

Beyond carbon accounting: integrating stand structure and community composition for understanding aboveground carbon storage in Sudanian drylands

Ngaba W. T. Caleb et al., 2026

Detailed statistical analysis process

Introduction

This analysis addresses the estimation of carbon stocks in protected areas of Chad, in Central Africa. The study site is located in Manda National Park (PNM), covering an area of 120,000 ha, in the Sudanian zone of south-eastern Chad. The data consist essentially of dendrometric measurements of the woody vegetation of PNM. They were collected from 14 November to 3 December 2016 along seven (7) transects, each 1.2 km long and 20 m wide. The linear transect method was applied for data collection. All trees/natural resprouts with a circumference ≥ 2 cm were systematically inventoried 10 m on either side of each transect. The sampling rate was 1.4% of the surface area of PNM, corresponding to 16.8 ha surveyed. The measurements taken were: circumference (cm) and diameter (cm) at breast height, as well as the height (m) of the trunk up to the base of the crown. The dataset contains 10 variables covering 5,968 observations (number of trees identified). It is distributed across 18 families, 44 species, and 38 genera, one of which is unidentified. Carbon stock estimation was based on diameter at breast height (DBH), calculated from the circumference at breast height of each inventoried individual. Basal area was also calculated from circumference (Circum_m), while aboveground biomass (AGB, here Wood_vol_m3) was derived from basal area. The carbon sequestered per individual was calculated using the allometric equations of Chave et al. (2005) and those of the IPCC.

Study objective

The objective of our analysis is to assess the atmospheric carbon sequestration potential of the woody vegetation of PNM, as a function of species diversity and the dendrometric parameters of the woody plants in our study area.

Research question

We would like to determine which of the two — species diversity or the dendrometric characteristics of the plants — better predicts the rate of carbon sequestered by the vegetation in PNM.

Hypothesis

We hypothesise that the rate of carbon sequestered by plants is independent of the genus or family of a species, but instead depends on the dendrometric characteristics of the plants. Thus, the higher a plant's dendrometric parameters, the more carbon it sequesters, independently of the genus or family of the species.

Step 1. Loading packages and data

The file PNM_flora_inventory_2016_fv.csv contains dendrometric measurements for 5,968 trees, together with the carbon values calculated for each individual following the allometric equations of Chave et al. (2005) and the IPCC.

```
library(ggplot2)
library(patchwork)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.2.0      v readr      2.2.0
v forcats    1.0.1      v stringr    1.6.0
v lubridate  1.9.5      v tibble     3.3.1
v purrr      1.2.1      v tidyr      1.3.2
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(stringr)
library(ggcorrplot)
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
library(car)
```

```
Loading required package: carData
Attaching package: 'car'
The following object is masked from 'package:dplyr':
  recode
The following object is masked from 'package:purrr':
  some
```

```
library(effectsize)
library(performance)
library(lmtest)
```

```
Loading required package: zoo
Attaching package: 'zoo'
The following objects are masked from 'package:base':
  as.Date, as.Date.numeric
```

```
library(lme4)
```

```
Loading required package: Matrix
Attaching package: 'Matrix'
The following objects are masked from 'package:tidyr':
  expand, pack, unpack
```

```
# Clear the global environment
```

```
rm(list=ls())
```

```
# Load the data (file in the Biostat_cours_14_projet folder)
```

```
dflora <- read.csv2("PNM_flora_inventory_2016_fv.csv", header = TRUE)
```

```
# Quick data exploration
```

```
str(dflora)
```

```
'data.frame':  1048575 obs. of  11 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Species : chr  "Acacia ataxacantha" "Acacia sieberiana" "Acacia sieberiana" "Acacia ..."
 $ Genus   : chr  "Acacia" "Vachellia" "Vachellia" "Vachellia" ...
 $ Family  : chr  "Fabaceae" "Fabaceae" "Fabaceae" "Fabaceae" ...
 $ Height_m : num  18 2.5 8 9.5 12 16 17 18 19 30 ...
 $ Circum_m : num  69 54 23 57 99 51 39 25 21 50 ...
 $ DBH     : num  21.97 17.2 7.32 18.15 31.53 ...
 $ Basal_area_G : num  0.03791 0.02322 0.00421 0.02587 0.07803 ...
 $ Wood_vol_m3 : num  398.4 244 44.3 271.9 820.1 ...
 $ Carbon_Eq_Chave : num  0.311 0.163 0.017 0.188 0.796 ...
 $ Carbon_Eq_GIEC : num  293.9 158.1 18.2 181.3 732.7 ...
```

```
head(dflora)
```

	id	Species	Genus	Family	Height_m	Circum_m	DBH
1	1	Acacia ataxacantha	Acacia	Fabaceae	18.0	69	21.974522
2	2	Acacia sieberiana	Vachellia	Fabaceae	2.5	54	17.197452
3	3	Acacia sieberiana	Vachellia	Fabaceae	8.0	23	7.324841
4	4	Acacia sieberiana	Vachellia	Fabaceae	9.5	57	18.152866
5	5	Acacia sieberiana	Vachellia	Fabaceae	12.0	99	31.528662
6	6	Acacia sieberiana	Vachellia	Fabaceae	16.0	51	16.242038

	Basal_area_G	Wood_vol_m3	Carbon_Eq_Chave	Carbon_Eq_GIEC
1	0.037906051	398.39260	0.31127992	293.94885
2	0.023216561	244.00605	0.16321887	158.10288
3	0.004211783	44.26584	0.01698691	18.24547
4	0.025867834	271.87094	0.18827797	181.27877
5	0.078033439	820.13145	0.79561105	732.72432
6	0.020708599	217.64737	0.14031512	136.81575

```
summary(dflora)
```

	id	Species	Genus	Family
Min. :	1	Length:1048575	Length:1048575	Length:1048575
1st Qu.:	1490	Class :character	Class :character	Class :character
Median :	2979	Mode :character	Mode :character	Mode :character
Mean :	2979			
3rd Qu.:	4468			
Max. :	5957			
NA's :	1042618			

	Height_m	Circum_m	DBH	Basal_area_G
Min. :	1.50	Min. : 2.00	Min. : 0.64	Min. :0.00
1st Qu.:	5.00	1st Qu.: 20.00	1st Qu.: 6.37	1st Qu.:0.00
Median :	8.00	Median : 31.00	Median : 9.87	Median :0.01
Mean :	9.76	Mean : 37.18	Mean : 11.84	Mean :0.02
3rd Qu.:	13.00	3rd Qu.: 47.00	3rd Qu.: 14.97	3rd Qu.:0.02
Max. :	65.00	Max. :350.00	Max. :111.46	Max. :0.98
NA's :	1042622	NA's :1042618	NA's :1042618	NA's :1042617

	Wood_vol_m3	Carbon_Eq_Chave	Carbon_Eq_GIEC
Min. :	0.33	Min. : 0.00	Min. : 0.04
1st Qu.:	33.47	1st Qu.: 0.01	1st Qu.: 12.81
Median :	80.41	Median : 0.04	Median : 38.83
Mean :	173.67	Mean : 0.14	Mean : 136.66
3rd Qu.:	184.85	3rd Qu.: 0.11	3rd Qu.: 111.27
Max. :	10250.60	Max. :16.94	Max. :17884.46
NA's :	1042618	NA's :1042618	NA's :1042618

```
# Checking dimensions
```

```
dim(dflora)
```

```
[1] 1048575      11
```

```
# Check column names
```

```
colnames(dflora)
```

[1] "id"	"Species"	"Genus"	"Family"
[5] "Height_m"	"Circum_m"	"DBH"	"Basal_area_G"
[9] "Wood_vol_m3"	"Carbon_Eq_Chave"	"Carbon_Eq_GIEC"	

```
# Identify missing values (NAs)
```

```
colSums(is.na(dflora))
```

	id	Species	Genus	Family	Height_m
	1042618	0	0	0	1042622
	Circum_m	DBH	Basal_area_G	Wood_vol_m3	Carbon_Eq_Chave
	1042618	1042618	1042617	1042618	1042618
	Carbon_Eq_GIEC				
	1042618				

Step 2. Data cleaning

Taxonomic consistency check

```
# Unique counts
n_species <- n_distinct(dflora$Species)
n_genus   <- n_distinct(dflora$Genus)
n_family  <- n_distinct(dflora$Family)
n_species
```

```
[1] 44
```

```
n_genus
```

```
[1] 38
```

```
n_family
```

```
[1] 18
```

```
# Individuals per species
species_count <- dflora %>%
  group_by(Species) %>%
  summarise(n_individuals = n()) %>%
  arrange(desc(n_individuals))
```

```
# Individuals per genus
genus_count <- dflora %>%
  group_by(Genus) %>%
  summarise(n_individuals = n()) %>%
  arrange(desc(n_individuals))
```

```
# Individuals per family
family_count <- dflora %>%
  group_by(Family) %>%
  summarise(n_individuals = n()) %>%
  arrange(desc(n_individuals))
```

