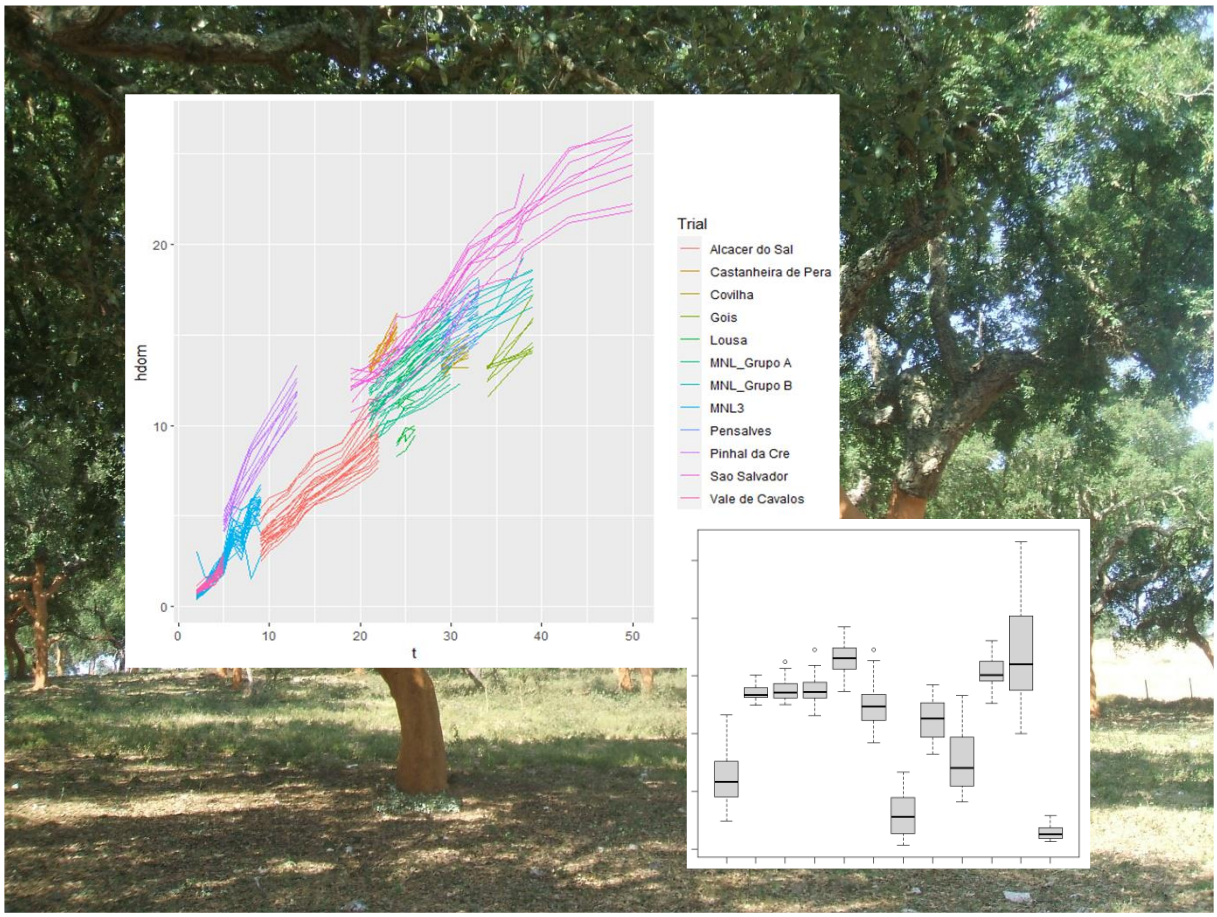
Universidade de Lisboa – Instituto Superior de Agronomia
Centro de Estudos Florestais

INTRODUÇÃO AO R

Exercícios

Margarida Tomé



Textos Pedagógicos TP 1/2024

ForChange Forest Ecosystem Management under Global Change

NOTA PRÉVIA

Este volume consiste num conjunto de exercícios selecionados com base em dados reais e pretendem ajudar os alunos a aprender como ler e fazer uma primeira visualização e análise de dados com recurso ao software R.

.

.

