## 2.9 Selecting optimum decay parameter values across all species

We then summarize different model criteria across all species runs. To look for the optimal value for decay parameters, we sum log likelihoods across all species for a given decay parameter combination and choose the resulting parameter combination with the highest summed log likelihood.

sums\_total <- sums %>%   
 filter(!sp %in% incomplete) %>%  
 group\_by(decay\_total, decay\_con, density\_type) %>%  
 summarise(nvalues = n(),  
 sumlogLik = sum(logLik),  
 meanlogLik = mean(logLik)) %>%  
 arrange(decay\_total, decay\_con, density\_type)  
  
sums\_total

# A tibble: 392 × 6  
# Groups: decay\_total, decay\_con [196]  
 decay\_total decay\_con density\_type nvalues sumlogLik meanlogLik  
 <chr> <chr> <chr> <int> <dbl> <dbl>  
 1 exp01 exp01 BA 4 -556. -139.  
 2 exp01 exp01 N 4 -554. -138.  
 3 exp01 exp03 BA 4 -553. -138.  
 4 exp01 exp03 N 4 -558. -140.  
 5 exp01 exp05 BA 4 -554. -138.  
 6 exp01 exp05 N 4 -561. -140.  
 7 exp01 exp07 BA 4 -557. -139.  
 8 exp01 exp07 N 4 -563. -141.  
 9 exp01 exp09 BA 4 -558. -140.  
10 exp01 exp09 N 4 -564. -141.  
# ℹ 382 more rows

We create a heatmap plot [Figure 2.4](#fig-optim) of summed log likelihoods for all parameter combinations across all species, with the optimal parameter combination marked with an X

# Find optimum value separately for N and BA  
 optimum <- sums\_total %>%  
 group\_by(density\_type) %>%  
 slice\_max(sumlogLik)  
   
# Plot heatmap of log likelihood values  
 ggplot(sums\_total, aes(x = decay\_total, y = decay\_con,   
 fill = sumlogLik)) +  
 geom\_tile(width = 0.9, height = 0.9, col = "black") +   
 scale\_fill\_viridis\_c() + # viridis color palette  
 geom\_label(data = optimum, label = "X") +   
 labs(x = "Decay total density", y = "Decay conspecific density",   
 fill = "sumlogLik") +   
 facet\_wrap(~density\_type, ncol = 1) +   
 theme\_bw(12) +   
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5,   
 hjust=1))

|  |
| --- |
| Figure 2.4: Heatmap of optimal values for decay constants |

For this data set, the following are the optimal decay parameter values across all species separately for neighborhood density calculated by abundance (N) and by basal area (BA)

knitr::kable(optimum, digits = 2) %>%  
 kable\_styling(latex\_options = c("striped", "scale\_down"))

| decay\_total | decay\_con | density\_type | nvalues | sumlogLik | meanlogLik |
| --- | --- | --- | --- | --- | --- |
| exp19 | exp25 | BA | 4 | -541.69 | -135.42 |
| exp03 | exp01 | N | 4 | -541.59 | -135.40 |

## 2.10 Selecting optimum decay parameter values separately for each species

Alternatively, it may be useful to determine optimum decay parameters on a species by species basis. To do this, we find the maximum log liklihood for each species separately across decay parameter combination and choose the resulting parameter combination with the highest log likelihood.

sums\_total\_by\_spp <- sums %>%   
 filter(!sp %in% incomplete) %>%  
 group\_by(decay\_total, decay\_con, density\_type, sp) %>%  
 summarise(nvalues = n(),  
 sumlogLik = sum(logLik),  
 meanlogLik = mean(logLik)) %>%  
 arrange(decay\_total, decay\_con, density\_type)  
  
sums\_total\_by\_spp

# A tibble: 1,568 × 7  
# Groups: decay\_total, decay\_con, density\_type [392]  
 decay\_total decay\_con density\_type sp nvalues sumlogLik meanlogLik  
 <chr> <chr> <chr> <chr> <int> <dbl> <dbl>  
 1 exp01 exp01 BA cappfr 1 -16.6 -16.6   
 2 exp01 exp01 BA des2pa 1 -489. -489.   
 3 exp01 exp01 BA micone 1 -43.7 -43.7   
 4 exp01 exp01 BA stylst 1 -6.84 -6.84  
 5 exp01 exp01 N cappfr 1 -14.3 -14.3   
 6 exp01 exp01 N des2pa 1 -489. -489.   
 7 exp01 exp01 N micone 1 -41.3 -41.3   
 8 exp01 exp01 N stylst 1 -8.64 -8.64  
 9 exp01 exp03 BA cappfr 1 -20.7 -20.7   
10 exp01 exp03 BA des2pa 1 -485. -485.   
# ℹ 1,558 more rows

We create a heatmap plot [Figure 2.5](#fig-optim-spp) of log likelihoods for all parameter combinations for each species separately, with the optimal parameter combination marked with an X. We only display the first three species here, because with data sets containing many species, it will be hard to visualize all the species on one graph and you may need to subdivide the plot further for visualization.

sums <- sums %>%  
 filter(sp %in% c("cappfr", "cordbi", "des2pa")) %>%  
 group\_by(sp) %>%  
 # Scale loglikihood by species to help with visualization  
 mutate(logLik\_scaled = scale(logLik)) %>%   
 ungroup()  
   
# Find optimum value separately for N and BA  
 optimum\_by\_sp <- sums %>%  
 group\_by(sp, density\_type) %>%  
 slice\_max(logLik\_scaled, with\_ties = FALSE)  
  
# Plot heatmap of log likelihood values  
 ggplot(sums, aes(x = decay\_total, y = decay\_con,  
 fill = logLik\_scaled)) +  
 geom\_tile(width = 0.9, height = 0.9, col = "black") +  
 geom\_label(data = optimum\_by\_sp, label = "X") +  
 labs(x = "Decay total density", y = "Decay conspecific density",   
 fill = "logLik") +  
 facet\_wrap(~sp + density\_type, ncol = 2, scales = "free") +  
 scale\_fill\_viridis\_c() + # viridis color palette  
 theme\_bw(12) +   
 theme(legend.position = "right") +   
 labs(fill = "Log likelihood\n(scaled)") +   
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5,   
 hjust=1))

|  |
| --- |
| Figure 2.5: Heatmap of optimal values for decay constants separately for each species |