Cleaning qualitative variable values

```
# Remove whitespace
dflora <- dflora %>%
  mutate(
    Species = str_trim(Species),
    Genus = str_trim(Genus),
    Family = str_trim(Family)
  )
```

```
# Standardise text case
dflora <- dflora %>%
  mutate(
    Species = str_to_sentence(Species),
    Genus = str_to_title(Genus),
    Family = str_to_title(Family)
  )
```

```
# Check correspondence between Species and Genus
dflora %>%
  mutate(Genus_from_species = word(Species, 1)) %>%
  filter(Genus != Genus_from_species)
```

	id	Species	Genus	Family	Height_m	Circum_m	DBH
1	2	Acacia sieberiana	Vachellia	Fabaceae	2.5	54	17.197452
2	3	Acacia sieberiana	Vachellia	Fabaceae	8.0	23	7.324841
3	4	Acacia sieberiana	Vachellia	Fabaceae	9.5	57	18.152866
4	5	Acacia sieberiana	Vachellia	Fabaceae	12.0	99	31.528662
5	6	Acacia sieberiana	Vachellia	Fabaceae	16.0	51	16.242038
6	7	Acacia sieberiana	Vachellia	Fabaceae	17.0	39	12.420382
7	8	Acacia sieberiana	Vachellia	Fabaceae	18.0	25	7.961783
8	9	Acacia sieberiana	Vachellia	Fabaceae	19.0	21	6.687898
9	10	Acacia sieberiana	Vachellia	Fabaceae	30.0	50	15.923567
10	11	Amphiloxus pentaphylla	Spermacoce	Rubiaceae	4.0	53	16.878981
11	1993	Unidentified	Unknown	Unknown	12.0	14	4.458599
12	1994	Unidentified	Unknown	Unknown	12.0	40	12.738853
	Basal_area_G	Wood_vol_m3	Carbon_Eq_Chave	Carbon_Eq_GIEC	Genus_from_species		
1	0.023216561	44.26584	0.016986905	158.10288	Acacia		
2	0.004211783	44.26584	0.016986905	18.24547	Acacia		
3	0.025867834	271.87094	0.188277974	181.27877	Acacia		
4	0.078033439	820.13145	0.795611047	732.72432	Acacia		
5	0.020708599	217.64737	0.140315119	136.81575	Acacia		
6	0.012109873	127.27476	0.068877193	69.40280	Acacia		
7	0.004976115	52.29896	0.021171788	22.53055	Acacia		
8	0.003511146	36.90215	0.013367054	14.49434	Acacia		
9	0.019904459	209.19586	0.133148164	130.13007	Acacia		
10	0.022364650	235.05247	0.155348002	150.80006	Amphiloxus		
11	0.001560510	16.40096	0.004645338	5.19622	Unidentified		
12	0.012738854	133.88535	0.073667668	73.99380	Unidentified		

Cleaning quantitative variables

```
num_vars <- dflora %>%
  select(Height_m, Circum_m, DBH, Basal_area_G,
         Wood_vol_m3, Carbon_Eq_Chave, Carbon_Eq_GIEC)
```

```
# Summary statistics
summary(num_vars)
```

(same summary statistics as reported for the full data frame above)

```
# Detection of negative or impossible values
```

```
dflora %>%
  filter(
    Height_m < 0 |
    Circum_m < 0 |
    DBH < 0 |
    Basal_area_G < 0 |
    Wood_vol_m3 < 0 )
```

```
[1] id          Species      Genus        Family        Height_m
[6] Circum_m    DBH          Basal_area_G Wood_vol_m3    Carbon_Eq_Chave
[11] Carbon_Eq_GIEC
<0 rows> (or 0-length row.names)
```

```
# Outlier detection using IQR
```

```
detect_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  x < (Q1 - 1.5 * IQR) | x > (Q3 + 1.5 * IQR) }
```

```
outliers <- dflora %>%
  mutate(across(where(is.numeric), detect_outliers))
```

Checking categorical variables

```
# Frequency tables
```

```
table(dflora$Species)
```

Acacia ataxacantha	Acacia sieberiana	Amphiloxus pentaphylla
1	9	1
Annona senegalensis	Anogeissus leiocarpus	Balanites aegyptiaca
32	588	1
Bombax costatum	Calotropis procera	Cedrela odorata
29	1	46
Combretum collinum	Crossopteryx febrifuga	Daniellia oliveri
1088	54	136
Diospyros mespiliformis	Ficus glumosa	Ficus gnaphalocarpa
5	1	1
Ficus ingens	Flueggea virosa	Gardenia erubescens
1	63	8
Gardenia ternifolia	Grewia venusta	Guiera senegalensis
68	230	45
Hexalobus monopetalus	Hibiscus squamosus	Hymenocardia acida
14	1	930
Khaya senegalensis	Mitracarpus scaber	Mitragyna inermis
202	23	1
Parkia biglobosa	Piliostigma reticulatum	Prosopis africana
384	203	87
Sarcocephalus latifolius	Sclerocarya birrea	Strychnos spinosa
8	1	227
Tamarindus indica	Tephrosia bracteolata	Terminalia avicennioides
12	1	16
Terminalia laxiflora	Unidentified	Vitellaria paradoxa
1351	2	39
Vitex doniana	Vitex simplicifolia	Ziziphus mauritiana
25	7	14
Ziziphus mucronata		
1		

```
table(dflora$Genus)
```

Acacia	Annona	Anogeissus	Balanites	Bombax	Calotropis	Cedrela
1	32	588	1	29	1	46
Combretum	Crossopteryx	Daniellia	Diospyros	Ficus	Flueggea	Gardenia
1088		54	136	5	3	63
Grewia	Guiera	Hexalobus	Hibiscus	Hymenocardia	Khaya	Mitracarpus
230	45	14	1	930	202	23
Mitragyna	Parkia	Piliostigma	Prosopis	Sarcocephalus	Sclerocarya	Spermacoce
1	384	203	87	8	1	1
Strychnos	Tamarindus	Tephrosia	Terminalia	Unknown	Vachellia	
227	12	1	1367	2	9	
Vitellaria	Vitex	Ziziphus				
39	32	15				

```
table(dflora$Family)
```

Anacardiaceae	Annonaceae	Apocynaceae	Combretaceae	Ebenaceae
1	1	1	3088	5
Fabaceae	Lamiaceae	Loganiaceae	Malvaceae	Meliaceae
833	32	227	260	248
Moraceae	Phyllanthaceae	Rhamnaceae	Rubiaceae	Sapotaceae
3	993	15	163	39
Unknown	Zygophyllaceae			
2	1			

```
# Identification of rare species
```

```
rare_species <-
```

```
species_count %>%
```

```
filter(n_individuals == 1)
```

Data validation checks

```
# Consistency check: DBH vs circumference ( $C = \pi \times D$ )
```

```
dflora %>%
  mutate(
    expected_DBH = Circum_m / pi,
    diff = abs(DBH - expected_DBH)
  ) %>%
  arrange(desc(diff)) %>% head(10)
```

	id	Species	Genus	Family	Height_m	Circum_m
1	5414	Terminalia laxiflora	Terminalia	Combretaceae	12.0	350
2	4944	Terminalia laxiflora	Terminalia	Combretaceae	8.0	310
3	5761	Terminalia laxiflora	Terminalia	Combretaceae	18.0	300
4	4217	Prosopis africana	Prosopis	Fabaceae	8.0	240
5	4611	Terminalia laxiflora	Terminalia	Combretaceae	4.0	235
6	1720	Combretum collinum	Combretum	Combretaceae	16.0	233
7	3403	Khaya senegalensis	Khaya	Meliaceae	11.0	233
8	2391	Guiera senegalensis	Guiera	Combretaceae	3.5	230
9	3195	Hymenocardia acida	Hymenocardia	Phyllanthaceae	15.0	230
10	4426	Strychnos spinosa	Strychnos	Loganiaceae	8.0	230

	DBH	Basal_area_G	Wood_vol_m3	Carbon_Eq_Chave	Carbon_Eq_GIEC	expected_DBH
1	111.46497	0.9753185	10250.597	16.943458	17884.455	111.40846
2	98.72611	0.7651274	8041.489	12.896157	13156.145	98.67606
3	95.54140	0.7165605	7531.051	11.968267	12108.779	95.49297
4	76.43312	0.4585987	4819.873	7.129785	6885.223	76.39437
5	74.84076	0.4396895	4621.137	6.783993	6528.077	74.80282
6	74.20382	0.4322373	4542.814	6.648181	6388.429	74.16620
7	74.20382	0.4322373	4542.814	6.648181	6388.429	74.16620
8	73.24841	0.4211783	4426.584	6.447149	6182.370	73.21127
9	73.24841	0.4211783	4426.584	6.447149	6182.370	73.21127
10	73.24841	0.4211783	4426.584	6.447149	6182.370	73.21127

	diff
1	0.05650804
2	0.05004993
3	0.04843541
4	0.03874834
5	0.03794108
6	0.03761818
7	0.03761818
8	0.03713382
9	0.03713382
10	0.03713382

Step 3. Exploratory data visualisation

```
# DBH distribution  
ggplot(dflora, aes(DBH)) + geom_histogram(bins = 30)
```

Warning: Removed 1042618 rows containing non-finite outside the scale range (``stat_bin()``).

