Proceedings of the  
  
Workshop on Data analysis with R for Lao PDR’s fourth National Forest Inventory cycle

01-05 Sep 2025, FIPD office, Vientiane, Lao PDR

Table of contents

# Context

A new training course: “**Training workshop on Data analysis with R for Lao PDR’s fourth National Forest Inventory cycle**”, was organised for FIPD from 01 to 05 September 2025, in FIPD office in Vientiane (Lao PDR), in order to continue capacity building on national forest inventory data analysis and to present the calculation chain for the cycle 4 updated method.

The training objective was to:

1. Continue build capacities of FIPD in forest data analysis with R and Rstudio.
2. Present the code base used for analyzing the 4th National Forest Inventory of Lao PDR, implemented in 2024 and 2025.
3. Explore opportunities to build a R community with research institutions and the civil society to help each other build analytical skills and deal with complex data analysis in the forest, ecology and environment sectors.

The workshop agenda and participants list are shown in [Figure 1](#fig-agenda1).

This document contains all the code presented and the exercises with solutions during the training.

# Initial Setup

Before going to loading data and perform the analysis, we need to prepare the environment by loading packages and setting a few options. Here we don’t load directly packages, but we source a script that load them.

## Useful when sourcing script from Quarto doc.  
if (!require(here)) install.packages("here"); library(here)  
  
source(here("R/00-load-packages.R"))  
source(here("R/fct-nfi-aggregate3-new.R"))

# 1. Session 01: Objectives

Session 1 presented the detailed objectives of the training, as a mixture of basic(refresher) and advanced analysis in R. The technical analysis objectives were to cover: Basic calculations, Aggregation with simple sampling, Ratio estimators, and Double sampling for post-stratification.

The institutional objective was to explore a pool of resources available in other institutions to improve analytical capacities of FIPD. This includes academia and research institutions in Lao PDR to bring them together for broader technical exchange in R and exploring their potential support to FIPD.

Finally, the workshop aimed at defining a work plan of continuous training to bring in additional pools and calculations.

To achieve the above objectives, the training focused on:

1. Read data into R and visualize trees characteristics
2. Manipulate tables (merge, split, rename)
3. Detect outliers and correct the data
4. Aggregate from tree to plot level
5. Aggregate plot level data to LC types
6. Aggregate to sub-population and national level

# 2. Session 02: Updated calculation chain for the NFI cycle 4 carbon analysis

The NFI cycle 4 of LAO PDR had an updated sampling methodology: two-phases sampling (or double sampling) for stratification with ratio estimators. The plot design also slightly changed, the subplots were positioned 60 m apart center to center on a L shape.

## 2.1 Sampling design

The updated sampling design consisted of two sampling frames and sizes:

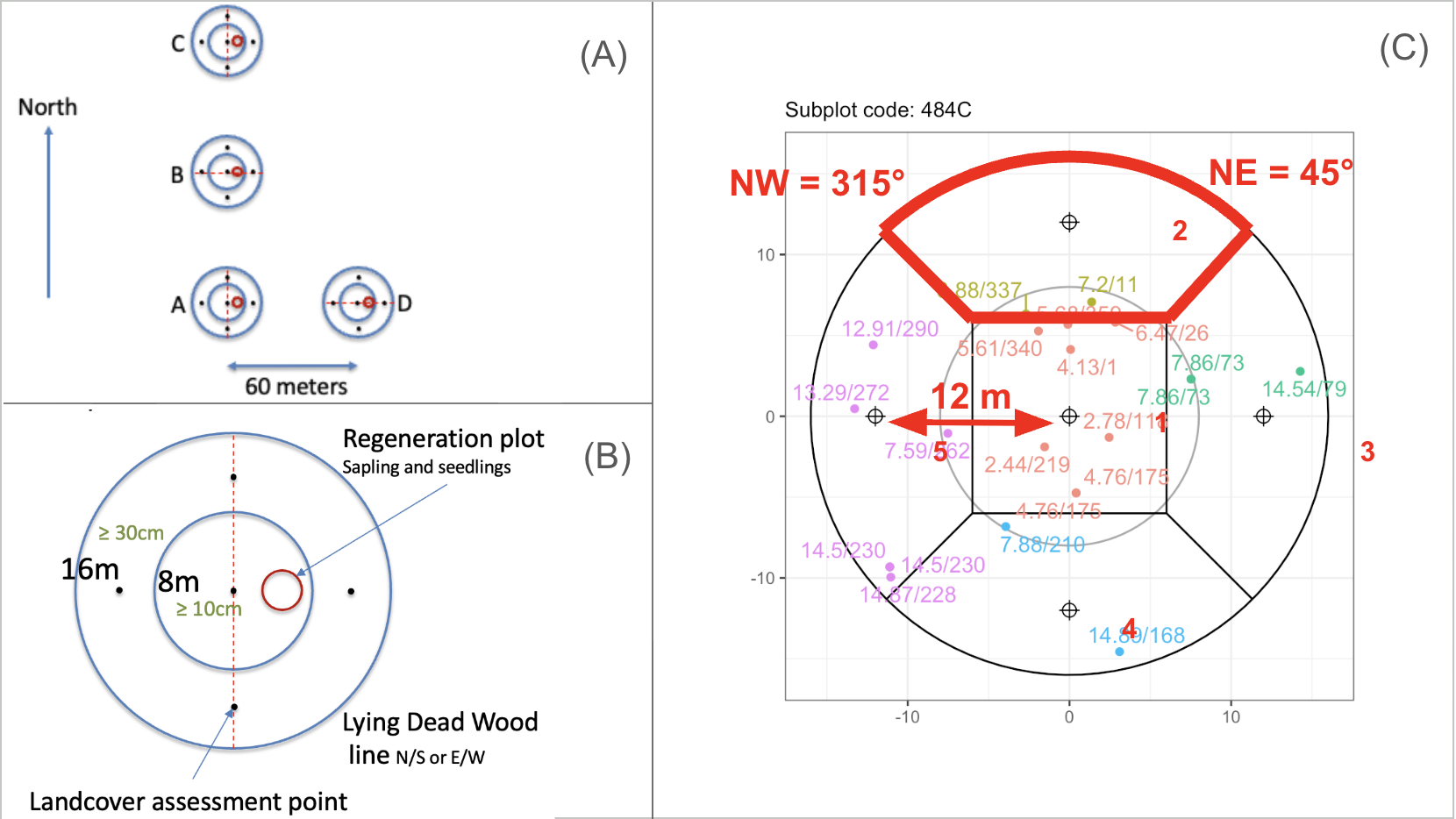
* Phase 1: a dense 6 x 6 km grid, on which all grid points were visually interpreted to identify the land cover class and stratum.
* Phase 2: an initial systematic subset of the phase 1, a 12 x 18 km grid, for field survey. Then not all plots on the second grid could be visited due to budget constraints, so around half of grid point were visited.

Important advantages of this design:

* The final estimates takes into account both phases when producing means and totals.
* Mis-classifications in phase 1 are corrected with phase 2 plots.
* **2-phases designs produce both areas and forest characteristics, which makes final estimates generally more robust than NFIs with one sampling and independent land cover maps (**[Westfall et al. 2019](https://www.fs.usda.gov/nrs/pubs/jrnl/2019/nrs_2019_westfall_001.pdf)**).**

## 2.2 Plot design

The plot design was also updated from the NFI previous cycles, where the subplots where placed randomly within a designated area. In the cycle 4 design, plots are composed by 4 circular subplots in a L shape, with 60 m between subplot centers. Subplots are nested circles of 8 and 16 m radius for trees between 10 and 30 cm DBH (30 excluded) and trees bigger or equal to 30 cm DBH respectively. Seedling and saplings are counted in a 2m radius circle positioned 6 m east of the subplot centers and lying deadwood are measured on a line transect of 32 m.



Plot design of Lao PDR NFI cycle 4. (A) Plot design, (B) subplot design and (C) land cover observation points and land cover sections.

A new specificity of the NFI cycle 4 was to observe land cover on 5 points located at the subplot centers and 12 m north, east, south and west from the center. Following equidistance, or Thiessen polygons, land cover sections (LCS) were defined as areas of:

* 12 m side square at the center,
* a quarter of the difference between the 16 m radius outer subplot circle and the 12 m side square for the other LCSs. The diagonals (NW-SE and SW-NE) were used to delineate the quarters.

## 2.3 Core measurements

The NFI cycle 4 primary remained Carbon stock estimation but it expanded to multi-purpose with additional measurements of tree location for time series and non-timber forest products (NTFPs). The full list of environmental, trees and other measurements can be found in the field manual. Here is a summary of core measurements used for the carbon stock analysis:

Phase 1:

* Administrative zone
* Land cover at plot center (subplot A) and stratum (natural forest, planted forest, regenerating vegetation, non-forest)

Subplot:

* Land cover at 5 points

Tree:

* Tree distance and azimuth from subplot center
* Tree DBH
* Tree species

Standing deadwood:

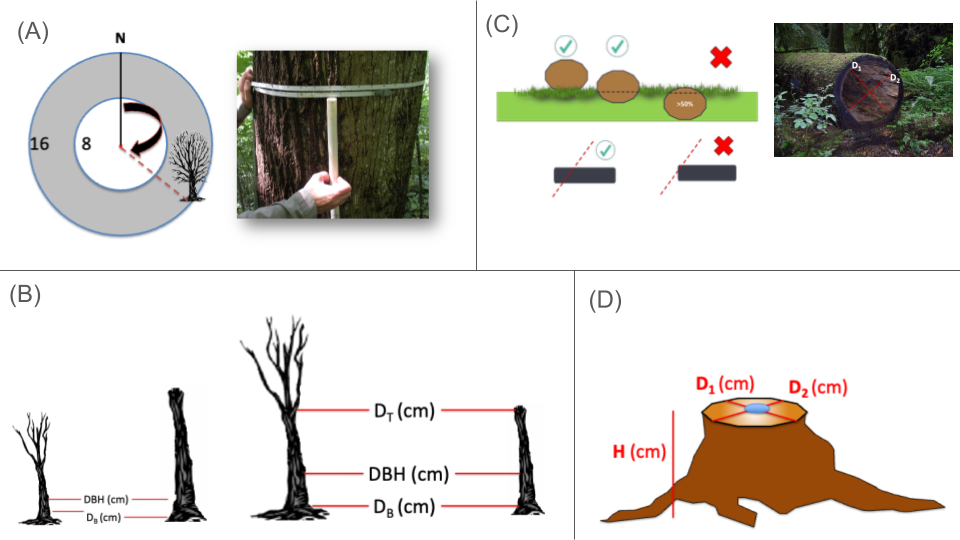
* Diameter at base, DBH and diameter at top (for short standing deadwood).

Stump:

* Stump height
* Stump diameters (2 perpendicular measurements)

Lying deadwood:

* Section length
* inner and outer ring in case of hollow stem.