For this data set, the following are the optimal decay parameter values for each species separately for neighborhood density calculated by abundance (N) and by basal area (BA):

knitr::kable(optimum\_by\_sp, digits = 2) %>%  
 kable\_styling(latex\_options = c("striped", "scale\_down"))

| run\_name | sp | decay\_total | decay\_con | density\_type | df | logLik | AIC | BIC | deviance | df.residual | nobs | logLik\_scaled |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| cappfr\_totalexp19\_conexp25\_BA | cappfr | exp19 | exp25 | BA | 11.09 | -8.74 | 43.29 | 90.63 | 17.48 | 278.91 | 290 | 2.9564926 |
| cappfr\_totalexp03\_conexp01\_N | cappfr | exp03 | exp01 | N | 11.03 | -11.80 | 49.64 | 97.43 | 23.59 | 278.97 | 290 | 2.1291275 |
| cordbi\_totalexp05\_conexp03\_BA | cordbi | exp05 | exp03 | BA | 7.61 | -12.46 | 43.10 | 58.90 | 24.93 | 34.39 | 42 | 0.2471958 |
| cordbi\_totalexp01\_conexp07\_N | cordbi | exp01 | exp07 | N | 14.84 | 0.00 | 31.83 | 59.48 | 0.00 | 27.16 | 42 | 4.5489731 |
| des2pa\_totalexp25\_conexp09\_BA | des2pa | exp25 | exp09 | BA | 8.36 | -481.80 | 984.07 | 1037.31 | 963.61 | 1337.64 | 1346 | 1.7645474 |
| des2pa\_totalexp03\_connodecay\_N | des2pa | exp03 | nodecay | N | 7.33 | -481.92 | 981.23 | 1026.49 | 963.84 | 1338.67 | 1346 | 1.6992914 |

# 3. Calculating Marginal Effects

## 3.1 Overview

In this section, we examine a *subset* of the Barro Colorado Island (BCI) seedling data to illustrate the impact of conspecific density on mortality probability. We calculate the Average Marginal Effect ([more about marginal effects in general here](https://marginaleffects.com/)) as our metric of the strength conspecific density dependence. We then estimate both the absolute Average Marginal Effect (AME) and the relative Average Marginal Effect (rAME). The **AME** represents the average absolute change in mortality probability due to a specified increase in conspecific density. In contrast, the **rAME** is the relative change in mortality probability compared to a baseline value.

The standard approach for modeling plant CDD patterns presumes a linear relationship between performance (e.g., mortality/survival) and conspecific neighborhood density metrics. However, in this section, we use [‘Generalized Additive Models (GAMs)’](https://multithreaded.stitchfix.com/blog/2015/07/30/gam/) as it offer flexibility for non-linear relationships between performance and predictors.

|  |
| --- |
| Note |
| Note: The code is adapted from the [latitudinalCNDD repository](https://github.com/LisaHuelsmann/latitudinalCNDD/tree/main/code) by [Lisa Hüelsmann](https://demographicecology.com/). |

## 3.2 Load libraries

# Load libraries  
library(boot)  
library(broom)  
library(dplyr)  
library(ggplot2)  
library(gratia)  
library(here)  
library(kableExtra)   
library(knitr)   
library(lubridate)  
library(mgcViz)  
library(mgcv)  
library(MASS)  
library(parallel)  
library(pbapply)  
library(readr)  
library(skimr)  
library(spatstat.geom)  
library(tidyr)

## 3.3 Load data

Our demonstration utilizes a subset of the Barro Colorado Island (BCI) seedling data, encompassing only 30 species. Please note that this analysis is *solely for demonstration purposes*; hence, no biological conclusions should be drawn from the results due to the limited data subset used.

The code below assumes the data is in a format where each row is an observation for an individual from a census. For this particular data set, the column descriptions are as follows:

* **Id**: unique identifier for each seedling
* **plot**: location
* **spp**: species
* **date:** date of observation
* **year:** year of observation
* **census**: census number
* **status:** status at census, 0 = alive, 1 = dead
* **height.last.census**: height.last.census
* **con\_dens:** number of conspecific neighbors
* **total\_dens:** number of total neighbors
* **het\_dens:** number of heterospecifics neighbors
* **time.since.last.census**: time interval between censuses

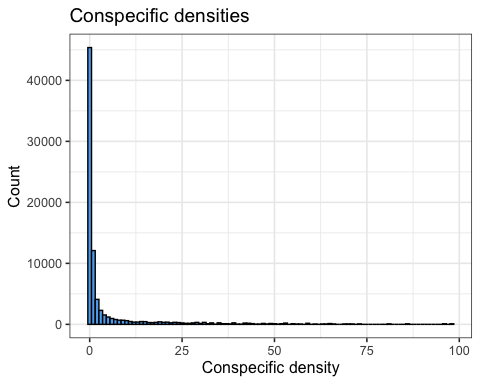
data\_BCI <- read\_csv("./data/BCI\_seedling\_data\_30\_spp\_2023.csv")  
colnames(data\_BCI)

[1] "id" "q20"   
 [3] "plot" "tag"   
 [5] "spp" "date"   
 [7] "year" "census"   
 [9] "status" "height.last.census"   
[11] "height.last.census.log.scaled" "con\_dens"   
[13] "total\_dens" "het\_dens"   
[15] "time.since.last.census"

# make sure variables are taken as factor  
data\_BCI$census <- factor(data\_BCI$census)  
data\_BCI$spp <- factor(data\_BCI$spp)  
data\_BCI$plot <- factor(data\_BCI$plot)  
data\_BCI$height=data\_BCI$height.last.census  
data\_BCI$interval=as.numeric(data\_BCI$time.since.last.census)

Let’s take a quick look at the data set we’ll be working with:

# Exploring data   
 # visualize   
 ggplot(data\_BCI, aes(x = con\_dens)) +  
 geom\_histogram(binwidth = 1, color = "black", fill = "steelblue2") +  
 labs(x = "Conspecific density", y = "Count",   
 title = "Conspecific densities") +   
 theme\_bw(12)



## 3.4 Handling “data deficient” species

We categorize a species as *data deficient* if it has fewer than four unique conspecific density values. At the end of this section of the code, a data frame named **‘nsp’** will be generated. This data frame will classify each species as either “data deficient” or “not data deficient”.

From our data set 9 species out of 30 were assigned as *data deficient*. Here, we used the arbitrary thresholds of 4 unique values of conspecific densities and a minimum range of 1 for conspecific as our thresholds for what constitutes a data deficient species. These exact values will depend on the data set and the questions of the study, so we suggest thinking carefully about what the appropriate values are for the purposes of a given study.

nval = 4 ## this is the number of 'unique' conspecific values   
minrange = 1 # minimum range for conspecific density  
  
data\_BCI %>%   
 group\_by(spp) %>%   
 summarise(  
 range\_con\_dens = max(con\_dens) - min(con\_dens),  
 max\_con\_dens = max(con\_dens),  
 unique\_con\_dens = length(unique(con\_dens)),  
 unique\_total\_dens = length(unique(total\_dens)),  
 unique\_height = length(unique(height.last.census))  
 ) %>%   
   
 # check if conspecific less than defined nval  
 mutate(issue\_nval = unique\_con\_dens < nval,  
 # range should at least be equal to minrange  
 issue\_range = range\_con\_dens < minrange,   
 trymodel = !(issue\_nval|issue\_range),  
 # Assignment of "data deficient" species  
 data\_deficient = !trymodel   
 ) -> nsp # Store the resulting dataframe in the object 'nsp'  
  
# Visualize the top rows of the table 'nsp' in a formatted manner  
nsp %>%  
 head() %>%  
 kable() %>%  
 kable\_styling(latex\_options = c("striped", "scale\_down"))

| spp | range\_con\_dens | max\_con\_dens | unique\_con\_dens | unique\_total\_dens | unique\_height | issue\_nval | issue\_range | trymodel | data\_deficient |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ACALDI | 7 | 7 | 8 | 35 | 379 | FALSE | FALSE | TRUE | FALSE |
| AEGICE | 4 | 4 | 5 | 28 | 499 | FALSE | FALSE | TRUE | FALSE |
| BEILPE | 98 | 98 | 69 | 72 | 894 | FALSE | FALSE | TRUE | FALSE |
| CAPPFR | 5 | 5 | 6 | 52 | 767 | FALSE | FALSE | TRUE | FALSE |
| CECRIN | 7 | 7 | 7 | 20 | 177 | FALSE | FALSE | TRUE | FALSE |
| CORDLA | 4 | 4 | 5 | 32 | 457 | FALSE | FALSE | TRUE | FALSE |

## 3.5 Function for fitting models

In the context of the Janzen-Connell Hypothesis, we focus on the differences between CDD and HDD (‘Stabilizing effect’), where conspecific neighbors’ negative effects surpass those from heterospecifics, leading to population stabilization (Broekman et al. 2019). This effect is vital for estimating species’ self-limitation.