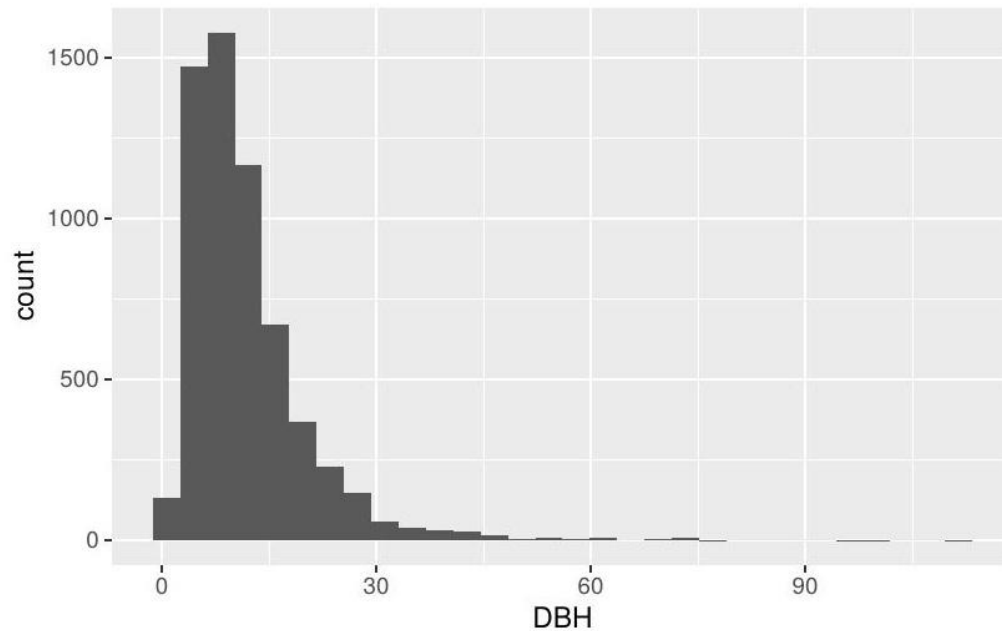


Figure 1. Distribution of diameter at breast height (DBH) for the 5,968 inventoried individuals.

```
# Height vs DBH  
ggplot(dflora, aes(DBH, Height_m)) + geom_point(alpha = 0.3)
```

Warning: Removed 1042622 rows containing missing values or values outside the scale range (``geom_point()``).

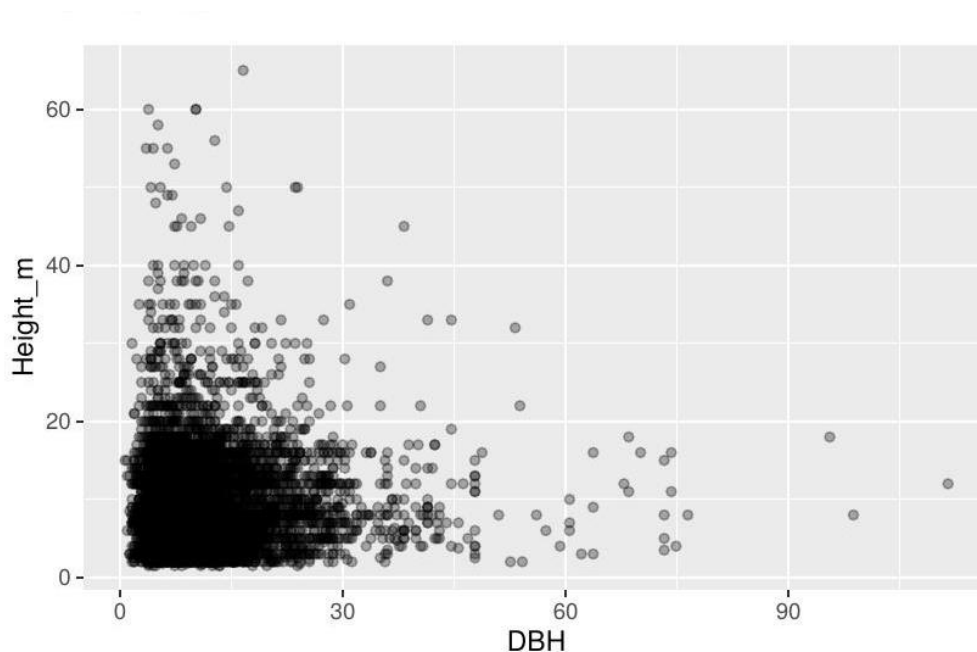


Figure 2. Relationship between DBH and total height.

Step 4. Exploratory analysis of inventory data: Observations – Ecological interpretation – Statistical implications

Graph 1. Distribution of diameters (DBH) of the 5,968 inventoried individuals

Observations

The distribution is strongly right-skewed. The majority of individuals have small diameters, generally below 20 cm. Large diameters (> 40 cm) are rare. A few trees show exceptionally high diameters (> 80 cm), constituting extreme values.

Ecological interpretation

This inverse-J shape is characteristic of naturally regenerating stands. PNM appears to be dominated by small-sized individuals. Large trees are few in number but represent a substantial share of biomass and stored carbon.

Implications for the carbon analysis

1. Sequestered carbon is likely to be very unevenly distributed among individuals. A small proportion of large trees could concentrate a large share of the total carbon stock.
2. Classical linear models risk being influenced by extreme values. A logarithmic transformation of DBH may be necessary to improve the statistical analyses.

Graph 2. Relationship between DBH and height

Observations

A general positive relationship exists between diameter and height. However, the dispersion is very large. For a given diameter, observed heights vary considerably. The relationship appears non-linear. Trees with very large diameters are not systematically the tallest.

Ecological interpretation

Growth in diameter and in height is not perfectly synchronised. Interspecific differences appear to be important: some species invest more in vertical growth while others develop more in diameter.

Implications for the carbon analysis

1. DBH alone does not capture all of the variability in carbon.
2. Height could provide important complementary information.
3. A non-linear relationship is likely between the dendrometric variables and carbon.

Step 5. Exporting the cleaned database

```
# Assign new name to the cleaned database
df_clean <- dflora

# Export as CSV spreadsheet
write.csv2(df_clean, "PNM_inventory_2016_cleandb.csv", row.names = FALSE)
```

Step 6. Logarithmic transformation of the data

```
df_clean <- df_clean %>%
  mutate(
    log_DBH = log(DBH + 1),
    log_Height = log(Height_m + 1),
    log_Basal = log(Basal_area_G + 0.001),
    log_WoodVol = log(Wood_vol_m3 + 1),
    log_Carbon_Chave = log(Carbon_Eq_Chave + 0.001),
    log_Carbon_GIEC = log(Carbon_Eq_GIEC + 1)
  )
```

Step 7. Examining the response variable "Carbon_Eq_Chave"

```
# Histogram
ggplot(df_clean,
  aes(Carbon_Eq_Chave)) +
  geom_histogram(bins = 40)
```

Warning: Removed 1042618 rows containing non-finite outside the scale range (`stat\_bin()`).