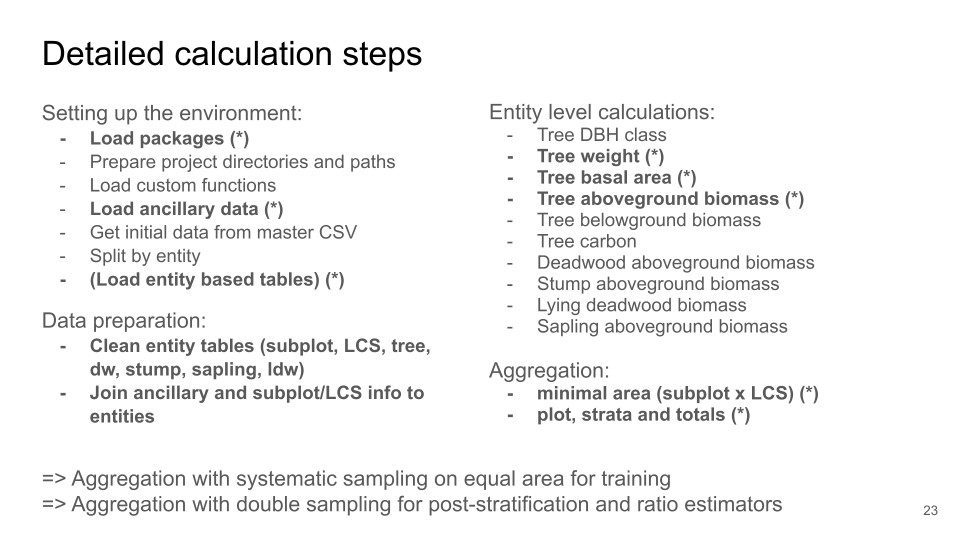
Core measurements for carbon. (A) Live tree measurements, (B) standing dead tree measurement, (C) Lying deadwood measurement, (D) stump measurement.

## 2.4 Analysis steps

The data analysis was performed with R and R studio and structured around 4 main steps:

1. **Setup** the R environment (inc. loading NFI and ancillary data)
2. Prepare **clean** data (clean, correct and join)
3. **Calculate entity** level variables
4. **Aggregate** measurements to minimal measurement area, plot, strata and totals

Each of the main steps was sub-divided into the following actions, noting that actions in bold with the symbol (\*) were the focus of this training:

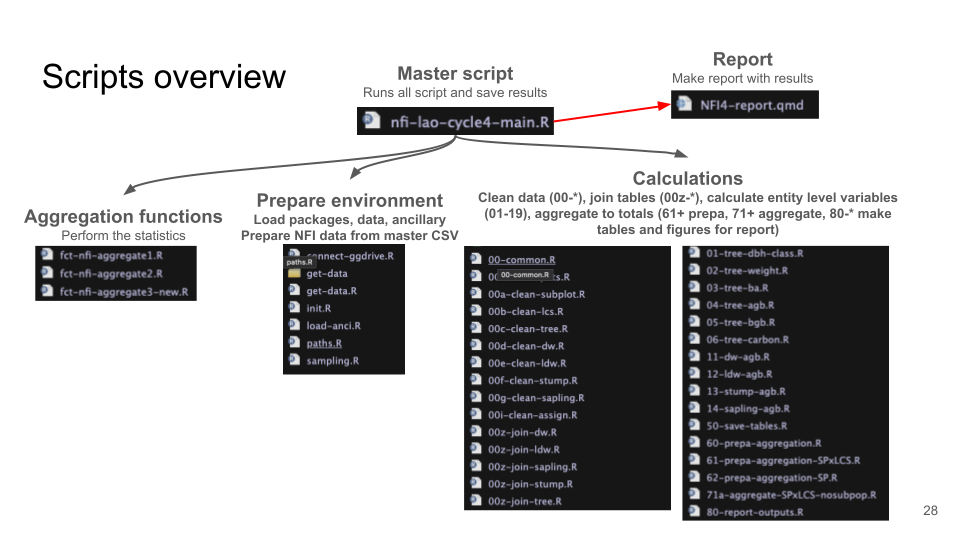


Detailed calculation steps.

The initial input data from the field work was stored into an ONA server and downloaded manually as a master CSV. Splitting the master CSV and renaming columns was relatively complex and not the focus of this training. Instead “semi”-cleaned tables **tree, subplot** and **lcs** tables were prepared for the participants.

Training practice was also limited to live trees aboveground biomass, as participants should be more familiar with this this carbon pool than deadwood, stump, sapling and lying deadwood.

An overview on how the detailed actions were handled in specific R scripts was also presented:



List of analytical R scripts.

## 2.5 Core results

Core results were shown from the result XLSX file that contained the following tables:

* **Phase 1 data**: strata, subpopulations
* **Phase 2 subplot level**: subplot x LCS values for core variables (tree AGB, tree BGB, sapling AGB, DW, stump AGB, LDW, Ctotal)
* **Plot summary**: aggregate subplots to partial plots
* **Plot unique**: for each plot with mixed LC keeps the section with max AGB, if equal, keeps the smallest LC code
* **Subpop x stratum**:  aggregate variables for each land cover (i.e domain d)
* **Subpop**: aggregate strata
* **Totals**: aggregate sub-populations

The initial result file was prepared with ERPA vs non-ERPA zone as subpopulations but the final estimates required a few corrections (see final report):

* Correction of phase 1 land cover class passed to phase 2 plots. Initially based on FTM2022, these need to be back to Phase 1 land cover.
* Correction Factor was missing in EG and DD allometric equations.
* Subplot level AGB for root-to-shoot ratio was using wrong weights,
* ERPA subpopulations influenced slightly the final estimates vs a scenario with no subpopulation.

# 3. Session 03: Discussion and remarks on the updated calculations

There were no questions on the calculation chain at the time of the presentation.

# 4. Session 04: Refresher on R for forest data analysis

TBD

# 5. Session 05: data preparation

**Goals**

1. Load initial tables and visualize core data
2. Correct errors in subplot table (duplicates in subplot\_id)

## 5.1 Load harmonized tables and ancillary data

### 5.1.1 Load subplot, lcs, tree tables from NFI

We can read tables or CSV files from the computer with the function read\_csv(). We can store the initial table in objects with the suffix ‘\_init’ to keep the initial version in the environment.

Example: read the table ‘training\_subplot.csv’ from the computer into the object subplot\_init:

subplot\_init <- read\_csv("data/training\_subplot.csv", show\_col\_types = FALSE)

Load the tables ‘training\_tree.csv’ and ‘training\_lcs.csv’ in the [Exercise 1](#exr-51).

|  |
| --- |
| **Exercise 1 (~ Load ‘training\_tree.csv’ and ‘training\_lcs.csv’ tables)**   * Load the table ‘training\_lcs.csv’ in the R objects ‘lcs\_init’. * Load the table ‘training\_tree.csv’ in the R objects ‘tree\_init’. |

Answer to [Exercise 1](#exr-51):

## !!! Solution  
lcs\_init <- read\_csv("data/training\_lcs.csv", show\_col\_types = FALSE)  
tree\_init <- read\_csv("data/training\_tree.csv", show\_col\_types = FALSE)

## 5.2 Load ancillary data

**Ancillary data are not collected in the field but useful for the analysis of field data.**

In the training we use plot level environment factor (E) from Chave et al. 2014 and the NFI Phase 1 plot data. Ancillary data are loaded into a list to keep the environment tidy.

anci <- list()  
anci$ph1 <- read\_csv("data/training\_anci\_phase1.csv", show\_col\_types = FALSE)  
anci$plot\_E <- read\_csv("data/training\_anci\_plotE.csv", show\_col\_types = FALSE)

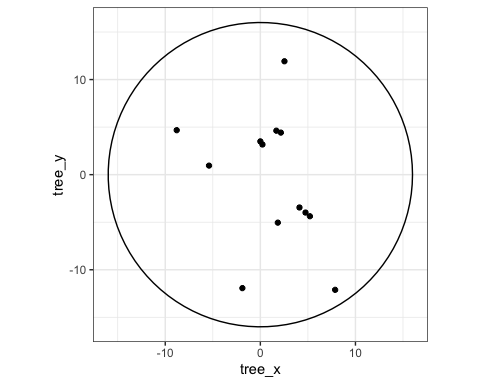
## 5.3 Visualize tree locations

We visualize tree locations from the tree\_init table with ggplot() and geom\_point() to see if any potential error.

Step-by-step for one subplot:

* Define circ16 as tibble of 100 rows to draw a circular plot in ggplot, based on a radius and an angle “theta” in radian.
* Draw a ggplot with trees from the table tree\_init in plot 631 and subplot “C”, using filter(),
* Use tree\_x and tree\_y to show trees based on their location from the subplot center,
* Add plot boundary with circ16.

circ16 <- tibble(  
 theta = seq(0, 2\*pi, length = 100),  
 x = 16 \* cos(theta),  
 y = 16 \* sin(theta)  
)  
  
tree\_init |>  
 filter(plot\_no == 631, subplot\_no == "C") |>  
 ggplot(aes(x = tree\_x, y = tree\_y)) +  
 geom\_point() +  
 geom\_path(data = circ16, aes(x = x, y = y)) +  
 coord\_fixed()

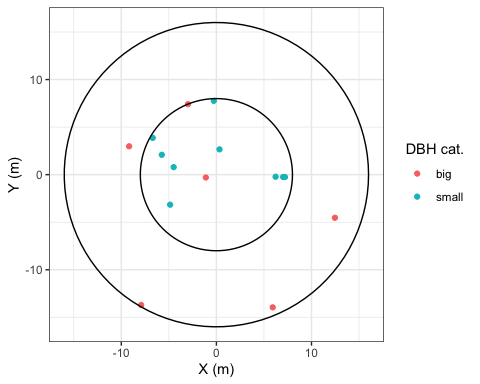


Adapt the previous code to make a tree location map for the subplot ‘1A’ and add different colors based on tree DBH size in [Exercise 2](#exr-52).

|  |
| --- |
| **Exercise 2 (~ Make a tree location figure)**   * create the table circ08 similar to circ16 but for a 8m radius circle. * Filter within tree the trees from subplot ‘1A’ with filter(). * Use mutate() and if\_else() to add a column tree\_dbh\_cat that has value “small” for trees with DBH < 30 cm and “big” for the other trees. * Make the figure of tree location with the 8m and 16m radius circles and change the color of trees based on tree\_dbh\_cat. * Use labs() to create clean labels in English or Lao. |

Answer to [Exercise 2](#exr-52):

## !!! Solution  
circ08 <- tibble(  
 theta = seq(0, 2\*pi, length = 100),  
 x = 8 \* cos(theta),  
 y = 8 \* sin(theta)  
)  
  
tree\_init |>   
 filter(plot\_no == 1, subplot\_no == "A") |>  
 mutate(tree\_dbh\_cat = if\_else(tree\_dbh < 30, "small", "big")) |>  
 ggplot(aes(x = tree\_x, y = tree\_y)) +  
 geom\_point(aes(color = tree\_dbh\_cat)) +  
 geom\_path(data = circ08, aes(x = x, y = y)) +  
 geom\_path(data = circ16, aes(x = x, y = y)) +  
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)", color = "DBH cat.")