In this section we illustrate the conspecific density impact, considering total tree density. The estimated effect (slope) in our analysis corresponds to the result of subtracting HDD from CDD.

We use here a Generalized Additive Model (GAM) with a complementary log-log (cloglog) link function to model the seedling status (‘alive’ or ‘dead’) as a function of conspecific density **con\_dens**, total density **total\_dens** and tree height or size, the latter serving as a potential confounder or precision covariate.

The cloglog link accounts for differences in observation time Δ𝑡 through an offset term, applicable to datasets where (0=alive and 1=alive)(Currie 2016).

After defining covariates for the GAM models, we specify the smooth term in the model formula. If a k value exceeds the number of unique values in the variable, it is adjusted to be two less than that number. Model convergence and warnings are also checked. which this set up here we define k=2 as minimum as K=1 is a linear relation.

After defining the covariates for the GAMs, we specify the smooth term in the model formula. The k value determines the complexity of the smooth. If k exceeds the number of unique values in a variable, it’s adjusted to be two less than that number. We also monitor model convergence and any warnings that may arise.

# Define a function to fit a model to the given data  
model\_fit = function(data, spinfo, reduced = F) {  
   
 # create new factor with correct factor levels per species   
 data$census = factor(data$census)  
   
   
 # # Determine if there's more than one unique value in 'census'   
 # and construct the relevant term for the model formula  
   
 term\_c = ifelse(length(unique(data$census)) > 1,   
 "+ s(census, bs = 're')", "")   
 #term\_p = "+ s(plot, bs = 're')"  
   
 if (reduced) {  
 form = as.formula(paste0("status ~ s(height, k = k1) +   
 s(total\_dens, k = k2)"  
 , term\_c)) # reduced model #,term\_p  
 } else {  
 form = as.formula(paste0("status ~ s(height, k = k1) +   
 s(total\_dens, k = k2) +   
 s(con\_dens, k = k3)"   
 , term\_c)) # full model #, term\_p  
 }  
   
 # Choose penalty  
 # set to default k=10   
 k1 = k2 = k3 = 10  
 if (k1 > spinfo$unique\_height) k1 = spinfo$unique\_height - 2  
 if (k2 > spinfo$unique\_total\_dens) k2 = spinfo$unique\_total\_dens - 2  
 if (k3 > spinfo$unique\_con\_dens) k3 = spinfo$unique\_con\_dens - 2  
   
 # k = 1 would force linearity for that term, and we aim to consider   
 # also potential non-linear relationships, so conspecific k is k=2   
 # minimum.  
   
 # Fit the Generalized Additive Model (GAM)  
 mod = try(gam(form  
 , family = binomial(link=cloglog)  
 , offset = log(interval)  
 , data = data  
 , method = "REML"  
 ) , silent = T  
 )  
 # Return the fitted model  
 return(mod)  
   
}  
  
# check model run  
  
# Define a function to check the convergence of the model  
model\_convergence = function(model) {  
   
 # gam not available  
 if (!any(class(model)=="gam")) {  
 print(paste(spp, "gam failed"))  
 } else {  
   
 # gam not converged  
 if (!model$converged) {  
 print(paste(spp, "no convergence"))  
 } else {  
   
   
# Explore warning "glm.fit: fitted probabilities numerically 0 or 1   
 # occurred (complete separation)"  
 eps <- 10 \* .Machine$double.eps  
 glm0.resids <- augment(x = model) %>%  
 mutate(p = 1 / (1 + exp(-.fitted)),  
 warning = p > 1-eps,  
 influence = order(.hat, decreasing = T))  
 infl\_limit = round(nrow(glm0.resids)/10, 0)  
   
 # Check if any of the warnings correspond to the 10% most   
 # influential observations. If not, then it is okay.  
  
 num = any(glm0.resids$warning & glm0.resids$influence <   
 infl\_limit)  
   
 # If there's complete separation  
 if (num) {  
 print(paste(spp, "complete separation is likely"))  
 } else {  
   
 # Check if the Vc component of the model is missing  
 if (is.null(model$Vc)) {  
 print(paste(spp, "Vc not available"))  
 } else {  
   
 # If everything is fine, return the model  
 return(model)  
 }  
 }  
 }  
 }  
}

## 3.6 Fit models

Here we fit the model for all species. *data deficient* species are treated as one group since there is not sufficient data to be treated independently. In this dataset, we have 9 species that are flagged as ‘data deficient’,

# Check the distribution of species marked as 'data deficient'   
# (T or F)   
# and determine the number of species for which we will try the model   
  
# We have 9 species that are flagged as 'data deficient'  
table(data\_deficient = nsp$data\_deficient, trymodel = nsp$trymodel)

trymodel  
data\_deficient FALSE TRUE  
 FALSE 0 21  
 TRUE 9 0

# Convert spp to character  
data\_BCI$spp <- as.character(data\_BCI$spp)  
  
# Extract species that are flagged as 'data deficient' from the nsp   
# dataframe  
data\_deficient\_species <- nsp$spp[nsp$data\_deficient == TRUE]  
  
# Replace species names with "data\_deficient\_seedling" for those   
# flagged as 'data\_deficient'  
data\_BCI2 <- data\_BCI %>%   
 mutate(spp = ifelse(spp %in% data\_deficient\_species,   
 "data\_deficient\_seedling", spp))  
  
# # Summarize attributes for each species in the modified dataframe   
# (including the new #"data\_deficient\_seedling" group)  
  
data\_BCI2 %>%  
 group\_by(spp) %>%  
 summarise(  
 range\_con\_dens = max(con\_dens, na.rm = TRUE) - min(con\_dens,   
 na.rm = TRUE),  
 max\_con\_dens = max(con\_dens, na.rm = TRUE),  
 unique\_con\_dens = length(unique(con\_dens)),  
 unique\_total\_dens = length(unique(total\_dens)),  
 unique\_height = length(unique(height.last.census))  
 ) %>%  
 mutate(  
 # less than nval unique values in consp densities  
 issue\_nval = unique\_con\_dens < nval,   
 # range should at least be equal to minrange  
 issue\_range = range\_con\_dens < minrange,   
 trymodel = !(issue\_nval | issue\_range),  
 # preliminary assignment of data\_deficient species  
 data\_deficient = !trymodel   
 ) -> nsp\_data\_deficient  
  