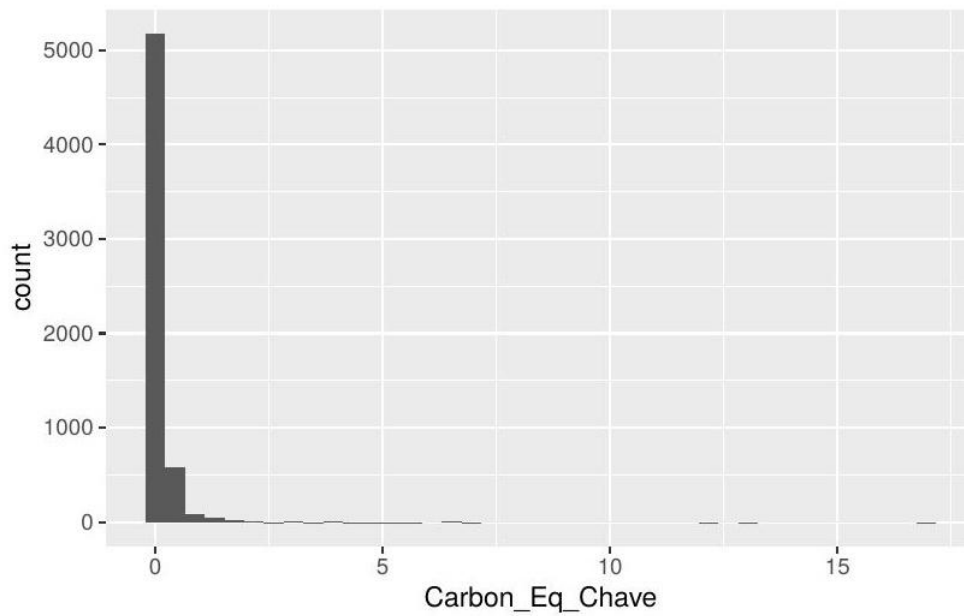


Figure 3. Histogram of raw Carbon_Eq_Chave values.

```
# Logarithmic version  
ggplot(df_clean,  
  aes(log_Carbon_Chave)) +  
  geom_histogram(bins = 40)
```

Warning: Removed 1042618 rows containing non-finite outside the scale range
(`stat\_bin()`).

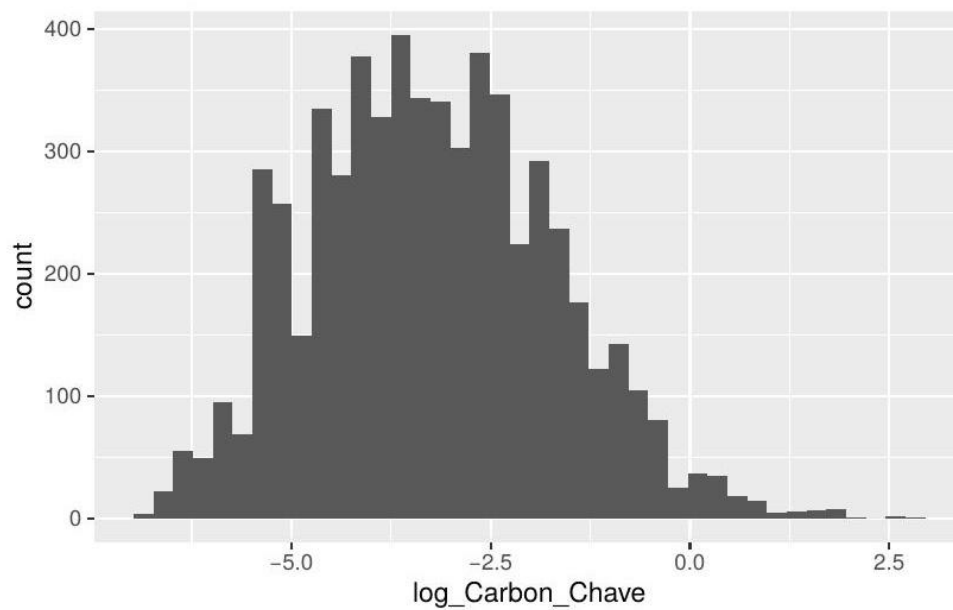


Figure 4. Histogram of log-transformed carbon (log_Carbon_Chave).

```
# Boxplot
ggplot(df_clean,
       aes(y = Carbon_Eq_Chave)) +
  geom_boxplot()
```

Warning: Removed 1042618 rows containing non-finite outside the scale range (`stat\_boxplot()`).

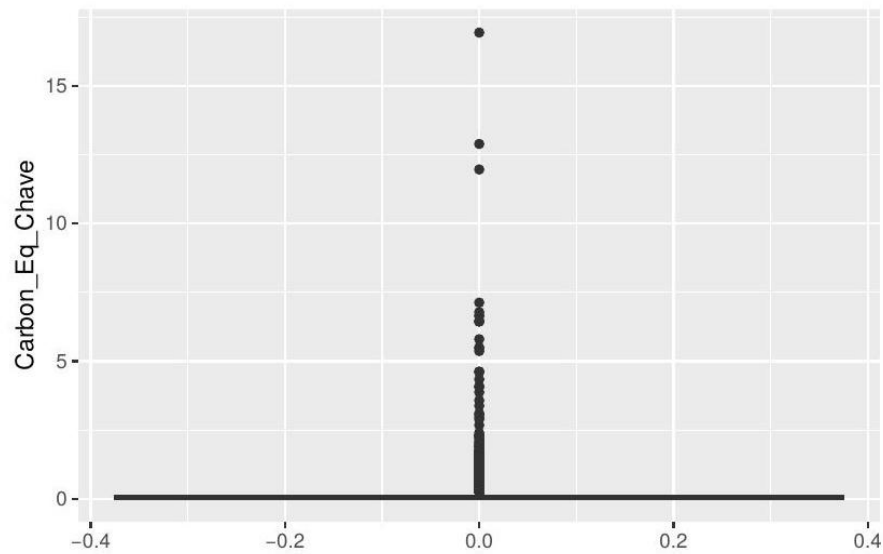


Figure 5. Boxplot of raw Carbon_Eq_Chave values (compressed by the extreme right-skew).

```
# Normality test (Shapiro-Wilk; the test requires a limited sample since n > 5000)
shapiro.test(
  sample(df_clean$log_Carbon_Chave,
        5000)
)
```

Shapiro-Wilk normality test

data: sample(df_clean\$log_Carbon_Chave, 5000)
W = 0.93103, p-value = 0.1028

Step 8. Correlation analysis among dendrometric variables

Selecting the dendrometric variables

```
dendro_vars <- df_clean %>%
  select(
    DBH,
    Height_m,
    Basal_area_G,
    Wood_vol_m3,
    Carbon_Eq_Chave
  )
```

Building the correlation matrix

```
cor_mat <- cor(
  dendro_vars,
  use = "complete.obs"
)
round(cor_mat, 3)
```

	DBH	Height_m	Basal_area_G	Wood_vol_m3	Carbon_Eq_Chave
DBH	1.000	0.012	0.876	0.876	0.771
Height_m	0.012	1.000	0.011	0.011	0.010
Basal_area_G	0.876	0.011	1.000	1.000	0.981
Wood_vol_m3	0.876	0.011	1.000	1.000	0.981
Carbon_Eq_Chave	0.771	0.010	0.981	0.981	1.000

Visualising the correlation matrix

Matrix 1 (with correlation values shown on the plot):

```
ggcorrplot(cor_mat,
  type = "lower",      # Show only the lower triangle to avoid duplicates
  lab = TRUE,          # Show correlation values
  digits = 3,          # Show 3 decimal places
  hc.order = TRUE,     # Group similar variables together
  colors = c("#6D9EC1", "white", "#E46726"),
  # Blue (negative), White (none), Orange (positive)
  title = "Pearson correlation matrix",
  ggtheme = ggplot2::theme_minimal())
```

Warning: `aes\_string()` was deprecated in ggplot2 3.0.0.

i Please use tidy evaluation idioms with `aes()`.

i See also `vignette("ggplot2-in-packages")` for more information.

i The deprecated feature was likely used in the ggcorrplot package.