Lisboa, Fevereiro 2024

ÍNDICE

1. Leitura de dados e primeira análise

Ficheiro de dados: “Pb_PP&Ensaios.xlsx”

Script: “1_Pb_PP&Ensaios”

a) Utilize a função *getwd()* para descobrir para qual pasta o R está apontado e a função *setwd()* para alterar o *workspace* para a pasta onde guardou o ficheiro “Pb_PP&Ensaios.xlsx”. Verifique com o comando *getwd()* se a alteração foi bem-sucedida.

b) Para ler ficheiros EXCEL é necessário instalar, por exemplo, a *package “xlsx”* ou a *package “readxl”*. Se já o tiver instalado, basta ir para a alínea seguinte, caso contrário deverá instalá-lo. A melhor maneira de aprender como instalar *packages* em R é simplesmente pesquisar no Google “R install packages” (ou algo semelhante); também pode usar a ajuda do R que está na janela “files” ou digitar *?install* na consola. Outra *package* muito útil para manipulação de dados é o *“dplyr”*. Deve instalá-la já se ainda não o tiver feito.

   O *script* “0_installPackages” exemplifica a instalação das duas *packages* com o comando *install.packages()*. Só é preciso instalar uma *package* uma vez. As *packages* também podem ser instaladas diretamente no RStudio.

c) Tem que carregar as *packages* que são necessários para a manipulação dos dados (*“xlsx”* ou *“readxl”* e *“dplyr”*) com a função *load()*.

   As *packages* têm que ser carrgadas em cada nova sessão de R.

d) Está agora pronto para ler (com os comandos *read.xlsx()* ou *read_xlsx()*, consoante a *package* que carregou) os dados do ficheiro EXCEL que contém os dados. O ficheiro tem duas sheets: uma com os dados (Pb) e outra com a descrição das variáveis (var). Queremos ler a *sheet* Pb e guardá-la numa *dataframe* de nome Pb (convém dar nomes pequenos às dataframes porque dentro do script cada variável é reconhecida pelo nome da *dataframe* seguido de $ e o nome da variável, por exemplo *Pb$G*).

e) O ficheiro Pb contém os dados de biomassa por componente (*Ww*-lenho, *Wb*-casca, *Wbr*-ramos, *Wl*-folhas). Calcule a biomassa total acima do solo (*Wa*)

f) O R tem diversas funções para descrever e resumir os dados que estão disponíveis numa *dataframe* ou em cada um dos vetores componentes, por exemplo: *head()*, *summary()*, *apply()*, *length()*, *tapply()*. Use o *help* do R e/ou faça uma pesquisa na web sobre esses comandos e experimente-os com a *dataframe* que acabou de criar.

g) Crie uma variável do tipo *factor* a partir da variável *ID_plot* (um *fator* só pode tomar um número limitado de valores)

h) As funções R da família *apply* são muito úteis para resumir dados classificados de acordo com os valores de um f*ator*. Use a função *tapply* (use o help para aprender sobre a sintaxe da função) para calcular o valor médio da variável *N* para cada uma das parcelass. Guarde os resultados numa variável chamada *Navg_plot*. Repita o cálculo, mas para o desvio padrão.

i) Utilize a função *count* para contra o número de medições em cada parcela.

j) Use a função *plot* para fazer um gráfico da altura dominante em função da idade.

k) Explore agora a função *ggplot*, mais "ponderosa" do que a função *plot*, para fazer alguns gráficos alternativos da altura dominante em função da idade:
   - Juntando com uma linha as medições de cada parcela
   - Juntando com uma linha as medições de cada parcela mas utilizando uma cor diferente para cada ensaio
   - Juntando com uma linha as medições de cada parcela utilizando uma cor diferente para cada parcela e acrescentando uma legenda

l) Use a função *filter* para criar um conjunto de dados que seja um subconjunto do original considerando apenas os gráficos com *Cod_Trial*="SS" e use a função *ggplot* para representar graficamente a altura dominante em função da idade para cada parcela usando uma cor diferente para cada parcela e uma legenda para identificar as parcelas.

m) Use a função *ggplot* para fazer o gráfico da biomassa de lenho (*Ww*) em função do volume total (*V*) e acrescente uma linha linear de tendência.

n) Use as funções *ggplot* e *boxplot* para fazer uma gráfico do tipo *box-plot* da altura dominante em função do fator *Cod_Trial*.

o) Faça um histograma do número de árvores por ha

p) Use a função *par*() com o argumento *mfrow* para fazer o gráfico da biomassa de cada componente em função da biomassa total numa matriz 2x2 de gráficos.

q) Crie uma *dataframe Pb_graf* apenas com os vetores das biomassas (*Ww*,*Wb*,*Wbr*,*Wl*) e use a função *plot()* para fazer o gráfico da dataframe. Qual o resultado?

r) Use a função *cor()* para calcular as correlações entre as variáveis da dataframe Pb, depois de lhe retirar as variáveis alfanuméricas.

s) Instale agora a *package "GGally"* e explore a função *ggcorr()* para fazer gráficos de correlações entre variáveis.