  
####  
## Fit model for each species  
###  
  
# Create lists to store results of the main and reduced model fits  
  
# List for main model fits  
res\_mod = list()   
# List for reduced model fits (for calculating Pseudo R2)  
res\_red\_mod = list()   
  
  
# Loop through species in the nsp\_data\_deficient dataframe for which   
# we will try modeling   
# (Here, the group of data\_deficient species is treated as a single   
# species)  
   
 for(spp in nsp\_data\_deficient$spp[nsp\_data\_deficient$trymodel]) {  
   
 # select data for individual species  
 dat\_sp = data\_BCI2[data\_BCI2$spp == spp, ]  
   
 # Fit the main and reduced models for the current species  
 mod = model\_fit(data = dat\_sp,   
 spinfo = nsp\_data\_deficient[nsp\_data\_deficient$spp   
 == spp, ])  
 mod\_red = model\_fit(data = dat\_sp,   
 spinfo = nsp\_data\_deficient[nsp\_data\_deficient$spp   
 == spp, ], reduced = T)  
   
 # Check the convergence of both models  
 res = model\_convergence(model = mod)  
 res\_red = model\_convergence(model = mod\_red)  
   
 # save result  
 if (is.character(res)) {  
 nsp$data\_deficient[nsp$spp == spp] = T   
 } else {  
 res\_mod[[spp]] = res  
 res\_red\_mod[[spp]] = res\_red  
 }  
}

## 3.7 Summarize model fits

Regression table via broom::tidy()

coefs = lapply(res\_mod, broom::tidy)  
coefs = Map(cbind, coefs, sp = names(coefs))  
coefs = do.call(rbind, coefs)

Model summary via broom::glance()

# For each model result in 'res\_mod', extract summary statistics   
# (like log-likelihood, AIC, BIC #, etc.) using the 'glance' function  
# from the 'broom' package  
  
# df logLik AIC BIC deviance df.residuals nobs   
sums = lapply(res\_mod, broom::glance)  
sums = Map(cbind, sums, sp = names(sums))  
sums = do.call(rbind, sums)  
head(sums)

df logLik AIC BIC deviance df.residual nobs  
ACALDI 7.738186 -428.3495 876.0873 924.9256 856.6990 1131.2618 1139  
AEGICE 8.763874 -524.7678 1070.5298 1123.3674 1049.5355 1125.2361 1134  
BEILPE 17.776666 -11142.3710 22323.8580 22478.1537 22284.7420 19697.2233 19715  
CAPPFR 10.232458 -2024.4980 4075.8754 4174.3285 4048.9960 11210.7675 11221  
CECRIN 6.221103 -143.4318 301.0402 326.2521 286.8635 252.7789 259  
CORDLA 6.840230 -531.0591 1079.4468 1125.3893 1062.1182 1477.1598 1484  
 sp  
ACALDI ACALDI  
AEGICE AEGICE  
BEILPE BEILPE  
CAPPFR CAPPFR  
CECRIN CECRIN  
CORDLA CORDLA

# AUC  
aucs = lapply(res\_mod, function(x) {  
 roc <- performance::performance\_roc(x, new\_data = x$model)  
 bayestestR::area\_under\_curve(roc$Spec, roc$Sens)  
})  
sums$AUC = unlist(aucs)  
  
  
# Pseudo R2  
sums$pseudoR2 = 1 - (unlist(lapply(res\_mod, function(x) x$deviance)) /  
 unlist(lapply(res\_red\_mod, function(x)   
 x$deviance)))

## 3.8 Plotting results

# plot splines in pdf ----------------------------------------------  
  
# # Specify the name of the PDF file where the plots will be saved  
pdf\_file <- "mortality.pdf"  
  
# Open the PDF file for writing  
pdf(pdf\_file)  
  
# Loop through all the model results stored in 'res\_mod'  
for (i in 1:length(res\_mod)) {  
   
 # Get the vizmod for the current species  
 vizmod <- getViz(res\_mod[[i]], post = T, unconditional = T)  
 pl <- plot(vizmod, nsim = 20, allTerms = T) +   
 # Add confidence interval line/ fit line/ simulation line  
 l\_ciLine() + l\_fitLine() + l\_simLine() +  
 #Add confidence interval bar/# Add fitted points  
 l\_ciBar() + l\_fitPoints(size = 1) +   
 l\_rug() + # Add rug plot  
 # Add title with the name of the current species  
 labs(title = names(res\_mod)[i])   
   
 # Print the plot to the R console only for the first 3 species  
 if (i <= 3) {  
 print(pl, pages = 1)  
 }  
   
 # Save the plot to the PDF  
 print(pl, pages = 1)  
}  
  
# Close the PDF file  
dev.off()

quartz\_off\_screen   
 2

## 3.9 AMEs Absolute and Relative

In this section we illustrate the calculation of the average marginal effect and both the absolute Average Marginal Effect (AME) and the relative Average Marginal Effect (rAME).

Here, **AMEs** are computed by determining the marginal effect (essentially the partial derivative or slope) of a predictor for a unit change change in conspecific density at each observed value, and then averaging these effects. This method yields a single, interpretable measure that offers an averaged estimate of the predictor’s impact. It’s worth noting that the AME, compared to traditional effect sizes, provides a clearer measure of a predictor’s influence by quantifying the average change in the outcome for each unit change in the predictor rather than the raw coefficient (effect size) that may be influenced by the scale of the variable or confounded by interactions with other predictors.

Furthermore, **rAMEs** enhance the level of interpretability by delivering a normalized measure of these averaged effects. They represent the percentage change in the response variable due to a unit change in the predictor relative to the base mortality, providing an intuitive, relative grasp of the predictor’s influence. This normalization process allows for the comparison of effects across different species, each with its own base level of mortality.

For **AMEs** we need the file “res\_model” that includes all the models results.

There is also alternative approaches for calculating average marginal effects using [‘marginaleffects’ package](https://marginaleffects.com/articles/slopes.html) that we encourage the user to explore.

### 3.9.1 Settings for AMEs

Here we provide three scenarios for calculating **AMEs** or **rAMEs** through the **change** argument.

**Equilibrium**: This scenario computes **AMEs** or **rAMEs** for a unit increase in conspecific density, providing insight into the marginal effect of density increase in an existing ecosystem. **Invasion**: This scenario models the effects of introducing species into a new habitat, transitioning the conspecific density from zero to a specified unit, helping understand the impact of sudden species introductions. **IQR**: This scenario evaluates **AMEs** or **rAMEs** within the middle 50% range of conspecific density, offering a perspective less influenced by extremes, hence providing a more robust understanding of effects within typical density ranges.