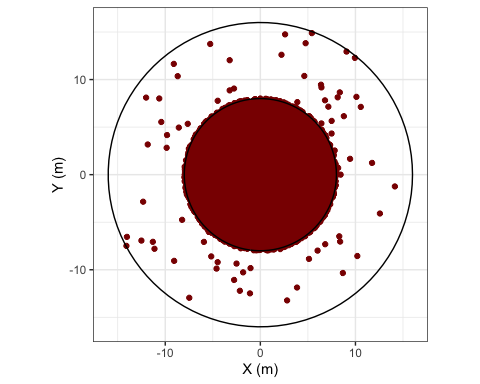


Visualizations can also be used to detect outliers. We use the example of small trees outside the nested subplot circle of 8m radius as guided exercise in [Exercise 3](#exr-53).

|  |
| --- |
| **Exercise 3 (~ (Optional) detect outliers on graph)** **Part 1: Find outliers**   * Make a figure with tree location of all trees smaller than or equal to 30 cm with the 2 circles of 8 and 16 meters. * Use filter() to get all trees with DBH <= 30 cm. * Make the ggplot of tree locations   **Part 2: highlight outliers**  We can highlight outliers by making a specific table for them.   * Create the table outliers containing trees with DBH <= 30 and distance > 8. Use filter(). * Make the same figure as before and add a geom\_point() with data = outliers, shape = 23 and size = 4 .   TIP: here we added alpha = 0.2 to the geometry of the trees. It make tree point semi-transparent and helps to detect potential measurement issues if we would detect patches with few trees.  **Part 3: explanations**  Now are these outliers measurement errors?   * Make the same graph as this exercise’s part 1, but this time with tree DBH < 30 and **not** DBH <= 30. |

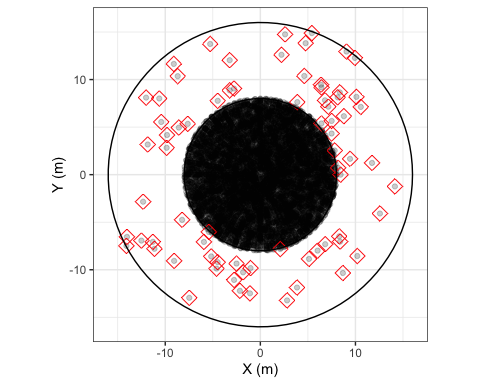
Type here answers to [Exercise 3](#exr-53) part1:

## !!! Solution  
tree\_init |>   
 filter(tree\_dbh <= 30) |>  
 ggplot(aes(x = tree\_x, y = tree\_y)) +  
 geom\_point(col = "darkred") +  
 geom\_path(data = circ08, aes(x = x, y = y)) +  
 geom\_path(data = circ16, aes(x = x, y = y)) +  
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)")



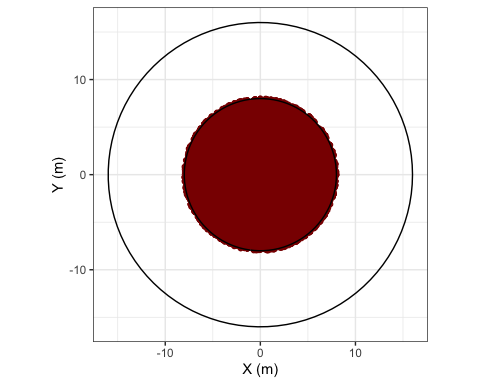
Type here answers to [Exercise 3](#exr-53) part2:

## !!! Solution  
outliers <- tree\_init |> filter(tree\_dbh <= 30, tree\_distance > 8)  
  
tree\_init |>   
 filter(tree\_dbh <= 30) |>  
 ggplot(aes(x = tree\_x, y = tree\_y)) +  
 geom\_point(alpha = 0.2) +  
 geom\_point(data = outliers, col = "red", shape = 23, size = 4) +  
 geom\_path(data = circ08, aes(x = x, y = y)) +  
 geom\_path(data = circ16, aes(x = x, y = y)) +  
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)")



Type here answers to [Exercise 3](#exr-53) part3:

## !!! Solution  
tree\_init |>   
 filter(tree\_dbh < 30) |>  
 ggplot(aes(x = tree\_x, y = tree\_y)) +  
 geom\_point(col = "darkred") +  
 geom\_path(data = circ08, aes(x = x, y = y)) +  
 geom\_path(data = circ16, aes(x = x, y = y)) +  
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)")



Conclusion of [Exercise 3](#exr-53): the trees found outside the small circle were trees with DBH = 30 cm. They were not measurement errors, they were part of the big trees and therefore could be found outside the 8m radius circle.

## 5.4 Find and correct errors

One of the main issue found in the raw data was the entry errors of subplot and tree IDs. It is very important in Database Management Systems (DBMS) to have unique IDs for all tables. In NFI it is particularly important for joining information from one table to the other. For example, land cover information is recorded at subplot level and needs to be passed on to tree level to apply the correct allometric equations. These errors can be detected programatically but often have to be solved manually.

This section introduces how to use group\_by() and summarise() to find duplicates in subplot IDs.

### 5.4.1 Identify duplicates in subplot IDs

Here we want to check within subplots if there are duplicates in subplot IDs (subplot\_id) and plot numbers (plot\_no) in the subplot\_init table. The objective is to create a vector of subplot Ids that are duplicated. If there are no duplicate the vector will be of length 0.

step-by-step:

* Create a vector vec\_dup and assign it the following code sequence.
* Group the subplot table by subplot ID with group\_by() and use summarise() to count the number of subplots that have each ID inside a new column count. Use the function n() to count.
* Filter all the subplot IDs that have a count bigger than 1 with filter().
* Pull the subplot IDs.

vec\_dup <- subplot\_init |>  
 group\_by(subplot\_id) |>  
 summarise(count = n(), .groups = "drop") |>  
 filter(count > 1) |>  
 pull(subplot\_id)  
vec\_dup

[1] "553D" "631C" "632C"

Notice the results as 3 subplots have duplicate subplot\_id.

Now practice the same code sequence to check if there are duplicates in plot\_no in the [Exercise 4](#exr-54). Note that each plot ID should appears 4 times in the subplot table, one for each subplot, so duplicates are when count > 4.

|  |
| --- |
| **Exercise 4 (~ Identify duplicates of plot number)**   * Create vec\_dup2 similarly to vec\_dup, but this time looking for plot\_no that have more than 4 subplot IDs, so you will need to group by plot\_no and filter count > 4 and pull plot\_no.   There should be no duplicates of plot\_no. |

Answer to [Exercise 4](#exr-54):

## !!! Solution  
vec\_dup2 <- subplot\_init |>  
 group\_by(plot\_no) |>  
 summarise(count = n(), .groups = "drop") |>  
 filter(count > 4) |>  
 pull(plot\_no)  
vec\_dup2

numeric(0)

There should be no duplicates of plot\_no.

### 5.4.2 Correct the subplot ID issues

To correct the duplicates in subplot\_id, we filter the subplots from one of the duplicated IDs and visually check what is wrong. In case one ID is missing and one is duplicated we can use time stamps to identify which of the duplicates should be renamed.

Then we can use mutate() and case\_when() to apply correction in R.

Step-by-step:

* Filter the subplots of one plot with duplicates: 631C.
* Visualize the table (run View(tt) or click tt in the Environment tab (top-right window in Rstudio).
  + In the case of ‘631C’ there are 2 subplots C but no B. we can use time stamps to identify that ONA\_index 109 should actually be for subplot B.
* Perform the correction using case\_when() function for subplot\_id 631C. At this stage only ONA\_index is unique so it can be used to correctly identify the subplot to correct.

tt <- subplot\_init |> filter(plot\_no == 631)  
  
# View(tt)  
  
subplot <- subplot\_init |>  
 mutate(  
 subplot\_no = case\_when(  
 subplot\_id == "631C" & ONA\_index == 109 ~ "B",  
 ## ADD more corrections,  
 TRUE ~ subplot\_no  
 )  
 )

You can now copy/paste the above R chunk and complete the correction for the 2 other subplot IDs with duplicates in [Exercise 5](#exr-55).

Note: we correct subplot\_no now, but we will need to remake subplot\_id once subplot\_no is correct.

|  |
| --- |
| **Exercise 5 (~ Correct duplicates in subplot IDs)**   * for each remaining duplicate, run the table tt with all subplots for the plot with duplicates * Visualize tt and identify ONA\_index of the subplot that should be corrected and what is the correct subplot\_no. * Copyu/paste the subplot <- ... sequence from above and fill in the case\_when() with all the correction for the 3 duplicated subplots. |

Solution to [Exercise 5](#exr-55):

## !!! Solution  
subplot <- subplot\_init |>  
 mutate(  
 subplot\_no = case\_when(  
 subplot\_id == "631C" & ONA\_index == 109 ~ "B",  
 subplot\_id == "632C" & ONA\_index == 113 ~ "B",  
 subplot\_id == "553D" & ONA\_index == 265 ~ "C",  
 TRUE ~ subplot\_no  
 )  
 )

Now we can recreate subplot\_id from the corrected subplot\_no. Since subplot\_id is a text field, we need to make it consistent for number of characters to ensure correct sorting of the rows. For this purpose we include texts “0” and “00” to make the number of characters consistent.

subplot <- subplot |>  
 mutate(  
 subplot\_id = case\_when(  
 plot\_no < 10 ~ paste0("00", plot\_no, subplot\_no),  
 plot\_no < 100 ~ paste0("0", plot\_no, subplot\_no),  
 TRUE ~ paste0(plot\_no, subplot\_no)  
 )  
 )

Now if we re-run the vec\_dup sequence with subplot instead of subplot\_init, vec\_dup should be of length 0.

vec\_dup <- subplot |>  
 group\_by(subplot\_id) |>  
 summarise(count = n(), .groups = "drop") |>  
 filter(count > 1) |>  
 pull(subplot\_id)  
vec\_dup

character(0)

### 5.4.3 Check outliers in tree table

The tree table we can check for outliers in numeric columns with the summary() function.

For example with DBH:

summary(tree\_init$tree\_dbh)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 10.00 13.50 19.80 25.79 34.20 199.00

There are no trees with DBH < 10 cm or unrealistically big DBH, all good.