2. Leitura de ficheiros de dados grandes e mais análise exploratória

Data files used: "IFN5_Pb.xlsx" or "IFN5_Ec.xlsx" (generic name "IFN5_Sp.xlsx")

Script: "4.2_NFIData_LargeDataset"

a) Use the function *getwd()* to find out to which directory/forder R is pointed and the function *setwd()* to change the workspace to the folder where you stored the IFN_Sp.xlsx file. Check with the command *getwd* that the change was successful

b) In order to read large EXCEL files, there is the need to use the library (*readxl*). You can also use the library *(RODBC) (ImportExport)*. To read the file you need to use the *odbcConnectExcel2007()* and *sqlFetch()* commands.

c) Now just use the commands that you already learned in the previous exercise to characterize and explore the data.

d) Create a data frame with the names of the variables changed for the names in your native language.

e) Create a file with a subset of the initial variables: id_plot, d, dclass, d0, h, hc, sample_tree

f) Calculate the tree basal area (g) for each tree and a variable with value=1 named n (each trees represents 1 tree within the plot)

g) Use the aggregate() function to compute plot basal area (Gplot) and number of trees per plot (Nplot)

h) Read the information in the Plot_Sp sheet and keep the information on the plot area. Merge the two data.frames (with tree and plot information) by Id_plot and expand the values of basal area and number of trees in the plot to the ha

i) Create a file filtering just the trees that are sample trees and that have a measurement of tree height.

j) It is often usefull to have an equation that allows to estimate diameter at breast height as a function of stump diameter (for instance if we need to estimate stand variables after the stand being harvested). Use the file created in the previous alinea to fit a simple linear model that estimates the diameter at breast height as a function of the diameter measured at the tree base.

3. Reading and preparing data organized in several EXCEL worksheets/files

Data file used: “Ec_StandGrowthData.xlsx”

Script: “4.3_PlayWith_Ec_GrowthData”

a) Use the function *getwd()* to find out to which directory/forder R is pointing and the function *setwd()* to change the workspace to the folder where you stored the Ec_StandGrowthData. Check with the command *getwd* that the change was succesfull

b) Start by loading the packages that will be needed for the data manipulation (“xlsx” and “dplyr”) with the *library()* function.

c) Read (with the *read.xlsx()* or *read_xlsx()* commands) the data from each one of the 4 worksheets, each worksheet into one dataframe (AV, AV_plot, EPE and EPE_plot).

d) Use the R functions *head()*, *summary()*, *apply()*, *length()*, *tapply()* to get familiar with the 4 dataframes you that you just built.

e) Using the *rbind()* function, create two dataframes, Ec (AV+EPE) and Ec_plot (AV_plot+EPE_plot), that make an union (add the rows of both dataframes in a unique dataframe)

f) Using the *merge()* function, create a dataframe (Ec) that merges the dataframes euca and Ec_plot by adding the variables available in Ec_plot to the dataframe Ec (merge), using the ID_plot and ID_rot variables to make the merge. Calculate the variable age (t) for each plot at time of measurement.

   Use the functions *head()* and *str()* to analyse the structure of the euca_all dataframe.

g) Write the data in the dataframe Ec in an EXCEL file and in an R file.

h) Study something about the %>% R function and how it can be used to simplify the R scripts.

i) Use the functions:

   a. *select()*, to select just some variables

   b. *rename()*, to rename some variables

   c. *filter()*, to select some individuals (rows)

   d. *mutate()*, to add variables computed from existing variables

j) As an alternative to the *mutate()* function, use just an assignment to add a variable dg to the dataframe euca_all. For that

k) Use the *aggregate()* function to calculate the means and standard deviations of the hdom and G variables by the combination (ID_plot,ID_rot) and place the results in a dataframe.

l) Use the *lag()* function to create a dataframe with variables that are lagged, as it is needed to fit growth functions in the form of difference equations.

m) Repete the previous exercise with the *lead()* function.

n) Split the data in the dataframe Ec in two datasets, to fit and to validate the models, each with approximately 50% of the plots (not of the data points).

4. More exploratory data analysis with graphics

The best way to start a data analysis is by explore the relationships that exist in the data set by displaying them in appropriate graphics.