#### chose predictors for local density -setting AMEs  
   
  
# Define a vector of predictors with their corresponding names  
predictors <- c(con\_dens = "con\_dens",   
 total\_dens = "total\_dens")  
  
# change in conspecific density for AME calculations----  
  
# One more neighbor seedling   
# or for adult trees: pi\*((dbh\_neighbor/1000)/2)^2 \*  
 #dec\_fun(decay\_con,dist\_neighbor, decay\_type)   
additive=1   
   
# different change settings for con-specific densities  
  
interval = 1  
  
# Specify how the conspecific density should be changed for   
# different scenarios:   
# 'equilibrium', 'invasion', and 'iqr' (interquartile range)  
  
change = list(equilibrium = data.frame(con\_dens =   
 "paste('+', additive)")  
 , invasion = data.frame(con\_dens = "c(0, additive)")   
 , iqr = data.frame(con\_dens = "c(q1, q3)")  
 )  
  
## Set the number of iterations for any subsequent calculations or   
# simulations  
iter = 500

### 3.9.2 Functions to calculate AMEs and rAME

This section defines two functions. setstep and get\_AME. Get\_AME is the function calculates the Average Marginal Effects for a given term in a model. The function first creates two copies of the data, **d0** and **d1**, and adjusts the term of interest based on the **change** argument (one of the three scenarios). It then calculates the predictions for the two data sets and computes the marginal effects.

# Define a function to set the step size for numerical derivatives  
# This function determines an appropriate step size using machine's   
# precision/machine epsilon to strike a balance between accuracy and   
# rounding errors.  
  
setstep = function(x) {  
 eps = .Machine$double.eps  
 return(x + (max(abs(x), 1, na.rm = TRUE) \* sqrt(eps)) - x)  
}  
  
  
# Function to compute Average and relative Marginal Effects   
# (AME and rAME) for a given model--------------------  
  
get\_AME = function(mod, data, term  
 , change = NULL  
 , at = NULL  
 , offset = 1  
 , relative = F  
 , iterations = 1000  
 , seed = 10  
 , samples = F) {  
   
 # Prepare two dataframes for different scenarios in marginal   
 # effect computation  
 d0 = d1 = data  
   
 # Adjust the 'term' in the data based on the 'change' parameter  
 if (is.null(change)) {  
   
 # If change is NULL, adjust the term for numerical derivative   
 # computation  
 d0[[term]] = d0[[term]] - setstep(d0[[term]])  
 d1[[term]] = d1[[term]] + setstep(d1[[term]])  
   
 }   
   
 # If change has an additive component, adjust the term accordingly  
 if (grepl("\\+", paste(change, collapse = "\_"))) {  
   
 d1[[term]] = d1[[term]] + as.numeric(gsub("\\+", "", change))  
   
 }   
   
 # If change is explicit with two values, set the term values   
 # directly  
 if (length(change) == 2) {  
   
 d0[[term]] = as.numeric(change[1])  
 d1[[term]] = as.numeric(change[2])  
   
 }  
   
 # If 'at' is specified, set predictor values in the data to these   
 # fixed values  
 # (allows the function to calculate the marginal effects at the   
 # specified values)  
 if (!is.null(at)) {  
 for (i in names(at))  
 d0[[i]] = at[[i]]  
 d1[[i]] = at[[i]]  
 }  
   
 # Create matrices for prediction based on the model  
 Xp0 <- predict(mod, newdata = d0, type="lpmatrix")  
 Xp1 <- predict(mod, newdata = d1, type="lpmatrix")  
   
 # Extract model parameters  
 ilink <- family(mod)$linkinv  
 beta <- coef(mod)  
 vc <- mod$Vc # covariance matrix   
   
  
 # Compute marginal effects based on the adjusted data  
 pred0 <- 1 - (1-ilink(Xp0 %\*% beta))^offset  
 pred1 <- 1 - (1-ilink(Xp1 %\*% beta))^offset  
 ME <- (pred1-pred0)  
   
 # Adjust for numerical derivative if change is NULL  
 if (is.null(change)) {  
 ME <- ME/(d1[[term]] - d0[[term]])  
 }   
   
 # convert to relative if requested  
 if (relative == T) ME = ME/pred0  
   
 # average marginal effect  
 AME = mean(ME)  
   
   
 # Simulate AMEs to compute uncertainty in the estimates  
   
 # Compute the variance of the average marginal effect through a   
 # "posterior" simulation.  
 # This involves simulating from a multivariate normal distribution   
 # using the model's   
 #coefficient means and covariance matrix  
   
 if (!is.null(seed)) set.seed(seed)  
 coefmat = mvrnorm(n = iterations  
 , mu = beta  
 , Sigma = vc)  
   
 # For each simulated coefficient vector, estimate the Average   
 # Marginal Effect (AME).  
 AMEs = apply(coefmat, 1, function(coefrow) {  
   
 # Calculate marginal effects based on the simulated coefficient  
 pred0 <- 1 - (1-ilink(Xp0 %\*% coefrow))^offset  
 pred1 <- 1 - (1-ilink(Xp1 %\*% coefrow))^offset  
 ME <- (pred1-pred0)  
   
 # if change is NULL, use numerical derivative  
 if (is.null(change)) {  
 ME <- ME/(d1[[term]] - d0[[term]])  
 }   
   
 # convert to relative if requested  
 if (relative == T) ME = ME/pred0  
   
 # average marginal effect  
 AME = mean(ME)  
 return(AME)  
 })  
   
 # Combine results  
 # If the 'samples' flag is FALSE, return the summary results.  
 # Otherwise, return both the summary and the sample results.  
   
 if (!samples) {  
 res = data.frame(term  
 , estimate = AME  
 , std.error = sqrt(var(AMEs))   
 , estimate.sim = mean(AMEs)   
 , offset  
 , change.value = paste(change, collapse = "\_"))  
 return(res)   
   
 } else {  
   
 res\_sums = data.frame(term  
 , estimate = AME  
 , std.error = sqrt(var(AMEs))   
 , offset  
 , change.value = paste(change, collapse = "\_"))  
   
 res\_samples = data.frame(term  
 , estimate = AMEs  
 , MLE = AME  
 , offset  
 , change.value = paste(change,   
 collapse = "\_"))  
 res = list(res\_sums, res\_samples)  
 return(res)   
   
 }  
}

### 3.9.3 Calculating Absolute Average Marginal Effect (AMEs)

At the end of this code segment, the **AME** data frame contains average marginal estimates for each predictor, each corresponding to different change scenarios. In essence, AME provides the average effect that a predictor has on the outcome across these scenarios. On the other hand, the **AMEsamples** data frame contains multiple samples of AME for each predictor, each aligned with a specific type of change scenario, which allows for an evaluation of the uncertainty inherent in the AME estimates.