We can check for outliers in tree distance and azimuth in [Exercise 6](#exr-56).

|  |
| --- |
| **Exercise 6 (~ Check azimuth and distance)**   * Check that azimuth is between 0 and 360. * Check that distance is between 0 and 16. |

Solution to [Exercise 6](#exr-56):

summary(tree\_init$tree\_azimuth)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0 90 179 179 268 360

summary(tree\_init$tree\_distance)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.000 4.335 6.280 6.785 7.900 16.000

## 5.5 Assigning clean data to objects

After resolving all errors, save the tables to the entity name without “\_init” suffix. This way we have both the initial and corrected version in the environment. If tehre were no errors the inital tables can be passed directly to the corrected tables:

tree <- tree\_init  
lcs <- lcs\_init

Finally we can remove the temporary objects that won’t be used later.

rm(vec\_dup, vec\_dup2, tt, outliers)

**NOTE: use of temporary object list tmp.**

In the main analysis the temporary objects are often stored in a list called tmp. This avoids cluttering the environment with temporary objects and makes it easy to remove them with rm(tmp).

## Create empty list  
tmp <- list()  
  
## Populate the list (example)  
tmp$outliers <- tree\_init |> filter(tree\_dbh <= 30, tree\_distance > 8)  
  
## Remove temporary objects  
rm(tmp)

# 6. Session 06: data joins

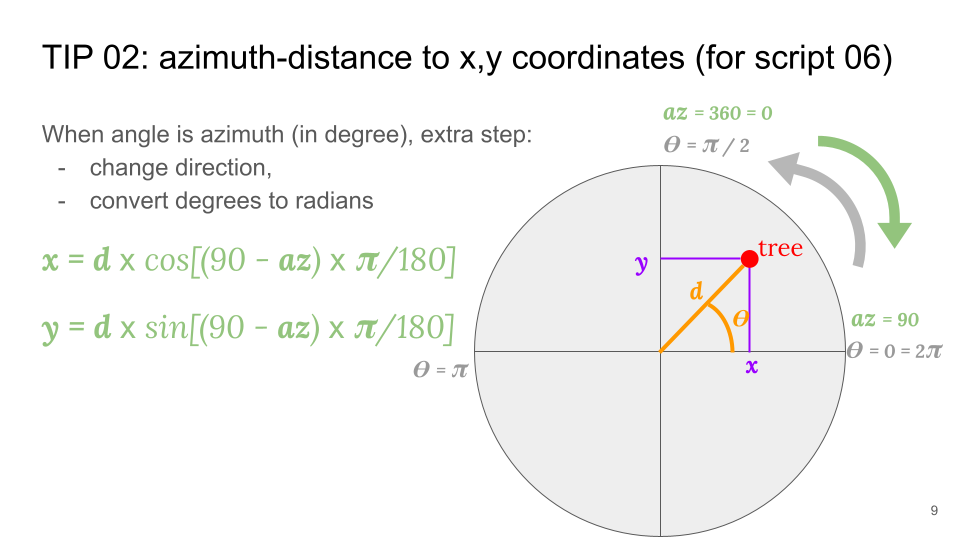
**Goals:**

1. Assign LCS number to trees
2. Join land cover and Chave E to trees

## 6.1 Assign LCS number to trees

Since the Land Cover Sections (LCS) are a mix of center square and outside disc portions, we need both tree position as distance and azimuth and as coordinates .

See the classroom presentation: “main calculations, TIP02” for explanations on converting distance and angle to coordinates for the subplot center, and converting distance and azimuth to coordinates .



Conversion of distance/azimuth to x, y

The R implementation is as follows:

tree <- tree |>   
 mutate(   
 tree\_x = cos((90 - tree\_azimuth) \* pi/180) \* tree\_distance,   
 tree\_y = sin((90 - tree\_azimuth) \* pi/180) \* tree\_distance,   
 )

Once we have tree coordinates, we can assign LCS with case\_when(), identifying trees in the center square first (with their coordinates), then the outer disc portion based on azimuth:

* To be inside the **center square**, trees must have their coordinates both between -6 and 6 meters, or their absolute value smaller or equal to 6:

$$
-6 \le x \le 6 \Leftrightarrow |x| \le 6 \\ -6 \le y \le 6 \Leftrightarrow |y| \le 6
$$

* To be in one of the **outer disc portions**, trees must not be in the center square and be respectively positioned between NW and NE lines, NE and SE, SE and SW or SW and NW for the north, east, south and west disc portions. These lines’ azimuth are 315, 45, 135, 215 degrees respectively for NW, NE, SE and SW lines.

R implementation:

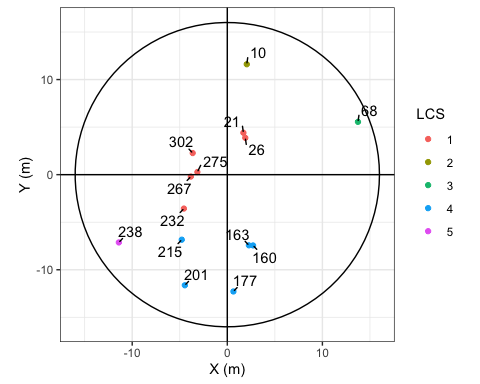
tree <- tree |>  
 mutate(  
 lcs\_no = case\_when(  
 abs(tree\_x) <= 6 & abs(tree\_y) <= 6 ~ 1,  
 tree\_azimuth > 315 | tree\_azimuth <=45 ~ 2,  
 tree\_azimuth > 45 & tree\_azimuth <=135 ~ 3,  
 tree\_azimuth > 135 & tree\_azimuth <=225 ~ 4,  
 tree\_azimuth > 225 & tree\_azimuth <=315 ~ 5,  
 TRUE ~ NA\_integer\_  
 )  
 )

We can now check if the code implementation worked with a visual check in [Exercise 7](#exr-61).

|  |
| --- |
| **Exercise 7 (~ Figure of tree locations)**   * Filter the tree data for subplot ‘**123B**’. * Make a base figure with geom\_point(). * Add subplot boundaries with geom\_path() and circ16. * **Add cardinal lines with geom\_hline(yintercept = 0) and geom\_vline(xintercept = 0).** * **Add azimuth to each tree as labels with geom\_label\_repel()**. * Add coord\_fixed() to make sure the x and y axis have the same unit length.   Check visually that the trees are positioned correctly based on their azimuth! |

Answer to [Exercise 7](#exr-61):

## !!! Solution  
tree |>  
 filter(plot\_no == 123, subplot\_no == "B") |>  
 ggplot(aes(x = tree\_x, y = tree\_y)) +  
 geom\_point(aes(color = as.character(lcs\_no))) +  
 geom\_path(data = circ16, aes(x = x, y = y)) +  
 geom\_hline(yintercept = 0) +  
 geom\_vline(xintercept = 0) +  
 geom\_text\_repel(aes(label = tree\_azimuth), min.segment.length = 0) +  
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)", color = "LCS")



We can observe that azimuths have been well converted to since azimuth labels are in their correct quadrants (0-90, 90-180, 180-270 ad 270-360 degree respectively).

## 6.2 Join tables

### 6.2.1 Prepare tables to join

We want to join 2 tables to the tree data:

* Chave et al. 2014 environment factor from the table anci$plot\_E,
* Land cover code from the table lcs,

Sometimes we don’t want all the columns from a tables to be passed on to tree. In this case we create a temporary tables with only the **key(s)** = the matching column(s) between two tables to join, and the information to pass.

For example, we don’t want lcs\_name from lcs to go to trees, we already have lcs\_no as key between the two tables. We can either remove the undesired column or keep all other columns. In this case, removing is faster, but for sustainability, it is preferable to write the columns you want to keep so that when you re-open this document in a few weeks you see directly what is kept.

tmp\_lcs <- lcs |> select(-lcs\_name)  
  
## Same but preferable:  
tmp\_lcs <- lcs |> select(plot\_no, subplot\_no, lcs\_no, lc\_no, lc\_code)

### 6.2.2 Make the joins

* **With changing object**

We join LCS (tmp\_lcs) with the tree table using their common keys. In this case there are three keys: plot\_no, subplot\_no and lcs\_no.

tree\_join <- tree |>   
 left\_join(tmp\_lcs, by = join\_by(plot\_no, subplot\_no, lcs\_no))

It is recommended to change object name when using join to avoid undesired suffixes. Undesired suffixes happen when we join the same tables twice or more by mistake. The second time we join two tables, since the new tables are already in the joined table, they will be passed again but this time with suffixes .x and .y. Observe undesired suffix in the [Exercise 8](#exr-62).

|  |
| --- |
| **Exercise 8 (~ Undesired suffixes)**   * Create tree\_err by joining tree and tmp\_lcs * Re-create tree\_err by joining tree\_err and tmp\_lcs * Run names(tree\_err) |

Answers to [Exercise 8](#exr-62):

## !!! Solution  
tree\_err <- tree |>   
 left\_join(tmp\_lcs, by = join\_by(plot\_no, subplot\_no, lcs\_no))  
tree\_err <- tree\_err |>   
 left\_join(tmp\_lcs, by = join\_by(plot\_no, subplot\_no, lcs\_no))  
names(tree\_err)

[1] "plot\_no" "subplot\_no" "lcs\_no"   
 [4] "tree\_no" "tree\_stem\_no" "tree\_dbh"   
 [7] "tree\_species\_code" "tree\_species\_binomial" "tree\_distance"   
[10] "tree\_azimuth" "tree\_x" "tree\_y"   
[13] "lc\_no.x" "lc\_code.x" "lc\_no.y"   
[16] "lc\_code.y"

* **With keeping the same object name**

Often time we don’t want to change object names for a join if we have a long sequence of modifications ahead. We can join and keep the same object names, but in this case we have to **control the suffixes**.

Step-by-step:

1. Identify the **new column** passed from the table to join.
2. **Create these columns in the receiving table** with NAs using mutate(),
3. **Join the tables and add suffixes** “\_rm” to the receiving table with NAs and “” to the joined columns,
4. **Remove the columns** that end with “\_rm” (since they contains NAs).