Data file used: “Ec_StandGrowthData.xlsx”

Script: “GraphicalDataAnalysis_Ec_GrowthData”

a) Prepare a R data file by:
   - reading the the data from each one of the 4 worksheets, each worksheet into one dataframe (AV, AV_plot, EPE and EPE_plot)
   - make the union of the data from the two data sets that contain stand data information (AV and EPE) and the same for the two data sets that contain plot data information (AV_plot and EPE_plot)
   - merge the information available in the two files previously created using ID_plot and ID_rot to make the merge and calculate plot age using the data of measurement and the data of regeneration
   - delete the variables that, in your opinion, are not relevant for the data analysis

b) Suppose that you want to develop a model to estimate the total aboveground biomass (Wa) and the biomass per tree component (stem, stem wood, stem bark, branches and leaves, respectively Ws, Ww, Wb, Wbr and Wl) from the stand variables that are usually available from forest inventories. In order to have an idea of the relationships between variables make an exploratory graphical analysis by ploting:

- Plot Wa, Ws, Ww, Wb, Wbr and Wl over each one of the stand variables t, hdom, N, G and V
- Plot graphs among the variables Wa, Ws, Ww, Wb, Wbr and Wl
- Plot Ww, Wb, Wbr and Wl over G with colors by rotation
- Plot Ww, Wb, Wbr and Wl over G with colors by N classes (N<1000,1000>=N<2000, N>2000)

5. Developing a model to estimate site index from site and stand variables for cork oak

Data file used: Sb_SiteIndex&SiteData.xlsx

Scripts: 6.6A_Sb_SiteIndex_DataPrep.R

6.6B_Sb_SiteIndex_ExploratoryAnalysis.R

6.6C_Sb_SiteIndex_SelectingSubsetsOfVar.R

6.6D_Sb_SiteIndex_AnalysingOneModel.R

6.6E_Sb_SiteIndex_ComparingCandidateModels.R

a) Read all the worksheets available in the file
b) Use the R function “merge” to obtain a unique data set with all the information available
c) Estimate the site index for each plot with the site index curves:

$$S = \frac{20.7216}{\left(1 - \left(1 - \frac{20.7216}{hdom}\right)\left(\frac{t}{80}\right)^{1.4486}\right)}$$

d) Analyse the frequency of observations per category in categorical variables and recode some if needed
e) Save the final data set as an R datafile for future use
f) Create dummy variables for each categorical variable: f1) using the *ifelse()* function; f2) automatically with the *dummy_cols()* function (requires library *fastDummie*); f3) check the frequency of data points in each category; group some categories with a low frequency in the data
g) Fit a linear regression between S and the set of dummy variables defined for each categorical variable and find out which variables/categories give a better prediction of S
h) Fit a linear regression between S and each categorical variable defined as a factor and compare the results with those obtained in the previous question. Try to find out the main difference between the two regressions (answer: with the dummies you can use just some of the categories, which is not possible if you use the categorical variable)
i) Use the *summary()* function to “look” at the variables and find out that there are some variables with missing values (NA). Create a dataframe that deletes the data points with missing values
j) Estimate the correlation coefficient between S and each one of the site and climate continuous variables available

k) Plot the site index (S) over the several site and climate continuous variables available and see the type of relationship that exists (positive, negative, linear, non-linear, strong, weak)
l) Fit a linear regression between S and each continuous variable and find out which variables give a better prediction of S
m) Use stepwise algorithms (e.g. functions of the family *ols_step_method_p()* ou *ols_step_method_aic()*, ou *step()*, ou *stepAIC()*) to select subsets of variables to e used as candidate models to estimate cork oak site index from environmental variables
n) Use now all possible regressions algorithms (e.g. *ols_step_all_possible()*, *ols_step_best_subset()*) and select some more candidate models
o) Compare the candidate models previously selected using the fitting and prediction statistics that can be used to characterize: the fitting; the prediction bias; the prediction precision. Check also if the models fulfil the regression assumptions. Propose one or, if justified, two models to be used for the estimation of cork oak stands

References

Bevan KJ, Kirkby MJ, 1979. A physically based, variable contributing area model of basin hydrology. Hydrol Sci Bull Sci Hydrol 24:43–69.

Paulo JA, Palma JHN, Gomes AA, Faias SP, Tomé J, Tomé M, 2015. Predicting site index from climate and soil variables for cork oak (*Quercus suber* L.) stands in Portugal. New Forests 46(2): 293-307. Doi: 10.1007/s11056-014-9462-4.