# Absolute AMEs ---------------------------------------------------  
  
# Initialize empty data frames to store the results  
AME = data.frame()  
AMEsamples = data.frame()  
  
# Loop through predictor names that match "con\_"  
for (i in names(predictors)[grepl("con\_", names(predictors))]) {   
  
# Loop through different change settings (e.g., equilibrium, invasion,  
 # iqr)  
 for (j in names(change)) {  
   
# Calculate the AME for each model in res\_mod  
 temp = lapply(res\_mod, function(x){  
   
# If the change is based on IQR (interquartile range), calculate the  
 # 1st and 3rd quartiles  
 if (j == "iqr") {  
 q1 = quantile(x$model$con\_dens, probs = 0.25)  
 q3 = quantile(x$model$con\_dens, probs = 0.75)  
 }  
  
# Use the get\_AME function to calculate the AME for the current model  
 get\_AME(x  
 , data = x$model  
 , offset = interval  
 , term = i  
 , change = eval(parse(text = change[[j]][,i]))  
 , iterations = iter  
 , samples = T  
 )  
 }  
 )  
   
 # AME  
 tempAME = lapply(temp, function(x) x[[1]])  
 tempAME = Map(cbind, tempAME, change = j, sp = names(tempAME))  
 tempAME = do.call(rbind, tempAME)  
 AME = rbind(AME, tempAME)  
   
 # AME samples  
 tempSamples = lapply(temp, function(x) x[[2]])  
 tempSamples = Map(cbind, tempSamples, change = j,   
 sp = names(tempSamples), iter = iter)  
 tempSamples = do.call(rbind, tempSamples)  
 AMEsamples = rbind(AMEsamples, tempSamples)  
 }  
}  
head(AME)

term estimate std.error offset change.value change sp  
ACALDI con\_dens 0.017532523 0.016546708 1 + 1 equilibrium ACALDI  
AEGICE con\_dens 0.089367793 0.025930769 1 + 1 equilibrium AEGICE  
BEILPE con\_dens 0.006043419 0.001062937 1 + 1 equilibrium BEILPE  
CAPPFR con\_dens 0.020944544 0.004310472 1 + 1 equilibrium CAPPFR  
CECRIN con\_dens 0.028621858 0.027053981 1 + 1 equilibrium CECRIN  
CORDLA con\_dens 0.013464775 0.030623821 1 + 1 equilibrium CORDLA

### 3.9.4 Calculating Relative Average Marginal Effect (rAMEs)

At the end of this code segment, the **rAME** data frame contains the**rAME** estimates for the predictor and each type of change, and the **rAMEsamples** data frame contains the **rAME** samples for each predictor and type of change.

# Relative rAMEs -----------------------------------------------------------  
  
  
# Initialize empty data frames to store the results  
rAME = data.frame()  
rAMEsamples = data.frame()  
  
# Loop through predictor names that match "con\_"   
for (i in names(predictors)[grepl("con\_", names(predictors))]) {   
   
# Loop through different change settings   
 # (e.g., equilibrium, invasion, iqr)  
 for (j in names(change)) {  
  
# Calculate the relative AME (rAME) for each model in res\_mod  
 temp = lapply(res\_mod, function(x){  
   
# If the change is based on IQR (interquartile range),   
 # calculate the 1st and 3rd quartiles  
 if (j == "iqr") {  
 q1 = quantile(x$model$con\_dens, probs = 0.25)  
 q3 = quantile(x$model$con\_dens, probs = 0.75)  
 }  
# Use the get\_AME function to calculate the rAME for the   
 # current model, setting the 'relative' argument to TRUE  
 get\_AME(x  
 , data = x$model  
 , offset = interval  
 , term = i  
 , change = eval(parse(text = change[[j]][, i]))  
 , iterations = iter  
 , relative = T  
 , samples = T  
 )  
 }  
 )  
   
 # rAME  
 tempAME = lapply(temp, function(x) x[[1]])  
 tempAME = Map(cbind, tempAME, change = j, sp = names(tempAME))  
 tempAME = do.call(rbind, tempAME)  
 rAME = rbind(rAME, tempAME)  
   
 # rAME samples  
 tempSamples = lapply(temp, function(x) x[[2]])  
 tempSamples = Map(cbind, tempSamples, change = j,   
 sp = names(tempSamples), iter = iter)  
 tempSamples = do.call(rbind, tempSamples)  
 rAMEsamples = rbind(rAMEsamples, tempSamples)  
 }  
}  
head(rAME)

term estimate std.error offset change.value change sp  
ACALDI con\_dens 0.12190548 0.122096554 1 + 1 equilibrium ACALDI  
AEGICE con\_dens 0.46332386 0.140120748 1 + 1 equilibrium AEGICE  
BEILPE con\_dens 0.02026319 0.003624488 1 + 1 equilibrium BEILPE  
CAPPFR con\_dens 0.42931069 0.087277501 1 + 1 equilibrium CAPPFR  
CECRIN con\_dens 0.04121409 0.042571758 1 + 1 equilibrium CECRIN  
CORDLA con\_dens 0.10440629 0.241547627 1 + 1 equilibrium CORDLA

## 3.10 Saving results

# Save results ---------------------------------------------------  
save(list = c("AME", "AMEsamples", "rAME", "rAMEsamples", "nsp",   
 "coefs", "sums") # "nsp\_data\_deficient"  
 , file = paste0( "./data/mortality.Rdata"))  
write.csv(AME, "./data/AME.csv")  
write.csv(rAME, "./data/rAME.csv")

# 4. Using Meta-regressions to Compare CDD across Species or Sites

## 4.1 Overview

Following from *Section 4. How does CDD vary across species, abiotic gradients or in time?* in the main text, here, we demonstrate one approach to comparing the strength of CDD across species using a meta-analysis framework. The same approach can be used to compare CDD across sites, plots, or any other unit of interest, as long as it is possible to generate reliable estimates of the strength of CDD (suitable sample sizes, etc.). As noted in the main text, “*Correct propagation of uncertainty in CDD estimates requires meta-regressions in frequentist or Bayesian frameworks. Weighted regressions (e.g., lm, lmer, gam) can also estimate how CDD varies e.g., across latitude or between species with different life-history strategies, but incorrectly estimate the associated uncertainty.*”

We use a subset of the BCI seedling data (Comita et al. 2023) of only 30 species to compare how species abundance is related to the strength of CDD at the species level. This is the same dataset used in section 2. This analysis is for demonstration purposes only, and biological conclusions should not be made about the results, given this is only a small subset of the data.

An advantage of using meta-regressions over simple weighted regressions is that the models are able to simultaneously account for uncertainty in species-specific CDD estimates as well as systematic differences in species’ CDD when regressed against a predictor via the inclusion of random effects. A simple weighted regression, on the other hand, assumes that there is only one error for which the relative strength is known when regressing the estimates against a predictor. The latter can lead to incorrect weighting of species in the metaregression.

In this tutorial, we use relative average marginal effect (𝑟𝐴𝑀𝐸) calculated in the previous chapter as our response variable, calculated separately for each species. 𝑟𝐴𝑀𝐸 in this case estimates the relative increase in the probability of annual mortality with the addition of one new conspecific neighbor, while keeping total densities constant. Positive numbers indicate a relative increase in mortality with an increase in conspecific density, a signature of NCDD. In principle, any metric of the strength of CDD can be used, though care must be taken to ensure that the metrics are comparable across species and sites (see main text for more information).