## 1. Columns to be passed are 'lc\_no' and 'lc\_code'  
tree <- tree |>  
 ## 2. Create these columns with NAs  
 mutate(lc\_no = NA, lc\_code = NA) |>   
 ## 3. join with controlled suffix names  
 left\_join(tmp\_lcs, by = join\_by(plot\_no, subplot\_no, lcs\_no), suffix = c("\_rm", "")) |>   
 ## 4. remove initial columns with NAs  
 select(-ends\_with("\_rm"))

Practice ‘joins with controlled suffix’ to join Chave E at plot level from anci$plot\_E with the [Exercise 9](#exr-63).

|  |
| --- |
| **Exercise 9 (~ Join Chave E to the trees table)**   * Identify which column from anci$plot\_E to pass to tree and which column is the key * from tree, create a column with NAs that has the same name as the column from anci$plot\_E that we want in tree * Join the 2 tables and control the suffixes ( add suffix = c("\_rm", "") ) * Remove the initial column with NAs (they have now suffix “\_rm”)   After the join, check that there is no NAs in plot\_E from tree with the function summary(). |

Answers to [Exercise 9](#exr-63):

## Your code  
# names(anci$plot\_E)  
  
tree <- tree |>  
 mutate(plot\_E = NA) |>  
 left\_join(anci$plot\_E, by = join\_by(plot\_no), suffix = c("\_rm", "")) |>  
 select(-ends\_with("\_rm"))  
  
summary(tree$plot\_E)

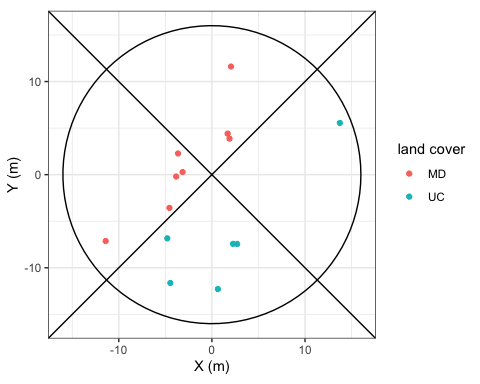
Min. 1st Qu. Median Mean 3rd Qu. Max.   
-0.08457 0.19793 0.24559 0.23280 0.28740 0.41069

Now that we have land cover information at tree level we can make a figure of tree location with color based on land cover code in [Exercise 10](#exr-64).

|  |  |
| --- | --- |
|  | **Exercise 10 (~ (optional) Figure of tree location with land cover)**   * Remake the figure of [Exercise 7](#exr-61) with these differences: * Change the color of trees based on lc\_code * Remove geom\_label\_repel() to stop showing the azimuths * Remove hline and vline * Add geom\_abline() with intercept = 0 and slope = 1 * Add geom\_abline() with intercept = 0 and slope = -1 * Check visually that different land cover are in different quadrants |

Type here answers to [Exercise 10](#exr-64):

## !!! Solution  
tree |>   
 filter(plot\_no == 123, subplot\_no == "B") |>   
 ggplot(aes(x = tree\_x, y = tree\_y)) +   
 geom\_point(aes(color = lc\_code)) +   
 geom\_path(data = circ16, aes(x = x, y = y)) +   
 geom\_abline(intercept = 0, slope = 1) +  
 geom\_abline(intercept = 0, slope = -1) +   
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)", color = "land cover")

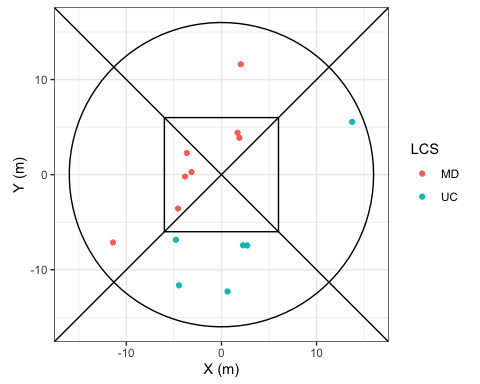


Subplot ‘123B’ is split between two land covers, Mixed deciduous (MD) in the land cover sections (LCS) 1, 2 and 5, and agriculture in LCS 3 and 4.

### 6.2.3 (optional) Final graph

We can add the square center LCS with a specific table

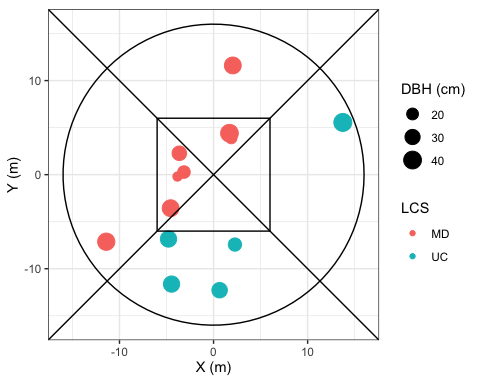
center\_sq <- tibble(x = c(-6, 6, 6, -6, -6), y = c(6, 6, -6, -6, 6))  
  
tree |>   
 filter(plot\_no == 123, subplot\_no == "B") |>   
 ggplot(aes(x = tree\_x, y = tree\_y)) +   
 geom\_point(aes(color = lc\_code)) +   
 geom\_path(data = circ16, aes(x = x, y = y)) +   
 geom\_abline(intercept = 0, slope = 1) +  
 geom\_abline(intercept = 0, slope = -1) +   
 geom\_path(data= center\_sq, aes(x = x, y = y)) +  
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)", color = "LCS")



### 6.2.4 Bonus: Tree DBH true to size on figures with the ggforce package.

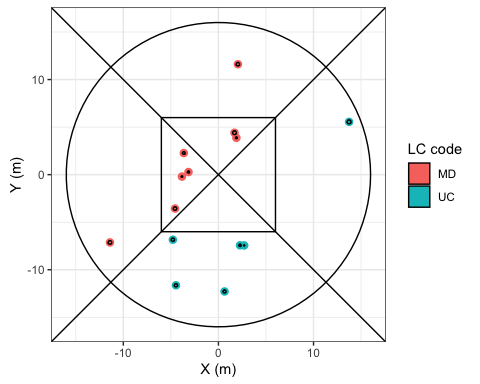
Graphs with ggplot() don’t natively intend to represent points true to size with geom\_point(), as size aesthetic is optimized for visibility. Here we represent the same figure as above but with point size change based on tree DBH

tree |>   
 filter(plot\_no == 123, subplot\_no == "B") |>   
 ggplot(aes(x = tree\_x, y = tree\_y)) +   
 geom\_point(aes(color = lc\_code, size = tree\_dbh)) +   
 geom\_path(data = circ16, aes(x = x, y = y)) +   
 geom\_abline(intercept = 0, slope = 1) +  
 geom\_abline(intercept = 0, slope = -1) +   
 geom\_path(data= center\_sq, aes(x = x, y = y)) +  
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)", color = "LCS", size = "DBH (cm)")



Now if we want to see trees’ DBH true to size, with the plot area, we can use geom\_circle() from the package ggforce.

tree |>   
 filter(plot\_no == 123, subplot\_no == "B") |>   
 ggplot(aes(x = tree\_x, y = tree\_y)) +   
 geom\_point(aes(fill = lc\_code), shape = 21, size = 3, stroke = 0) +   
 geom\_circle(aes(x0 = tree\_x, y0 = tree\_y, r = tree\_dbh/200, fill = lc\_code)) +   
 geom\_path(data = circ16, aes(x = x, y = y)) +   
 geom\_abline(intercept = 0, slope = 1) +  
 geom\_abline(intercept = 0, slope = -1) +   
 geom\_path(data= center\_sq, aes(x = x, y = y)) +  
 coord\_fixed() +  
 labs(x = "X (m)", y = "Y (m)", fill = "LC code", size = "DBH (cm)")



Here we kept the points as the trees are small and wouldn’t be visible otherwise.

# 7. Session 07: tree weights and basal area

**Goals**

1. Calculate tree weights for ratio estimators and for subplot per ha values
2. Calculate tree basal area

## 7.1 Tree weights for nested circles adjustment

### 7.1.1 Tree weight for ratio estimator

In the context of **ratio estimator,** tree weight converts small size trees’ variables to their equivalent if small trees were measured in the same area as the bigger sized trees.

For Lao PDR, the ratio between small and big trees is 4:

tree <- tree |> mutate(tree\_weight = if\_else(tree\_dbh < 30, 4, 1))

Let’s calculate where this value 4 comes from in [Exercise 11](#exr-71).

|  |
| --- |
| **Exercise 11 (~ Where does weight 4 come from?)**   * Calculate ‘Asmall’ as the area of a circle of 8m radius * Calculate ‘Abig’ as the area of a circle of 16m radius * Calculate Abig / Asmall |

Type here answers to [Exercise 11](#exr-71):

## Solution  
Asmall <- pi \* 8^2   
Abig <- pi \* 16^2  
Abig / Asmall

[1] 4

### 7.1.2 Tree weight for subplot per ha values

To get per ha at the subplot level, we first assign to each tree its subplot nested circle area, then the weight is the inverse of this area

tree <- tree |>   
 mutate(   
 tree\_spha = if\_else(tree\_dbh < 30, pi \* 16^2 / 10000, pi \* 16^2 / 10000),  
 tree\_weight\_spha = 1 / tree\_spha  
 )

There might be an error in the above code snippet, let’s solve it in [Exercise 12](#exr-72).

|  |
| --- |
| **Exercise 12 (~ Find the error)**   * Copy/paste the previous code snippet * Find and correct the error in tree\_spha |

Type here answers to [Exercise 12](#exr-72):

## !!! Solution  
tree <- tree |> mutate(   
 tree\_spha = if\_else(tree\_dbh < 30, pi \* 8^2 / 10000, pi \* 16^2 / 10000),   
 tree\_weight\_spha = 1 / tree\_spha   
 )

### 7.1.3 Calculate tree basal area

The basal area of a tree is its stem area footprint based on the measurement of it diameter at breast height.

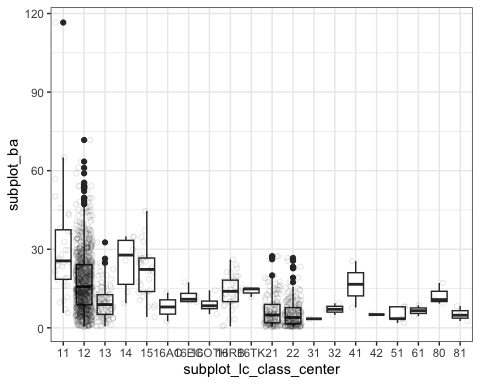
tree <- tree |> mutate(tree\_ba = pi \* (tree\_dbh/200)^2)

Let’s check subplot level basal area per land cover class in [Exercise 13](#exr-73).

|  |
| --- |
| **Exercise 13 (~ (optional) Basal area per subplot)**   * From tree, use group\_by() and summarise() to create subplot\_ba, the sum of tree\_ba \* tree\_weight\_spha per plot and subplot number, * Join lc\_class\_center from the subplot table using suffix control, * Make a boxplot of subplot\_ba against lc\_class\_center. |

Type here answers to [Exercise 13](#exr-73):

## !!! Solution  
tmp\_sp <- subplot |> select(plot\_no, subplot\_no, subplot\_lc\_class\_center)   
  
tree |>   
 group\_by(plot\_no, subplot\_no) |>   
 summarise(subplot\_ba = sum(tree\_ba \* tree\_weight\_spha), .groups = "drop") |>   
 left\_join(tmp\_sp, by = join\_by(plot\_no, subplot\_no)) |>   
 ggplot(aes(x = subplot\_lc\_class\_center, y = subplot\_ba)) +   
 geom\_boxplot() +   
 geom\_jitter(alpha = 0.1, shape = 21)



BONUS: The function geom\_jitter() is a nice addition to boxplot figures to see all the subplot values on top of their distribution.

