Laboratory 2: Managing Data with R

Felipe Bravo Oviedo
Universidad de Valladolid, Spain

In this Lab we will learn how to:

- Generate new variables in a dataset in R
- Select and rename columns from a dataset
- Subset data in R
- Merge data in R
- Export a dataset

We will use also the knowledge we acquire in our previous lab including the use of our working directory, loading and viewing datasets in R.

Generating new variables in a dataset in R

We will use the data sets (data0, data1, data2, data3 and data4) that we uploaded in the previous lab (see the document "Lab 01: An Introduction to R") We can use different operators, both arithmetic and logic, to generate new variables.

Arithmetic operators

+	Adition
-	Substraction
*	Multiplication
/	Division
^ or **	Power

Logic operators

<	lower than
<=	lower than or equal to
>	greater than
>=	greater than or equal to
==	equal to
!=	not equal to
x | y	x or y (alternatively)
x & y	x and y (simultaneously)
isTRUE(x)	test if x is true

For instance, we can generate a new variable in *data4* as the ratio of the basal area of *Pinus sylvestris* L. (G21) over the total basal area (GTOTAL)

```
data4$ratio <- data4$G21/data4$GTOTAL

# as you can see the structure is name_of_dataset$name_of_variable
```

Also is possible to generate new variables from two different datasets if we type something similar to this code:

```
newdata$ratio <- data3$variable1/data4$variable3

# with this code you are generating a new data file (newdata) where you will store
# the variable ratio which the division between variable 1 (from data3)
# and variable3 (from data4)
```

Select and rename columns from a dataset

First we will observe the name of the columns of the dataframe by using the names( ) function:

```
# view the name of the columns of the dataframe
names (name_of_the_dataframe)
```

if we test this structure with the dataframe data4 as follow:

```
names (data4)
```

we will obtain the following:

```
[1] "PARCELA" "SI4"     "GTOTAL"  "G21"     "VTOTAL"  "V21"     "NTOTAL"  "N21"     "Dg"
[10] "Hdom"    "Ddom"    "Hart"    "Hg"      "SDI"     "Dg21"    "ratio"
```

Now we can select keep some of the columns by using this code. It is important to know that as we will use same name as the original data, we are overriding the original file in R (not the original file we uploaded and still is our working directory). Using the following code we will keep the first 8 columns:

```
# selecting a range of  columns
# note by doing this you will overwrites the original R dataframe
# and you will need to upload again the original to R if are wrong
data5 <- data4[,c(1:6)]
names(data5)
```

We should obtain the following:

```
[1] "PARCELA" "SI4"     "GTOTAL"  "G21"     "VTOTAL"  "V21"
```

Now we will keep only some specific columns (the number 1, 3 and 8) by coding as follow:

```
# selecting some specific  columns
# note by doing this you will overwrites the original R dataframe
# and you will need to upload again the original to R if are wrong
data6 <- data4[,c(1, 7, 8)]
names(data6)
```

We should obtain the following:

```
[1] "PARCELA" "NTOTAL"  "N21"
```

Now we can change the name of the variable on one by one basis:

```
# changing the name of one individual column
names(data6)[2] <- "Treesperha"
```

or all the variables together with just one line of code:

```
# changing the names of every columns
names(data6) <- c("PARCELA", "Treesperha", "Treesperha_psyl")
names(data6)
```

Subseting data in R

We can use the subset ( ) function to select some variables and observation according to defined criteria. We will keep the observations with basal area equal or greater to 20 m^2/ha and (simultaneously) proportion (ratio) of Pinus sylvestris basal area equal or greater that 80%. We can do this task with this code:

```
# using subset function
# be aware that you need to upload again the data to R and calculate the variable ratio
newdata <- subset(data4, GTOTAL >= 20 & ratio >= 0.80)
View(newdata)
```

Merging data in R

Sometime we need to join two different data sets to generate a new one which contains the desired variables. To do this first we need that both files share at least one column (for instance an variable, let say plot or tree, ID). We'll use the dataframes data5 and data6 we created in the previous select and rename columns section that have in common the column named "PARCELA". In both files each observation has an unique ID the number of the plot ("PARCELA"). Now we are going to create a new dataset by joining data5 and data6

```
# merge data5 and data6 to create a new object called "merged_data"
merged_data <- merge(data5, data6, by="PARCELA")
```

Exporting a dataset in R

Now we will save the file merged_data in our working directory with a new name (join_data) in format csv with the following code:

```
# write the dataframe merged_data to a csv named join_data in your working directory
write.csv(merged_data, "join_data.csv", row.names = F)
```