We use the popular [metafor package](https://www.metafor-project.org/doku.php/metafor) to fit the meta-regression models.

|  |
| --- |
| Note |
| The following code is adapted from the [latitudinalCNDD repository](https://github.com/LisaHuelsmann/latitudinalCNDD/tree/main/code) by [Lisa Hülsmann](https://demographicecology.com/). |

## 4.2 Load libraries and data

# Load libraries  
library(tidyr)  
library(dplyr)  
library(readr)  
library(ggplot2)  
library(here)  
library(metafor)  
  
# Load in species abundances for BCI data subset  
abundances <- read\_csv(  
here("./data/BCI seedling data - 30 species - abundance 2023\_05\_18.csv")  
 )  
  
# Load marginal effects calculations from previous section  
load(here("./data/mortality.Rdata"))  
  
# Subset down to just equilibrium change  
rAME <- rAME %>%  
 filter(change == "equilibrium")  
  
# Join marginal effects and abundance data  
rAME <- left\_join(rAME, abundances, by = c("sp" = "spp"))  
  
# Add in average abundance for 'rare' species  
rAME$abundance[rAME$sp == "data\_deficient\_seedling"] <-   
 mean(abundances$abundance[abundances$spp %in%   
 nsp$spp[nsp$data\_deficient]])  
  
# Log transform abundance to use in models  
rAME$log\_abundance <- log(rAME$abundance)

Let’s take a quick look at the data set we’ll be working with:

head(rAME, n = 10)

term estimate std.error offset change.value change sp  
1 con\_dens 0.12190548 0.122096554 1 + 1 equilibrium ACALDI  
2 con\_dens 0.46332386 0.140120748 1 + 1 equilibrium AEGICE  
3 con\_dens 0.02026319 0.003624488 1 + 1 equilibrium BEILPE  
4 con\_dens 0.42931069 0.087277501 1 + 1 equilibrium CAPPFR  
5 con\_dens 0.04121409 0.042571758 1 + 1 equilibrium CECRIN  
6 con\_dens 0.10440629 0.241547627 1 + 1 equilibrium CORDLA  
7 con\_dens 0.11658356 0.200294935 1 + 1 equilibrium DAVINI  
8 con\_dens 0.01113154 0.123539647 1 + 1 equilibrium DESMAX  
9 con\_dens 0.23375216 0.227252222 1 + 1 equilibrium HEISCO  
10 con\_dens -0.06498043 0.061086207 1 + 1 equilibrium JUSTGR  
 abundance log\_abundance  
1 169 5.129899  
2 153 5.030438  
3 5574 8.625868  
4 1604 7.380256  
5 51 3.931826  
6 248 5.513429  
7 27 3.295837  
8 55 4.007333  
9 56 4.025352  
10 136 4.912655

## 4.3 Reformat data for model fitting

First, we use the ‘escalc’ function in the metafor package to essentially repackage our data frame into a format used in the meta-regression model fitting. Since we already calculated our effect size (𝑟𝐴𝑀𝐸), we just pass through the 𝑟𝐴𝑀𝐸 estimate and corresponding standard error using the ‘GEN’ option for the ‘measure’ argument, rather than calculating an effect size within the ‘escalc’ function.

# Reformat data for model fitting  
# Set measure to generic, which passes the observed effect sizes or   
# outcomes via the yi argument and the corresponding sampling   
# variances via the vi argument (or the standard errors via the sei   
# argument) to the function.  
 dat\_meta = metafor::escalc(measure = "GEN",   
 yi = estimate, # observed outcomes  
 sei = std.error, # standard errors  
 slab = sp, # label for species  
 data = rAME)

## 4.4 Fit meta-regression model

Next, we use the ‘rma’ function to fit a meta-regression model, where 𝑟𝐴𝑀𝐸 is our response variable (renamed as yi in the previous step) and log species abundance is our predictor. While not shown here, it is possible to fit mixed effects meta-regression models with the ‘rma.mv’ function. We suggest consulting the [extensive documentation for the metafor package](https://www.metafor-project.org/doku.php/metafor) for further details.

# Fit model  
metamod = metafor::rma(yi = yi,  
 vi = vi,  
 mods = ~ log\_abundance,  
 method = "REML",  
 data = dat\_meta)

## 4.5 Print model summary

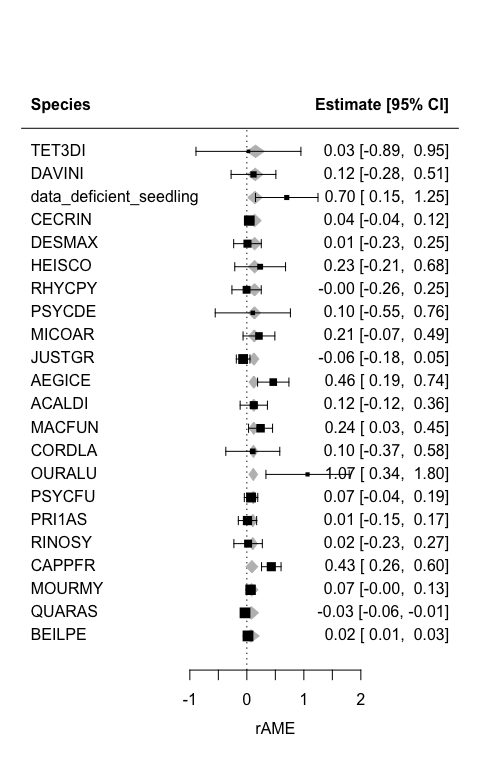
summary(metamod)

Mixed-Effects Model (k = 22; tau^2 estimator: REML)  
  
 logLik deviance AIC BIC AICc   
 2.2364 -4.4727 1.5273 4.5145 3.0273   
  
tau^2 (estimated amount of residual heterogeneity): 0.0166 (SE = 0.0090)  
tau (square root of estimated tau^2 value): 0.1287  
I^2 (residual heterogeneity / unaccounted variability): 91.39%  
H^2 (unaccounted variability / sampling variability): 11.62  
R^2 (amount of heterogeneity accounted for): 0.00%  
  
Test for Residual Heterogeneity:  
QE(df = 20) = 74.0420, p-val < .0001  
  
Test of Moderators (coefficient 2):  
QM(df = 1) = 0.3751, p-val = 0.5402  
  
Model Results:  
  
 estimate se zval pval ci.lb ci.ub   
intrcpt 0.1935 0.1401 1.3815 0.1671 -0.0810 0.4681   
log\_abundance -0.0141 0.0231 -0.6125 0.5402 -0.0594 0.0311   
  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 4.6 Plot the estimated of rAME for all species with a forest plot

In this case, a forest plot shows the estimates of the strength of CDD for individual species, here ordered by least to most abundant going from top to bottom.