# 8. Session 08: Tree aboveground biomass

**GOALS:**

1. add AGB at tree level for each forest type
2. Add AGB from chave 05 and 14 without height
3. Make a figure for natural forest and one for plantations, showing forest type AGB, and chave models

## 8.1 List of models

* "EG" ~ 0.3112 \* tree\_dbh^2.2331,
* "MD" ~ 0.523081 \* tree\_dbh^2,
* "DD" ~ 0.2137 \* tree\_dbh^2.2575,
* "CF" ~ 0.1277 \* tree\_dbh^2.3944,
* "MCB" ~ 0.1277 \* tree\_dbh^2.3944,
* "P\_AC" ~ 0.1173 \* tree\_dbh^2.454,
* "P\_EC" ~ 0.199 \* tree\_dbh^2.185,
* "P\_RB" ~ 0.0082 \* (pi\*tree\_dbh)^2.5623, ## Rubber model uses circumference
* "P\_TK" ~ 0.077 \* tree\_dbh^2.546,
* "P\_OTH" ~ 0.3112 \* tree\_dbh^2.2331,
* "RV" ~ 0.6 \* exp(-1.499 + 2.148 \* log(tree\_dbh) + 0.207 \* (log(tree\_dbh))^2 - 0.0281\*(log(tree\_dbh))^3),
* "B" ~ 0.6 \* exp(-1.499 + 2.148 \* log(tree\_dbh) + 0.207 \* (log(tree\_dbh))^2 - 0.0281\*(log(tree\_dbh))^3),
* Chave14 = round(exp(-1.803 - 0.976plot\_E + 0.976log(0.6) + 2.673log(tree\_dbh) -0.0299(log(tree\_dbh))^2)
* Chave05 = round(0.6 \* exp(-1.499 + 2.148log(tree\_dbh) + 0.207(log(tree\_dbh))^2 - 0.0281\*(log(tree\_dbh))^3)

## 8.2 AGB per forest type

We can assign different AGB models to trees for different forest types with mutate() and case\_when():

#table(tree$lc\_code)  
  
tree <- tree |>   
 mutate(   
 tree\_agb\_final = case\_when(  
 lc\_code == "EG" ~ 0.3112 \* tree\_dbh^2.2331,   
 ## ADD OTHER equations here   
 TRUE ~ 0  
 )   
 )

Let’s fill in the other cases in the [Exercise 14](#exr-81).

|  |
| --- |
| **Exercise 14 (~ Complete tree\_agb\_final)**   * For all forest land covers add their equations as shown at the beginning of the document. * For land cover with no equation, assign 0 with TRUE ~ 0. * Make a figure with ‘tree\_agb\_final’ against ‘tree\_dbh’ as points (optional) * Make a figure with ‘tree\_agb\_final’ against ‘tree\_dbh’ as line with color based on ‘lc\_code’ (optional) |

Type here answers to [Exercise 14](#exr-81):

## !!! Solution  
tree <- tree |>  
 mutate(  
 tree\_agb\_final = case\_when(  
 lc\_code == "EG" ~ 0.3112 \* tree\_dbh^2.2331,  
 lc\_code == "MD" ~ 0.523081 \* tree\_dbh^2,  
 lc\_code == "DD" ~ 0.2137 \* tree\_dbh^2.2575,  
 lc\_code == "CF" ~ 0.1277 \* tree\_dbh^2.3944,  
 lc\_code == "MCB" ~ 0.1277 \* tree\_dbh^2.3944,  
 lc\_code == "P\_AC" ~ 0.1173 \* tree\_dbh^2.454,  
 lc\_code == "P\_EC" ~ 0.199 \* tree\_dbh^2.185,  
 lc\_code == "P\_RB" ~ 0.0082 \* (pi\*tree\_dbh)^2.5623,   
 ## > Rubber model uses circumference  
 lc\_code == "P\_TK" ~ 0.077 \* tree\_dbh^2.546,  
 lc\_code == "P\_OTH" ~ 0.3112 \* tree\_dbh^2.2331,  
 lc\_code == "RV" ~ 0.6 \* exp(-1.499 + 2.148 \* log(tree\_dbh) + 0.207 \* (log(tree\_dbh))^2 - 0.0281\*(log(tree\_dbh))^3),  
 lc\_code == "B" ~ 0.6 \* exp(-1.499 + 2.148 \* log(tree\_dbh) + 0.207 \* (log(tree\_dbh))^2 - 0.0281\*(log(tree\_dbh))^3),  
 TRUE ~ 0  
 )  
 )  
  
# tree |>   
# ggplot(aes(x = tree\_dbh, y = tree\_agb\_final)) +   
# geom\_point()  
  
# tree |>   
# ggplot(aes(x = tree\_dbh, y = tree\_agb\_final)) +   
# geom\_line(aes(color = lc\_code))

## 8.3 AGB from Chave 2005 and 2014

Let’s add 3 models to all trees, regardless of their forest type to compare with the forest type based models, in [Exercise 15](#exr-82).

|  |
| --- |
| **Exercise 15 (~ Add AGB allometric equations for all trees)**   * Add ‘tree\_agb\_chave05’ with Chave et al. 2005 equation * Add ‘tree\_agb\_chave14’ with Chave et al. 2014 equation * Add ‘tree\_agb\_EG’ with the evergreen forest model |

Type here answers to [Exercise 15](#exr-82):

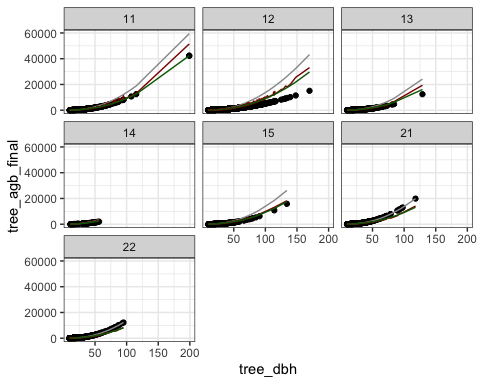
## !!! Solution  
tree <- tree |>   
 mutate(  
 tree\_agb\_chave05 =   
 0.6 \* exp(  
 -1.499 + 2.148\*log(tree\_dbh) +   
 0.207\*(log(tree\_dbh))^2 - 0.0281\*(log(tree\_dbh))^3  
 ),  
 tree\_agb\_chave14 =   
 exp(  
 -1.803 - 0.976\*plot\_E + 0.976\*log(0.6) +   
 2.673\*log(tree\_dbh) - 0.0299\*(log(tree\_dbh))^2  
 ),  
 tree\_agb\_EG = 0.3112 \* tree\_dbh^2.2331  
 )

## 8.4 Compare AGB models

### 8.4.1 Natural forest

Let’s compare our forest type specific tree aboveground with the Chave’s models and the model for Evergreen forests. Since there are many forest type let’s look at natural forests first with filter().

tree |>  
 filter(tree\_agb\_final != 0, lc\_no <= 29) |>  
 ggplot(aes(x = tree\_dbh, y = tree\_agb\_final)) +  
 geom\_point() +  
 geom\_line(aes(y = tree\_agb\_chave05), color = "grey60") +  
 geom\_line(aes(y = tree\_agb\_chave14), color = "darkred") +  
 geom\_line(aes(y = tree\_agb\_EG), color = "darkgreen") +  
 facet\_wrap(~lc\_no)



We can observe in this figure that the forest type specific models are quite conservative, since our AGB final model predicts smaller AGB for the same DBH compared to both Chave’s models.

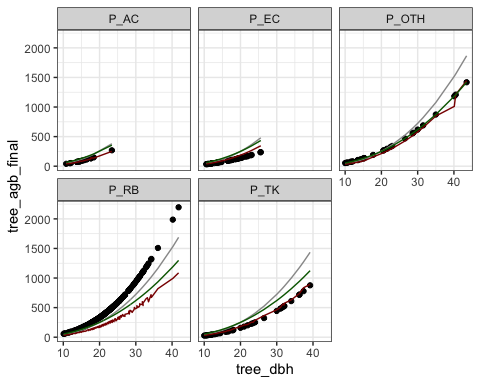
### 8.4.2 (optional) Planted forest

Let’s produce the same figure for planted forests in [Exercise 16](#exr-83).

|  |  |
| --- | --- |
|  | **Exercise 16 (~ Compare models for plantations)**   * Make a similar graph than 3.1 but for planted forest only |

Type here answers to the [Exercise 16](#exr-83):

## Your code here  
  
## !!! SOL  
tree |>  
 filter(tree\_agb\_final != 0, lc\_no >= 160) |>  
 ggplot(aes(x = tree\_dbh, y = tree\_agb\_final)) +  
 geom\_point() +  
 geom\_line(aes(y = tree\_agb\_chave05), color = "grey60") +  
 geom\_line(aes(y = tree\_agb\_chave14), color = "darkred") +  
 geom\_line(aes(y = tree\_agb\_EG), color = "darkgreen") +  
 facet\_wrap(~lc\_code)



## !!!

Plantation models selected for AGB final are also conservative in general, except for the rubber model.

# 9. Session 09: Aggregate tree to forest type AGB

**GOALS:**

1. Calculate the mean aboveground biomass for all forest types using **plot level** land cover and **systematic sampling**.
2. Calculate the mean aboveground biomass for all forest types using **ratio estimators** and **2-phases sampling for post-stratification**.
3. Compare estimates.

## 9.1 AGB with plot level land cover and systematic sampling

**Step-by-step**

1. Sum tree AGB to subplots.
2. Make plot majority LC class.
3. Create plot level AGB based on mean subplot AGB.
4. Create Forest type AGB as the mean of plot AGB per forest type.
5. Calculate sampling error.

### 9.1.1 Sum tree AGB to subplot

We will use group\_by() and summarise() to bring tree characteristics to the subplot level.

subplot\_agb <- tree |>  
 group\_by(plot\_no, subplot\_no) |>  
 summarise(  
 sp\_agb = sum(tree\_agb\_final \* tree\_weight\_spha / 1000),  
 sp\_ba = sum(tree\_ba \* tree\_weight\_spha),  
 sp\_dens = sum(tree\_weight\_spha),  
 tree\_meas = n(),  
 .groups = "drop"  
 )

### 9.1.2 Load plot level main land cover

Plot level land cover was generated, from the subplots center LC, with the following rules:

1. For plots with one unique LC, they were directly assigned to plot level.
2. For plots with multiple land cover, we used the majority LC for the whole plot.
3. If some plots have equal importance of two land covers, we used to minimum land cover class.

plot\_lc <- read\_csv("data/training\_plot\_lc.csv", show\_col\_types = FALSE)

### 9.1.3 Aggregate subplot to plot

We take the mean of subplot variables (except tree measured count).