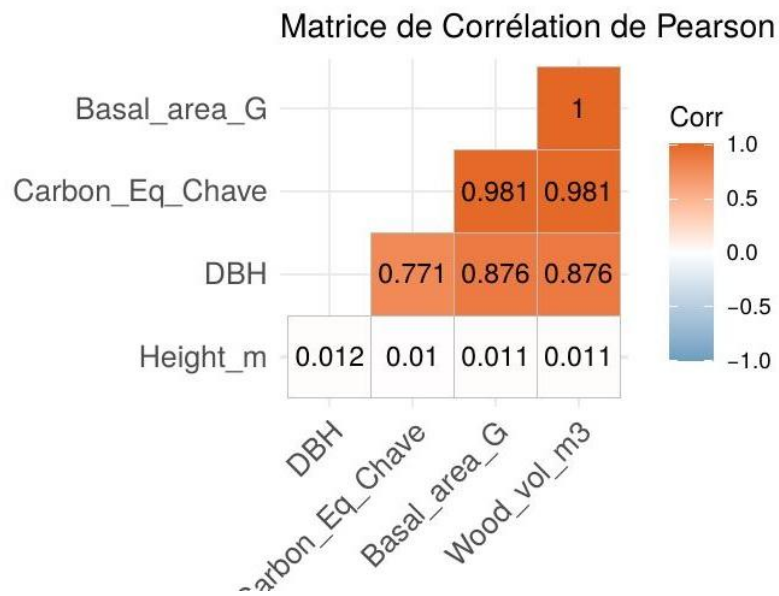


Figure 6. Pearson correlation matrix (ggcorrplot, lower triangle, values labelled).

Matrix 2:

```
corrplot(
  cor_mat,
  method = "color",
  type = "upper",
  tl.cex = 0.9
)
```

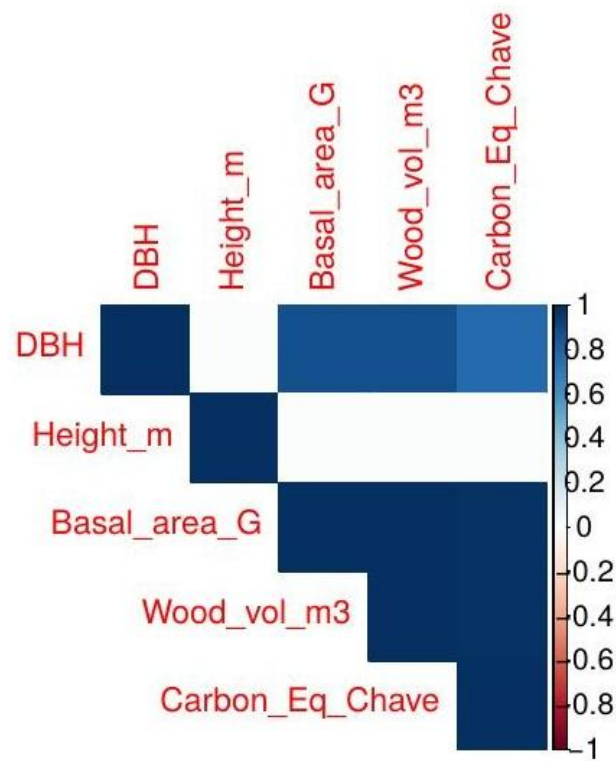


Figure 7. Pearson correlation matrix (corrplot, upper triangle, colour-coded).

Checking collinearity (car package)

```
mod_vif <- lm(
  log_Carbon_Chave ~
    log_DBH +
    log_Height +
    log_Basal +
    log_WoodVol,
  data = df_clean
)
vif(mod_vif)
```

```
      log_DBH   log_Height   log_Basal log_WoodVol
10403.328795    1.000624    521.301355    6661.719156
```

```
#-----
# Expected values:
# *VIF < 5: acceptable*
# *VIF > 10: major problem*
#-----
```

Step 9. Visualising relationships with carbon

Carbon vs DBH

```
ggplot(df_clean,
  aes(DBH,
      Carbon_Eq_Chave)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "lm")
```

`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 1042618 rows containing non-finite / out-of-range values.

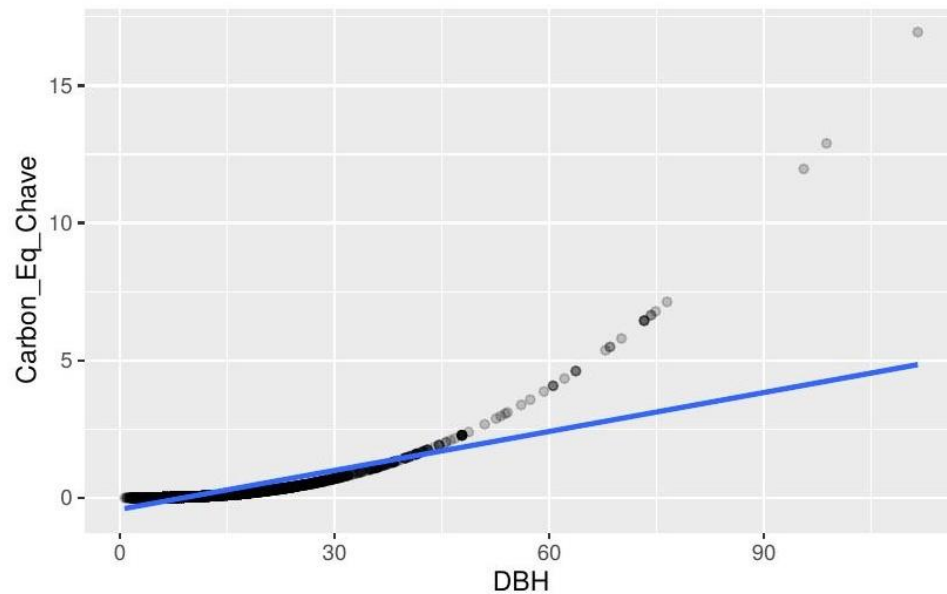


Figure 8. Carbon (Carbon_Eq_Chave) as a function of DBH, with linear fit.

Log-log version

```
ggplot(df_clean,
  aes(log_DBH,
      log_Carbon_Chave)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "lm")
```

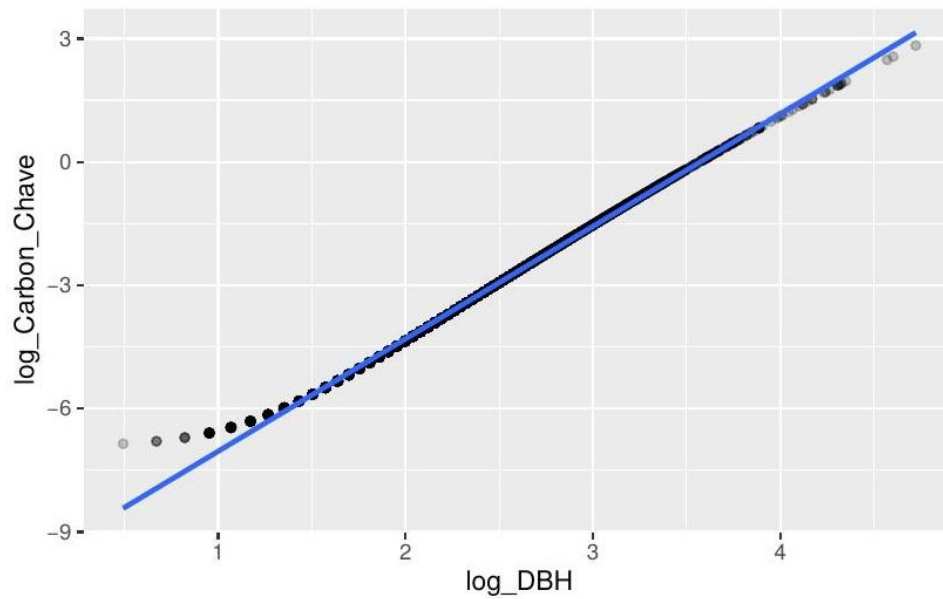



Figure 9. Log-log relationship between DBH and carbon.

Carbon vs height

```
ggplot(df_clean,
  aes(Height_m,
      Carbon_Eq_Chave)) +
  geom_point(alpha = 0.2) +
  geom_smooth()
```

`geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

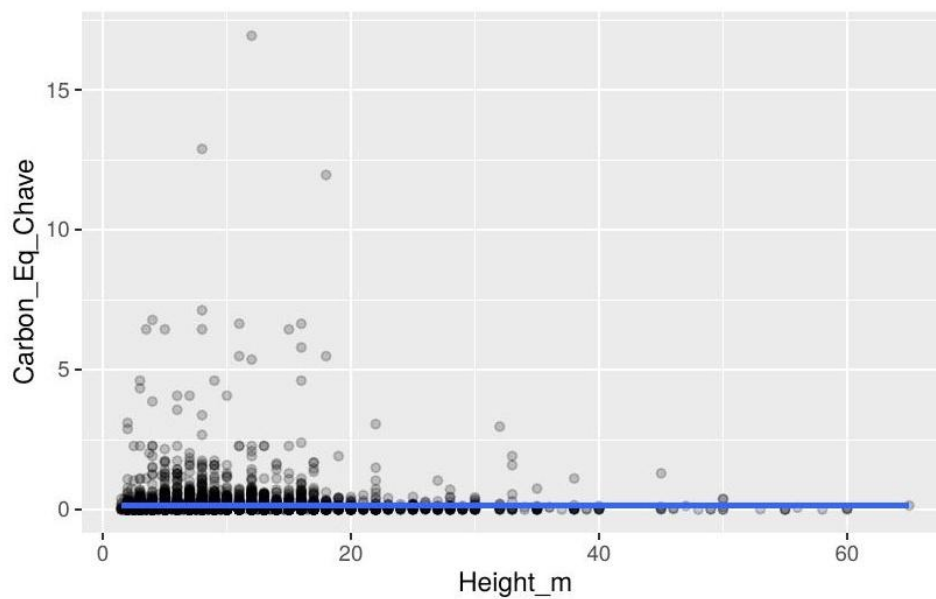


Figure 10. Carbon as a function of height, with GAM smooth.