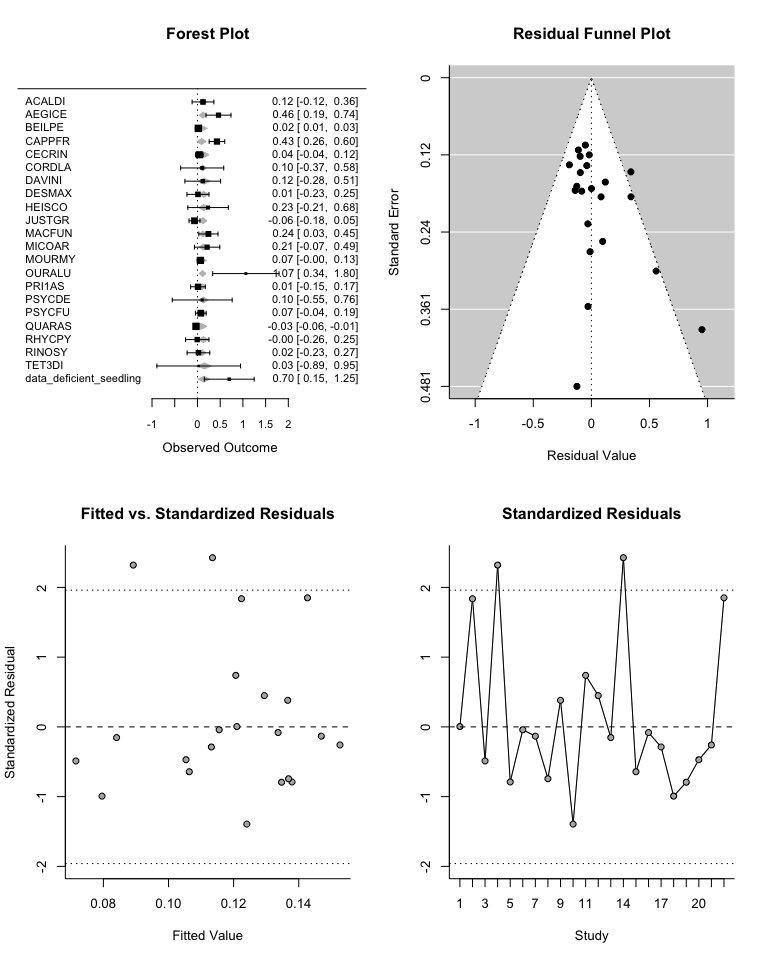
forest(metamod,   
 header = "Species",   
 xlab = "rAME",  
 order = log\_abundance)



## 4.7 Model diagnostics

The plot method displays model diagnostics of the meta-regression model ([more info here](https://wviechtb.github.io/metafor/reference/plot.rma.html)) in addition to a forest plot.

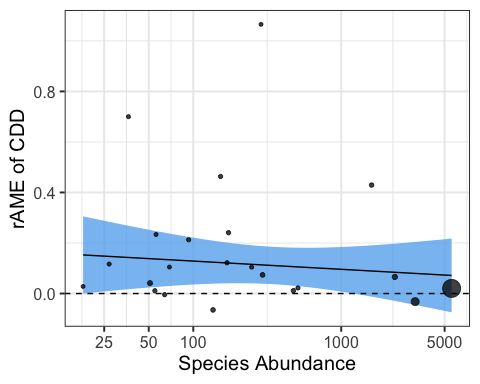
plot(metamod)



## 4.8 Model predictions of how species abundance is related to strength of CDD

Here, we generate predictions and corresponding confidence intervals for how our predictor of interest is related to the related to strength of CDD using the ‘predict’ function. The y-axis here indicates the relative increase in annual mortality probablity with the addition of one conspecific neighbor. Higher values indicate stronger negative conspecific density dependence. In this example, our predictor of interest is species abundance. We also scale the size of the points based on their weight in the meta-analysis, with larger points indicating higher weights. Note that this analysis is for demonstration purposes only, and biological conclusions should not be made about the results, given this is only a small subset of the data.

# Generate a prediction dataframe  
pred <- expand\_grid(log\_abundance = seq(min(dat\_meta$log\_abundance,   
 na.rm = TRUE),   
 max(dat\_meta$log\_abundance,   
 na.rm = TRUE),  
 length.out = 50))  
  
pred$abundance <- exp(pred$log\_abundance) # Back transform abundance  
  
# Bind predictions to dataframe  
pred <- cbind(pred, predict(object = metamod,   
 newmods = pred$log\_abundance))  
  
# Extract observed values  
observed\_values <- broom::augment(metamod)  
  
# Add in variance estimates to be able to scale size of points by   
# amount of variance in estimate  
observed\_values <- left\_join(observed\_values,  
 dat\_meta %>%   
 dplyr::select(sp, vi, log\_abundance),  
 by = c(".rownames" = "sp"))  
  
  
  
# Set abundance values for x axis  
abundances\_x\_axis <- c(25, 50, 100, 1000, 5000)  
  
# Plot prediction  
ggplot(pred, aes(x = log\_abundance, y = pred)) +   
 geom\_ribbon(aes(ymin = ci.lb, ymax = ci.ub),   
 fill = "steelblue2", alpha = 0.75) +   
 geom\_line() +   
 geom\_hline(yintercept = 0, lty = 2) +   
 labs(x = "Species Abundance", y = "rAME of CDD") +   
 scale\_x\_continuous(breaks = log(abundances\_x\_axis),  
 labels = abundances\_x\_axis) +   
 # Add observed points  
 geom\_point(data = observed\_values,  
 aes(x = log\_abundance, y = .observed, size = 1/vi),   
 alpha = 0.75) +   
 theme\_bw(15) +  
 theme(legend.position = "none")



In principle, the approach outlined in this appendix of estimating marginal effects followed by a meta-regression analysis can be used to test for a correlation between strength of CDD and other species-level traits (*e.g.*, wood density, leaf area, etc) or to determine how site-level CDD varies with site-level variables, such as latitude (Hülsmann et al. 202x), precipitation, soil fertility, etc.

# References

Barber, Cristina, Andrii Zaiats, Cara Applestein, Lisa Rosenthal, and T Trevor Caughlin. 2022. “Bayesian Models for Spatially Explicit Interactions Between Neighbouring Plants.” *Methods in Ecology and Evolution*.

Broekman, Maarten JE, Helene C Muller-Landau, Marco D Visser, Eelke Jongejans, SJ Wright, and Hans de Kroon. 2019. “Signs of Stabilisation and Stable Coexistence.” *Ecology Letters* 22 (11): 1957–75.

Canham, Charles D, Michael J Papaik, Marı́a Uriarte, William H McWilliams, Jennifer C Jenkins, and Mark J Twery. 2006. “Neighborhood Analyses of Canopy Tree Competition Along Environmental Gradients in New England Forests.” *Ecological Applications* 16 (2): 540–54.

Comita, Liza S, Salomón Aguilar, Stephen P Hubbell, and Rolando Pérez. 2023. “Long-Term Seedling and Small Sapling Census Data from the b Arro c Olorado i Sland 50 Ha f Orest d Ynamics p Lot, p Anama.” *Ecology* 104 (9): e4140.

Currie, Iain D. 2016. “On Fitting Generalized Linear and Non-Linear Models of Mortality.” *Scandinavian Actuarial Journal* 2016 (4): 356–83.

Uriarte, Marı́a, Nathan G Swenson, Robin L Chazdon, Liza S Comita, W John Kress, David Erickson, Jimena Forero-Montaña, Jess K Zimmerman, and Jill Thompson. 2010. “Trait Similarity, Shared Ancestry and the Structure of Neighbourhood Interactions in a Subtropical Wet Forest: Implications for Community Assembly.” *Ecology Letters* 13 (12): 1503–14.