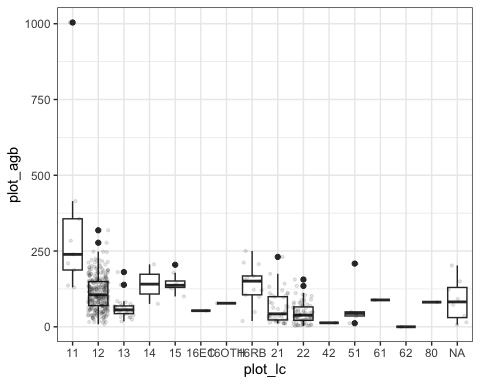
plot\_agb <- subplot\_agb |>  
 group\_by(plot\_no) |>  
 summarise(  
 plot\_agb = mean(sp\_agb),  
 plot\_ba = mean(sp\_ba),  
 plot\_dens = mean(sp\_dens),  
 plot\_tree\_meas = sum(tree\_meas)  
 ) |>  
 left\_join(plot\_lc, by = join\_by(plot\_no))

Let’s look at plot AGB per land cover in [Exercise 17](#exr-91).

|  |
| --- |
| **Exercise 17 (~ Boxplot of plot AGB)**   * Make a boxplot of plot AGB against land cover classes |

Type here the answers to [Exercise 17](#exr-91):

## !!! Solution  
plot\_agb |>  
 ggplot(aes(x = plot\_lc, y = plot\_agb)) +  
 geom\_boxplot() +  
 geom\_jitter(alpha = 0.1, size = 0.8)



### 9.1.4 Aggregate to forest type AGB

Now we can take averages of plot AGB for each forest type. For the uncertainty, we also need **standard deviation** of AGB and the **number of plots**.

ftype\_agb <- plot\_agb |>  
 group\_by(plot\_lc) |>  
 summarise(  
 plot\_count = n(),  
 total\_tree\_meas = sum(plot\_tree\_meas),  
 ftype\_dens = mean(plot\_dens),  
 ftype\_ba = mean(plot\_ba),  
 ftype\_agb = mean(plot\_agb),  
 ftype\_agb\_sd = sd(plot\_agb)  
 )

We can calculate the uncertainty for systematic design with this formula:

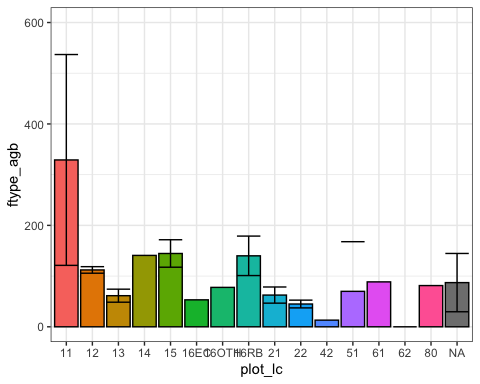
The student’s law function in R is qt(), and uncertainty can be calculated from the AGB mean and standard deviation as follows:

ftype\_agb <- ftype\_agb |>  
 mutate(  
 ftype\_agb\_se = ftype\_agb\_sd / sqrt(plot\_count),  
 ftype\_agb\_me = qt(0.975, plot\_count - 1) \* ftype\_agb\_se,  
 ftype\_agb\_U = ftype\_agb\_me / ftype\_agb \* 100,  
 ftype\_agb\_cilower = ftype\_agb - ftype\_agb\_me,  
 ftype\_agb\_ciupper = ftype\_agb + ftype\_agb\_me  
 )

Warning: There was 1 warning in `mutate()`.  
ℹ In argument: `ftype\_agb\_me = qt(0.975, plot\_count - 1) \* ftype\_agb\_se`.  
Caused by warning in `qt()`:  
! NaNs produced

We can then make a plot of forest type AGB with sampling uncertainty. We added a limit on the y-axis to avoid uncertainty of Coniferus Forest (CF, code 14) condensing the whole figure with ylim(0, 600).

ftype\_agb |>  
 ggplot(aes(x = plot\_lc, y = ftype\_agb)) +  
 geom\_col(aes(fill = plot\_lc), col = "black") +  
 geom\_errorbar(aes(ymin = ftype\_agb\_cilower, ymax = ftype\_agb\_ciupper)) +  
 theme(legend.position = "none") +  
 ylim(0, 600)



## 9.2 AGB with ratio estimator and 2-phases sampling for post-stratification

Step-by-step:

1. Prepare subplot x LCS level table
2. Use the aggregate function ‘nfi\_aggregate3()’ to get all aggregation levels
3. Compare results with equal plot statistiscs (table ‘ftype’)

### 9.2.1 Prepare the subplot x LCS table for aggregation

#### 9.2.1.1 Initial data preparation

For the second approach, all subplots must be accounted for, visited or not, accessible or not. This table has been prepared and can be loaded into the ancillary data list anci.

anci$ph2 <- read\_csv("data/training\_anci\_phase2.csv", show\_col\_types = F)  
  
ph2\_subplot <- anci$ph2  
  
# format(nrow(anci$ph2) / 20, big.mark = ",")

We can observe that anci$ph2 contains 1,067 *(1,067)* plots, as initially planned in the 12 x 18 km grid.

To generate this table, instead of starting from tree aggregates or from the subplot table, we start from the vector of planned Phase 2 plots and we recreate all combinations of subplot x LCS with expand\_grid() .

vec\_ph2 <- anci$ph1 |>   
 filter(!is.na(plot\_id)) |>  
 pull(plot\_id) |>  
 sort()  
  
test\_ph2 <- expand\_grid(  
 plot\_id = vec\_ph2,  
 subplot\_no = c("A", "B", "C", "D"),  
 lcs\_no = 1:5  
) |>  
 mutate(subplot\_id = paste0(subplot\_no, lcs\_no))

We can check that this table is identical to anci$ph2 with all.equal():

tmp\_ph2 <- anci$ph2 |>   
 select(plot\_id, subplot\_no, lcs\_no, subplot\_id)  
  
all.equal(tmp\_ph2, test\_ph2)

[1] TRUE

We then need to add **accessibility** from subplot and subpopulation and **strata** from phase 1 data with left\_join() in [Exercise 18](#exr-92).

|  |  |
| --- | --- |
|  | **Exercise 18 (~ (optional) Join core info to phase 2 table)**   * Prepare ‘tmp\_sp’ a table containing plot\_no, subplot\_no and subplot\_access from the subplot table and rename plot\_no into plot\_id. * Prepare ‘tmp\_ph1’ a table containing plot\_id, subpop, stratum , ph1\_prov\_no and ph1\_prov\_name from the table anci$ph1 and rename ph1\_prov\_no into prov\_no and ph1\_prov\_name into prov\_name * Join these 2 tables to ph2\_subplot with suffix control method. |

Type here answers to [Exercise 18](#exr-92):

## !!! Solution  
tmp\_sp <- subplot |>  
 select(plot\_id = plot\_no, subplot\_no, subplot\_access)  
  
tmp\_ph1 <- anci$ph1 |>  
 select(plot\_id, subpop, stratum, prov\_no = ph1\_prov\_no, prov\_name = ph1\_prov\_name)  
  
  
ph2\_subplot <- ph2\_subplot |>  
 mutate(  
 subplot\_access = NA,  
 subpop = NA,  
 stratum = NA,  
 prov\_no = NA,  
 prov\_name = NA  
 ) |>  
 left\_join(tmp\_ph1, by = join\_by(plot\_id), suffix = c("\_rm", "")) |>  
 left\_join(tmp\_sp, by = join\_by(plot\_id, subplot\_no), suffix = c("\_rm", "")) |>  
 select(-ends\_with("\_rm"))

In the final preparation phase 2 subplot table, correction for shifted plots was implemented beforehand.

#### 9.2.1.2 (NOTE) Investigating phase 2 plots in stratum 4: non-forest

The Phase 2 focused on forest and RV land cover, but some plots from Phase 1 and planned for phase 2 were non-forest in the Phase 1 interpretation and still visited.

tmp\_ph2\_nonforest <- anci$ph1 |>   
 filter(!is.na(plot\_id), stratum == 4)  
  
vec\_ph2nf <- tmp\_ph2\_nonforest |> pull(ph1\_plot\_no)   
  
tmp\_ph2nf\_visited <- subplot |>  
 filter(plot\_no %in% vec\_ph2nf, subplot\_no == "A")   
  
## Save the tables in a new folder  
if (!"results" %in% list.files()) dir.create("results")  
  
write\_csv(tmp\_ph2\_nonforest, "results/ph2\_nonforest\_planned.csv")  
write\_csv(tmp\_ph2nf\_visited, "results/ph2\_nonforest\_visited.csv")

#### 9.2.1.3 **Re-coding accessibility**

Accessibility plays an important role in the final estimates. Since not all the initial 1,067 plots were visited, there is missing information on **many non-forest plots** and **some forest plots**. To simplify the calculations, it was considered that:

* **all non-visited forest plots would be considered as non-accessible,**
* **all non-visited non-forest plots would be considered as accessible with biomass 0.**

This is important to keep in mind in case biomass of trees outside forest becomes relevant for future NFI cycles.

Code for re-coding accessibility:

ph2\_subplot <- ph2\_subplot |>  
 mutate(  
 access = case\_when(  
 plot\_id <= 636 & subplot\_access == "accessible" ~ TRUE,  
 plot\_id <= 636 & subplot\_access != "accessible" ~ FALSE,  
 plot\_id <= 636 & is.na(subplot\_access) ~ FALSE,  
 stratum %in% 1:3 ~ FALSE,  
 stratum == 4 ~ TRUE,  
 TRUE ~ NA  
 )  
 )

#### 9.2.1.4 Subplot largest area

For aggregation with ratio estimator we need to keep track of the areas where trees are measured. Since we have a tree weight for small trees, we are only interested in the larger area:

* 12 m side square of a quarter
* difference between the 16m radius circle and the 12m side square.

ph2\_subplot <- ph2\_subplot |>  
 mutate(  
 sp\_area = if\_else(lcs\_no == 1, 12^2, (pi\*16^2 - 12^2)/4) / 10000  
 )

#### 9.2.1.5 Aggregate tree to subplot x LCS

We can now aggregate trees to the subplot x LCS level and join with the phase 2 subplot table.

ph2\_sp\_tree <- tree |>  
 select(plot\_id = plot\_no, subplot\_no, lcs\_no, tree\_weight, tree\_ba, tree\_agb\_final) |>  
 mutate(  
 subplot\_id = paste0(subplot\_no, lcs\_no)  
 ) |>  
 group\_by(plot\_id, subplot\_id) |>  
 summarise(  
 dens = sum(tree\_weight),  
 ba = sum(tree\_ba \* tree\_weight),  
 agb = sum(tree\_agb\_final \* tree\_weight) / 1000, ## Convert kg to tons  
 .groups= "drop"  
 )

ph2\_subplot <- ph2\_subplot |>  
 mutate(  
 dens = NA,  
 ba = NA,  
 agb = NA  
 ) |>  
 left\_join(ph2\_sp\_tree, by = join\_by(plot\_id, subplot\_id), suffix = c("\_rm", "")) |>  
 select(-ends\_with("\_rm")) |>  
 mutate(  
 dens = if\_else(!is.na(dens), n\_tree, 0),  
 ba = if\_else(!is.na(ba), ba, 0),  
 agb = if\_else(!is.na(agb), agb, 0)  
 )

### 9.2.2 Run the aggregation function

The aggregation function 3, for 2-phases sampling with ratio estimator, takes phase 1 and phase 2 data as input as well as the attributes and for the ratio. can change based on what attribute we want to aggregate, while is always the minimum measured area.