Carbon vs biomass

```
ggplot(df_clean,  
  aes(Wood_vol_m3,  
    Carbon_Eq_Chave)) +  
  geom_point(alpha = 0.2) +  
  geom_smooth()
```

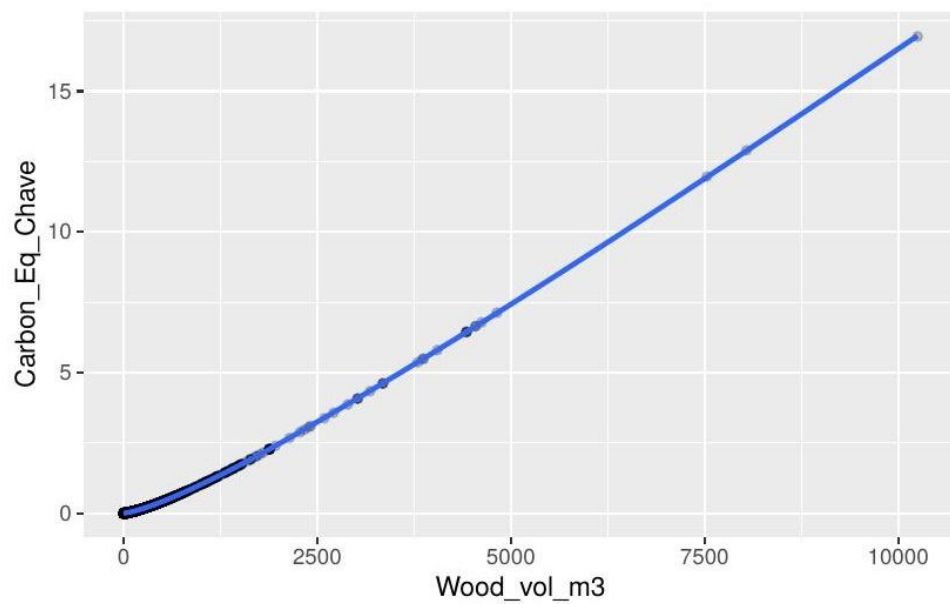


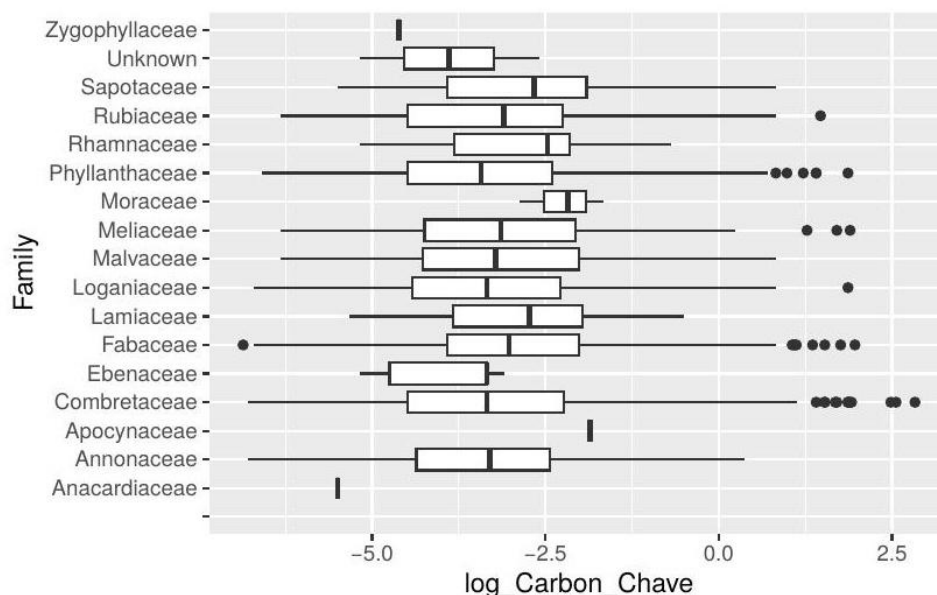
Figure 11. Carbon as a function of wood volume (AGB proxy), with GAM smooth.

Step 10. Effect of taxonomic groups on carbon storage capacity

Carbon stock vs taxonomic family

Visualisation:

```
ggplot(df_clean,
  aes(Family,
    log_Carbon_Chave)) +
  geom_boxplot() +
  coord_flip()
```



ANOVA test:

```
anova_family <- aov(
  log_Carbon_Chave ~ Family,
  data = df_clean
)
summary(anova_family)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Family	16	112	6.993	2.99	5.29e-05 ***
Residuals	5940	13893	2.339		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1042618 observations deleted due to missingness

Effect size:

```
eta_squared(anova_family)
```

For one-way between subjects designs, partial eta squared is equivalent to eta squared. Returning eta squared.

Effect Size for ANOVA

Parameter	Eta2	95% CI

Family | 7.99e-03 | [0.00, 1.00]
 - One-sided CIs: upper bound fixed at [1.00].

Carbon stock vs taxonomic genus

Visualisation:

```
ggplot(df_clean,
  aes(Genus,
    log_Carbon_Chave)) +
  geom_boxplot() +
  coord_flip()
```

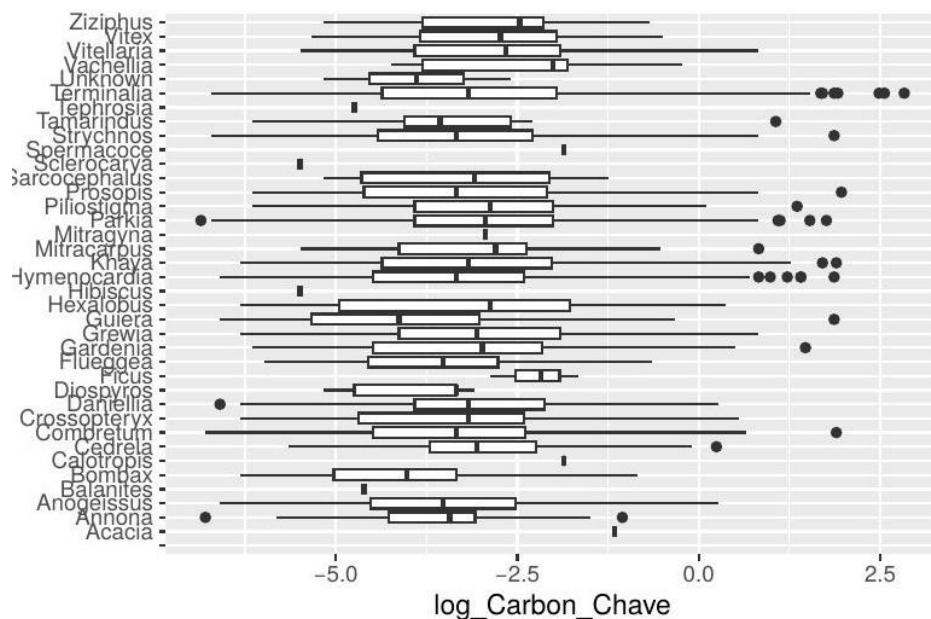


Figure 13. log-transformed carbon by taxonomic genus.

ANOVA test:

```
anova_genus <- aov(
  log_Carbon_Chave ~ Genus,
  data = df_clean
)
summary(anova_genus)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Genus	36	257	7.132	3.071	2.12e-09 ***
Residuals	5920	13748	2.322		

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 1042618 observations deleted due to missingness

```
eta_squared(anova_genus)
```

For one-way between subjects designs, partial eta squared is equivalent to eta squared. Returning eta squared.
 # Effect Size for ANOVA

Parameter	Eta2	95% CI

Genus | 0.02 | [0.01, 1.00]
- One-sided CIs: upper bound fixed at [1.00].

Robust alternative test (Kruskal-Wallis)

Family:

```
kruskal.test(  
  log_Carbon_Chave ~ Family,  
  data = df_clean  
)
```

Kruskal-Wallis rank sum test

data: log_Carbon_Chave by Family
Kruskal-Wallis chi-squared = 51.128, df = 16, p-value = 1.513e-05

Genus:

```
kruskal.test(  
  log_Carbon_Chave ~ Genus,  
  data = df_clean  
)
```

Kruskal-Wallis rank sum test

data: log_Carbon_Chave by Genus
Kruskal-Wallis chi-squared = 109.92, df = 36, p-value = 2.072e-09

Step 11. Building competing models to test our main hypothesis

Creating a common dataset (building a subset containing all variables needed before modelling):

```
df_model <- df_clean %>%  
  select(  
    log_Carbon_Chave,  
    log_DBH,  
    log_Height,  
    Family,  
    Genus  
  ) %>%  
  na.omit()
```

```
# Check
```

```
dim(df_model)
```

```
[1] 5953    5
```

```
colSums(is.na(df_model))
```

```
log_Carbon_Chave    log_DBH    log_Height    Family  
                0          0          0          0  
      Genus  
                0
```

Simple linear modelling

Model 1 (correlation between taxonomy and carbon stock)

```
mod_taxo <- lm(  
  log_Carbon_Chave ~ Family + Genus,  
  data = df_model  
)  
summary(mod_taxo)
```

Call:

```
lm(formula = log_Carbon_Chave ~ Family + Genus, data = df_model)
```

Residuals:

```
      Min       1Q   Median       3Q      Max  
-3.8939 -1.1391 -0.0145  0.9958  5.9337
```

Coefficients: (16 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-5.490e+00	1.524e+00	-3.602	0.000319	***
FamilyAnnonaceae	2.484e+00	1.578e+00	1.575	0.115416	
FamilyApocynaceae	3.634e+00	2.156e+00	1.686	0.091845	.
FamilyCombretaceae	2.386e+00	1.525e+00	1.565	0.117641	
FamilyEbenaceae	1.549e+00	1.670e+00	0.928	0.353701	
FamilyFabaceae	4.326e+00	2.156e+00	2.007	0.044799	*
FamilyLamiaceae	2.637e+00	1.548e+00	1.703	0.088543	.
FamilyLoganiaceae	2.165e+00	1.528e+00	1.418	0.156382	
FamilyMalvaceae	-4.013e-12	2.156e+00	0.000	1.000000	
FamilyMeliaceae	2.312e+00	1.528e+00	1.513	0.130292	
FamilyMoraceae	3.255e+00	1.760e+00	1.849	0.064481	.
FamilyPhyllanthaceae	2.149e+00	1.525e+00	1.409	0.158750	
FamilyRhamnaceae	2.679e+00	1.574e+00	1.702	0.088821	.
FamilyRubiaceae	3.634e+00	2.156e+00	1.686	0.091845	.
FamilySapotaceae	2.704e+00	1.544e+00	1.752	0.079897	.
FamilyUnknown	1.604e+00	1.867e+00	0.859	0.390200	
FamilyZygophyllaceae	8.770e-01	2.156e+00	0.407	0.684137	

GenusAnnona	-5.995e-01	4.884e-01	-1.228	0.219681
GenusAnogeissus	-3.500e-01	7.523e-02	-4.652	3.36e-06 ***
GenusBalanites	NA	NA	NA	NA
GenusBombax	1.518e+00	1.550e+00	0.979	0.327636
GenusCalotropis	NA	NA	NA	NA
GenusCedrela	1.809e-01	2.490e-01	0.727	0.467498
GenusCombretum	-2.917e-01	6.195e-02	-4.709	2.54e-06 ***
GenusCrossopteryx	-1.522e+00	1.538e+00	-0.989	0.322644
GenusDaniellia	-1.893e+00	1.530e+00	-1.237	0.216096
GenusDiospyros	NA	NA	NA	NA
GenusFicus	NA	NA	NA	NA
GenusFlueggea	-2.027e-01	1.984e-01	-1.021	0.307086
GenusGardenia	-1.256e+00	1.534e+00	-0.819	0.413030
GenusGrewia	2.432e+00	1.528e+00	1.592	0.111399
GenusGuiera	-8.347e-01	2.309e-01	-3.614	0.000304 ***
GenusHexalobus	NA	NA	NA	NA
GenusHibiscus	NA	NA	NA	NA
GenusHymenocardia	NA	NA	NA	NA
GenusKhaya	NA	NA	NA	NA
GenusMitracarpus	-1.251e+00	1.557e+00	-0.803	0.421724
GenusMitragyna	-1.088e+00	2.156e+00	-0.505	0.613746
GenusParkia	-1.800e+00	1.526e+00	-1.179	0.238303
GenusPiliostigma	-1.765e+00	1.528e+00	-1.155	0.247980
GenusProsopis	-1.941e+00	1.533e+00	-1.266	0.205589
GenusSarcocephalus	-1.374e+00	1.617e+00	-0.850	0.395486
GenusSclerocarya	NA	NA	NA	NA
GenusSpermacoce	NA	NA	NA	NA
GenusStrychnos	NA	NA	NA	NA
GenusTamarindus	-2.083e+00	1.586e+00	-1.313	0.189323
GenusTephrosia	-3.582e+00	2.156e+00	-1.662	0.096590 .
GenusTerminalia	NA	NA	NA	NA
GenusUnknown	NA	NA	NA	NA
GenusVachellia	-1.324e+00	1.607e+00	-0.824	0.409761
GenusVitellaria	NA	NA	NA	NA
GenusVitex	NA	NA	NA	NA
GenusZiziphus	NA	NA	NA	NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.524 on 5916 degrees of freedom
Multiple R-squared: 0.01829, Adjusted R-squared: 0.01232
F-statistic: 3.062 on 36 and 5916 DF, p-value: 2.369e-09

Model 2 (correlation between dendrometric parameters and carbon stock)

```
mod_dendro <- lm(
  log_Carbon_Chave ~ log_DBH + log_Height,
  data = df_model
)
summary(mod_dendro)
```

Call:

```
lm(formula = log_Carbon_Chave ~ log_DBH + log_Height, data = df_model)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.31401	-0.03618	-0.00481	0.02511	1.56992

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.7765306	0.0053213	-1837.237	<2e-16 ***
log_DBH	2.7357130	0.0016238	1684.811	<2e-16 ***
log_Height	0.0002698	0.0016294	0.166	0.869

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07014 on 5950 degrees of freedom
Multiple R-squared: 0.9979, Adjusted R-squared: 0.9979
F-statistic: 1.42e+06 on 2 and 5950 DF, p-value: < 2.2e-16

Model 3 (correlation between carbon stock and the full set of predictive variables)

```
mod_full <- lm(
  log_Carbon_Chave ~
    log_DBH +
    log_Height +
    Family +
    Genus,
  data = df_model
)
summary(mod_full)
```

Call:

```
lm(formula = log_Carbon_Chave ~ log_DBH + log_Height + Family +
    Genus, data = df_model)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.31161	-0.03721	-0.00542	0.02453	1.56697

Coefficients: (16 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.800371	0.070265	-139.478	< 2e-16 ***
log_DBH	2.735757	0.001636	1672.561	< 2e-16 ***
log_Height	0.002338	0.001855	1.260	0.2077
FamilyAnnonaceae	0.038411	0.072502	0.530	0.5963
FamilyApocynaceae	0.052560	0.099078	0.530	0.5958
FamilyCombretaceae	0.015925	0.070061	0.227	0.8202
FamilyEbenaceae	-0.007341	0.076711	-0.096	0.9238
FamilyFabaceae	0.054714	0.099048	0.552	0.5807
FamilyLamiaceae	0.017853	0.071139	0.251	0.8019
FamilyLoganiaceae	0.024798	0.070202	0.353	0.7239
FamilyMalvaceae	0.004616	0.099083	0.047	0.9628
FamilyMeliaceae	0.014572	0.070203	0.208	0.8356
FamilyMoraceae	0.038116	0.080869	0.471	0.6374
FamilyPhyllanthaceae	0.016188	0.070079	0.231	0.8173
FamilyRhamnaceae	0.021475	0.072386	0.297	0.7667
FamilyRubiaceae	0.052038	0.099068	0.525	0.5994
FamilySapotaceae	0.011859	0.070937	0.167	0.8672
FamilyUnknown	0.002852	0.085757	0.033	0.9735
FamilyZygophyllaceae	-0.035875	0.099017	-0.362	0.7171
GenusAnnona	0.009559	0.022456	0.426	0.6704
GenusAnogeissus	-0.006302	0.003462	-1.820	0.0688 .
GenusBalanites	NA	NA	NA	NA
GenusBombax	0.002882	0.071312	0.040	0.9678
GenusCalotropis	NA	NA	NA	NA
GenusCedrela	0.001248	0.011495	0.109	0.9135
GenusCombretum	0.013141	0.002895	4.539	5.76e-06 ***
GenusCrossopteryx	-0.026782	0.070673	-0.379	0.7047
GenusDaniellia	-0.029892	0.070289	-0.425	0.6706
GenusDiospyros	NA	NA	NA	NA
GenusFicus	NA	NA	NA	NA
GenusFlueggea	-0.007719	0.009169	-0.842	0.3999
GenusGardenia	-0.038708	0.070477	-0.549	0.5829
GenusGrewia	0.020165	0.070193	0.287	0.7739
GenusGuiera	0.009635	0.010711	0.900	0.3684
GenusHexalobus	NA	NA	NA	NA
GenusHibiscus	NA	NA	NA	NA
GenusHymenocardia	NA	NA	NA	NA