Let’s see an example with tree density and no subpopulation.

ph2\_data <- ph2\_subplot |> mutate(subpop = 1)  
  
res3\_dens <- nfi\_aggregate3(  
 .ph1\_df = anci$ph1,   
 .ph2\_sp = ph2\_data,   
 .class\_d = lc\_no,   
 .attr\_y = dens,   
 .attr\_x = sp\_area,   
 .aoi\_area = 23680000  
 )  
  
res3\_dens\_tot <- res3\_dens$totals\_short

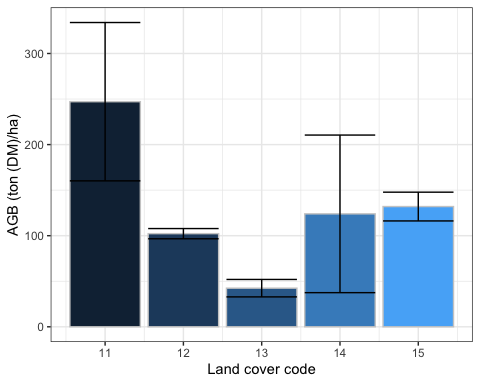
The function output is a list with different levels of data aggregation.

Practice now the function nfi-aggregate3() to get basal area and aboveground biomass in [Exercise 19](#exr-93).

|  |  |
| --- | --- |
|  | **Exercise 19 (~ Aggregate basal area and AGB)**   * (optional) Create res3\_ba as the output of the aggregation function with basal area as .attr\_y . * Create res3\_agb as the output of the aggregation function with aboveground biomass as .attr\_y . * Save the totals\_short output from res3\_agb to a new object res3\_agb\_tot. * Make a barplot with AGB per ha against land cover with their error bars, for natural forest only. |

Type here answers to [Exercise 19](#exr-93):

## !!! Solution  
ph2\_data <- ph2\_subplot |> mutate(subpop = 1)  
  
res3\_agb <- nfi\_aggregate3(  
 .ph1\_df = anci$ph1,   
 .ph2\_sp = ph2\_data,   
 .class\_d = lc\_no,   
 .attr\_y = agb,   
 .attr\_x = sp\_area,   
 .aoi\_area = 23680000  
 )  
  
res3\_agb\_tot <- res3\_agb$totals\_short  
  
res3\_agb\_tot |>  
 filter(lc\_no < 20) |>  
 ggplot(aes(x = lc\_no, y = Rd)) +  
 geom\_col(aes(fill = lc\_no), col = "grey80") +  
 theme(legend.position = "none") +  
 geom\_errorbar(aes(ymin = Rd - Rd\*Rd\_mep/100, ymax = Rd + Rd\*Rd\_mep/100)) +  
 labs(  
 x = "Land cover code",  
 y = "AGB (ton (DM)/ha)"  
 )



We can use an additional package: kableExtra, to customize tables with adding extra headers and even colors. However this package focuses on HTML and PDF outputs, for DOCX, we stick to simple kable() function. Let’s show the totals for AGB:

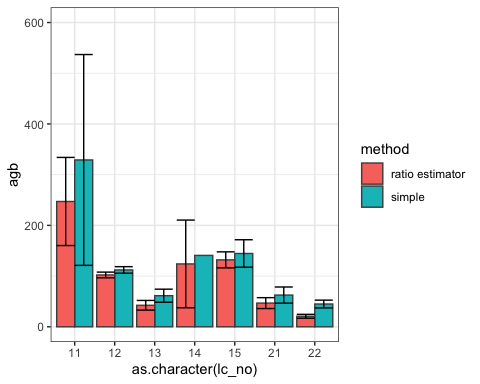
tab\_agb <- res3\_agb$totals\_d |>  
 select(lc\_no, Xd, Xd\_mep, Xtot, Yd, Yd\_mep, Rd, Rd\_mep, Ytot) |>  
 filter(lc\_no <= 22 | lc\_no >= 160)  
  
  
if (knitr::is\_html\_output() | knitr::is\_latex\_output()){  
   
 kbl(  
 tab\_agb,  
 col.names = c(  
 "Land cover", "prop.", "U (perc)", "tot.", "prop.", "U (perc)",   
 "ton (DM)/ha", "U (perc)", "tot."   
 ),  
 format.args = list(big.mark = ",", digits = 2),  
 booktabs = F  
 ) |>  
 kable\_styling(  
 bootstrap\_options = c("bordered", "condensed", "responsive"),  
 full\_width = FALSE  
 ) |>  
 add\_header\_above(  
 c(" " = 1, "Area" = 3, "AGB" = 5), border\_left = T, border\_right = T  
 ) |>  
 column\_spec(1, border\_left = T) |>  
 column\_spec(9, border\_right = T)  
   
} else {  
   
 kable(  
 tab\_agb,  
 col.names = c(  
 "Land cover", "Area prop.", "Area U (perc)", "Area tot. (ha)", "AGB prop.",  
 "AGB prop. U (perc)", "AGB (ton DM/ha)", "AGB U (perc)", "AGB tot. (tons)"   
 ),  
 format.args = list(big.mark = ",", digits = 2)  
 )  
  
}

| Land cover | Area prop. | Area U (perc) | Area tot. (ha) | AGB prop. | AGB prop. U (perc) | AGB (ton DM/ha) | AGB U (perc) | AGB tot. (tons) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 11 | 0.00525 | 47.4 | 124,260 | 1.297 | 60.0 | 247 | 35.2 | 3.1e+07 |
| 12 | 0.15406 | 6.2 | 3,648,029 | 15.750 | 8.1 | 102 | 5.5 | 3.7e+08 |
| 13 | 0.01993 | 23.9 | 471,951 | 0.845 | 35.3 | 42 | 22.6 | 2.0e+07 |
| 14 | 0.00107 | 101.4 | 25,424 | 0.133 | 141.4 | 124 | 69.8 | 3.2e+06 |
| 15 | 0.00456 | 45.9 | 107,889 | 0.601 | 49.5 | 132 | 12.0 | 1.4e+07 |
| 21 | 0.02898 | 18.7 | 686,166 | 1.354 | 29.2 | 47 | 23.0 | 3.2e+07 |
| 22 | 0.07202 | 10.7 | 1,705,431 | 1.495 | 21.1 | 21 | 18.3 | 3.5e+07 |
| 161 | 0.00041 | 116.5 | 9,725 | 0.017 | 143.5 | 42 | 73.6 | 4.1e+05 |
| 162 | 0.00159 | 84.0 | 37,725 | 0.029 | 164.7 | 18 | 125.0 | 6.9e+05 |
| 164 | 0.00736 | 34.5 | 174,181 | 1.025 | 40.6 | 139 | 22.5 | 2.4e+07 |
| 165 | 0.00035 | 105.5 | 8,177 | 0.028 | 101.2 | 81 | 11.7 | 6.6e+05 |
| 169 | 0.00132 | 84.8 | 31,237 | 0.041 | 132.9 | 31 | 78.8 | 9.7e+05 |

## 9.3 Compare AGB estimates from the 2 aggregation methods

Check AGB from 2 methods

tmp\_agb1 <- ftype\_agb |>   
 filter(plot\_lc %in% c("11", "12", "13", "14", "15", "21", "22")) |>  
 mutate(  
 plot\_lc = as.numeric(plot\_lc),  
 method = "simple"  
 ) |>  
 select(method, lc\_no = plot\_lc, agb = ftype\_agb, agb\_U = ftype\_agb\_U)  
  
tmp\_agb2 <- res3\_agb$totals\_short |>  
 filter(lc\_no <= 22) |>  
 mutate(method = "ratio estimator") |>  
 select(method, lc\_no, agb = Rd, agb\_U = Rd\_mep)  
  
agb\_compa <- tmp\_agb1 |> bind\_rows(tmp\_agb2)  
  
agb\_compa |>  
 ggplot(aes(x = as.character(lc\_no), y = agb)) +  
 geom\_col(aes(fill = method), col = "grey30", position = position\_dodge()) +  
 geom\_errorbar(aes(group = method, ymin = agb - agb\*agb\_U/100, ymax = agb + agb\*agb\_U/100), position = position\_dodge()) +  
 ylim(0, 600)



The ratio estimator method had big impact on EG and RV. Also big influence on uncertainty of CF.

# 10. Session 10: Building a R community of practice in Lao PDR

**TBD**

# Conclusion

The training aimed to continue FIPD capacity building to handle the NFI data analysis with R. This training session focused on core parts of full calculation chain for the NFI cycle 4. The trainings elements were:

* Solving subplot coding errors manually based on time stamps
* Joining subplot level and phase 1 information to trees
* Calculating tree weights, basal area and aboveground biomass
* Aggregating tree aboveground biomass to domain average AGB per ha based on equal area plots and simple averages
* Practicing a custom function to get areas and AGB estimates per ha and totals following double sampling for post-stratification with ratio estimator.

This cycle adopted a more complex method than previous cycles with the introduction of ratio estimators and the inclusion of the Phase 1 observations in the estimation of domain estimates of core variables’ averages and totals. This method is more sensitive to stratum assignment and few plots from stratum 4 (non-forest) were found to be forested and have tree records. This may explain the large differences between simple averages and weighted averages for two-phases sampling, in particular for evergreen forest and regenerating vegetation and requires further investigation. These plots may be shifted plots where the land cover was re-assigned but stratum was not.

In terms of capacity building, FIPD team expressed interest in adopting R for forest management and small scale forest estimations, which are more frequent requests for them, in order to have more opportunities to practice. Additional training was also suggested to share and practice the calculations for other pools.

# ANNEX: Agenda

|  |
| --- |
| Figure 1: Agenda part 1 |

|  |
| --- |
| Figure 2: Agenda part 2 |

# ANNEX: participants list

**TBD**