GenusKhaya	NA	NA	NA	NA
GenusMitracarpus	-0.043216	0.071534	-0.604	0.5458
GenusMitragyna	-0.032046	0.099026	-0.324	0.7462
GenusParkia	-0.033681	0.070115	-0.480	0.6310
GenusPiliostigma	-0.034855	0.070215	-0.496	0.6196
GenusProsopis	-0.044973	0.070443	-0.638	0.5232
GenusSarcocephalus	-0.041390	0.074274	-0.557	0.5774
GenusSclerocarya	NA	NA	NA	NA
GenusSpermacoce	NA	NA	NA	NA
GenusStrychnos	NA	NA	NA	NA
GenusTamarindus	-0.038840	0.072907	-0.533	0.5942
GenusTephrosia	-0.088318	0.099073	-0.891	0.3727
GenusTerminalia	NA	NA	NA	NA
GenusUnknown	NA	NA	NA	NA
GenusVachellia	-0.032124	0.073808	-0.435	0.6634
GenusVitellaria	NA	NA	NA	NA
GenusVitex	NA	NA	NA	NA
GenusZiziphus	NA	NA	NA	NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07001 on 5914 degrees of freedom

Multiple R-squared: 0.9979, Adjusted R-squared: 0.9979

F-statistic: 7.5e+04 on 38 and 5914 DF, p-value: < 2.2e-16

Step 12. AIC/BIC model comparison

```
AIC(
  mod_taxo,
  mod_dendro,
  mod_full
)
```

```
      df      AIC
mod_taxo  38 21950.74
mod_dendro  4 -14739.28
mod_full  40 -14723.94
```

```
BIC(
  mod_taxo,
  mod_dendro,
  mod_full
)
```

```
      df      BIC
mod_taxo  38 22205.02
mod_dendro  4 -14712.52
mod_full  40 -14456.28
```

mod_taxo and mod_dendro are not nested models. They explain the response variable using different, non-overlapping sets of predictors.

```
### Summary table
compare_performance(
  mod_taxo,
  mod_dendro,
  mod_full,
  rank = TRUE
)
```

Comparison of Model Performance Indices

Name	Model	R2	R2 (adj.)	RMSE	Sigma	AIC weights
mod_dendro	lm	0.998	0.998	0.070	0.070	1.000
mod_full	lm	0.998	0.998	0.070	0.070	4.66e-04
mod_taxo	lm	0.018	0.012	1.519	1.524	0.00e+00

Name	AICc weights	BIC weights	Performance-Score
mod_dendro	1.000	1.00	100.00%
mod_full	3.54e-04	2.28e-56	57.15%
mod_taxo	0.00e+00	0.00e+00	0.00%

Step 13. Model validation

Residuals

```
par(mfrow=c(2,2))
plot(mod_full)
```

Warning: not plotting observations with leverage one:
1, 11, 631, 2424, 3580, 4262

Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

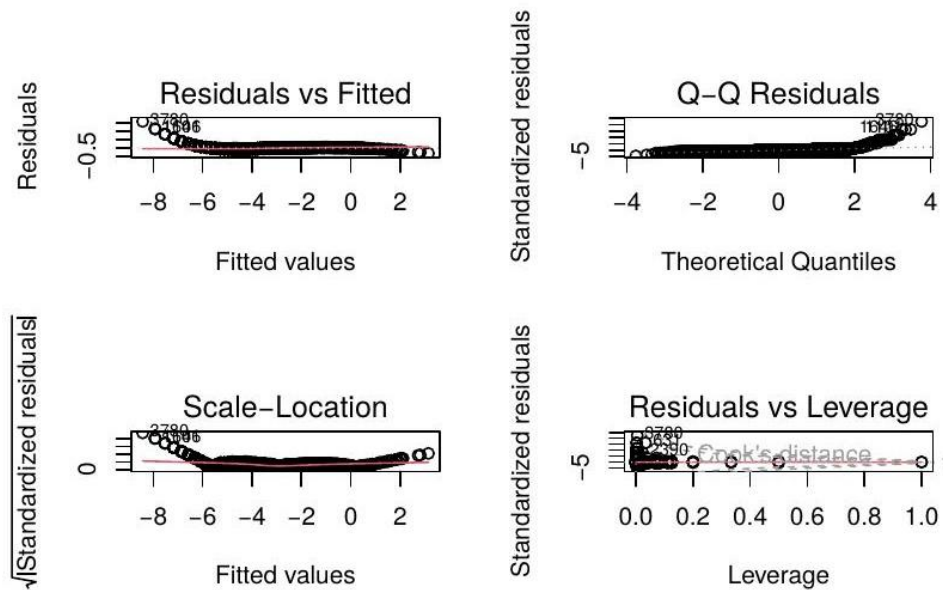


Figure 14. Residual diagnostic plots for `mod_full`: residuals vs fitted, Q-Q residuals, scale-location, and residuals vs leverage.

Normality

```
shapiro.test(
  residuals(mod_full)[1:5000]
)
```

Shapiro-Wilk normality test

```
data:  residuals(mod_full)[1:5000]
W = 0.50776, p-value < 2.2e-16
```

Homoscedasticity

```
bptest(mod_full)
```

studentized Breusch-Pagan test

```
data:  mod_full
BP = 261.58, df = 38, p-value < 2.2e-16
```

Step 14. Interpretation, revisions, and description of limitations

Implications for the research question

The results unambiguously confirm that it is the dendrometric characteristics, and not species diversity, that determine the rate of carbon sequestered in PNM. The purely dendrometric model (mod_dendro) explains 99.8% of the variance in carbon ($R^2 = 0.998$) with only two predictors — log_DBH and log_Height — and has the best AIC (−14,739). In contrast, the taxonomic model (mod_taxo) explains only 1.8% of this variance ($R^2 = 0.018$), confirming that the family or genus of a species is not a good predictor of stored carbon. Adding taxonomic variables in mod_full does not significantly improve performance relative to mod_dendro alone ($\Delta AIC \approx 15$ in favour of mod_dendro). DBH emerges as the dominant predictor ($\beta = 2.74$, $p < 2 \times 10^{-16}$), while height does not contribute significantly ($p = 0.87$). These results therefore validate the initial hypothesis: the larger a tree's diameter, the more carbon it sequesters, independently of its taxonomic identity.

Areas for improvement

A mixed model (lme4) with species or transect as a random effect would allow the hierarchical structure of the data to be taken into account (individuals nested within species, themselves nested within transects). A model robust to assumption violations (quantile regression, a GLS model with a variance structure) would address the heteroscedasticity. Removing the derived variables (Basal_area_G, Wood_vol_m3) from the predictor set would eliminate the structural multicollinearity.

Conclusions

1. DBH is the near-exclusive predictor of carbon. The log-log relationship between DBH and carbon is near-perfect, which is mechanically explained: the allometric equation of Chave et al. (2005) is itself built from DBH. The model therefore recovers the structure of the equation.
2. Taxonomy has a statistically significant but ecologically negligible effect. The ANOVAs show differences between families ($p = 5.3 \times 10^{-5}$, $\eta^2 = 0.008$) and genera ($p = 2.1 \times 10^{-9}$, $\eta^2 = 0.02$), but these effect sizes are minimal — taxonomy explains only 1–2% of the variance in carbon.
3. Stand structure prevails over floristic composition. For the management of PNM, this means it is preferable to protect large trees regardless of species, rather than targeting particular species.