Sánchez-González M, Tomé M, Montero G, 2005. Modelling height and diameter growth of dominant cork oak trees in Spain. Ann For Sci 62:633–643.

Tomé, M., Ribeiro, R. P., Marques, M., 1999. Inventário Florestal do concelho da Chamusca. Relatórios técnico-científicos do GIMREF, nº 1/1999, Instituto Superior de Agronomia, Lisboa, Portugal.

Tomé, M., Oliveira, T., Soares, P., 2006. O modelo Globulus 3.0. Publicações GIMREF - RC2/2006. Departamento de Engenharia Florestal, Instituto Superior de Agronomia, Lisboa.

ANNEXE 1 – Description of data files used as support to the course

Before solving each exercise, be sure that you are familiar with the data files needed for the exercise, that are described in the following items.

a) Pb_PP&Trials.xlsx

This file includes stand level variables of a set of permanent plots and trials – thinning and pruning trials, spacing trials – established in maritime pine stands.

b) IFN5_Arv_Pb.xlsx

This file was extracted from the Portuguese National Forest Inventory data base (IFN5) and includes measurements undertaken in plots pure or dominated by maritime pine.

c) IFN5_Arv_Ec.xlsx

This file was extracted from the Portuguese National Forest Inventory data base (IFN5) and includes all the measurements undertaken in plots pure or dominated by eucalyptus.

d) SinglePlotVolumeData.xlsx

Volume growth data for an eucalypt plot

e) Ec_ StandGrowthData_1.xlsx

This file includes the evolution of volume for an eucalypt plot over time in the first spread-sheet and a set of permanent plots with the evolution of basal area and dominant height over time on the second spread-sheet.

f) Ec_StandGrowthData_2.xlsx

This file is a sub-set of the data used to develop the GLOBULUS 3 model (Tomé et al. 2006).

The file includes stand level data from trials and permanent plots established in eucalyptus plantations in Portugal.

The same file includes several sub-sets, one for the permament plots (EPE) and the others each corresponding to one trial (AV, QP). The data from each sub-set includes two worksheets, one with the plot level information and another with the values of the stand variables over time. Variables names are according to the IUFRO standards.

g) File “3.1. BiomassData_Ec_trees.xlsx” contains data from the destructive sampling of eucalyptus young trees representing different spacings and clones

- Plot total biomass over diameter at 0.5 m, using different symbols according to the clone. Discuss the results
- The same as a) but according to the spacing. Discuss the results, comparing them with the results obtained in a)

h) Sb_SiteIndex&SiteData.xlsx

This data file is a sub-set of the data used by Paulo et al. (2015) to develope a model to estimate site index in cork oak stands in Portugal.

The file includes several worksheets: 1) Property – with the cod and name of the properties (Cod_property and property) where the plots are installed and the code of the closest metereologic station (Cod_Meteo); 2) StandVariables – with the information, for each plot, of the dominant diameter and height (dudom and hdom) and of stand age (t); 3) Soil_twi – with the information on soil and litology characteristics and topographic wetness index (Bevan and Kirkby 1979); 4) Climate – with information on several climate variables for each metereologic station.

It was not possible to open a soil pit in some of the plots, therefore the variables Soil_depth and Soil_depth_A are not available for all the plots.

Soil, litology and topographic variables:

Variable	**Description**
Data from the Portuguese Agency for the Environment website (http://sniamb.apambiente.pt/webatlas/).	
Litology	Litology according to Silva(1983)
Soil_FAO	FAO soil group according to the IUSS Working Group WRB (2006) classification
Observation of the soil profile (soil pit)	
Soil_depth	Soil depth until the R/C or C/R horizon was reached
Soil_depth_A	Thickness of the A horizon
Observation of the soil	
Soil_texture	Soil textural class (fine, medium and coarse)
Soil_texture_ A	Soil textural class (fine, medium and coarse) of the A horizon
Computed from the Jarvis et al. (2008) digital terrain model	
Twi	topographic wetness index developed by Bevan and Kirkby (1979)

Climate variables (average monthly data over the 30-year period 1961–1990:

Variable	Description
Tmin	minimum temperature (ºC)
T	mean temperature (ºC)
Tmax	maximum temperature (ºC)
HR_9	relative humidity at 9 hours
HR_15_18	relative humidity at 15-18 hours
P	mean monthly precipitation (mm)
NdaysP	number of days with precipitation per month
Evap	monthly evaporation with Piche evaporimeter (mm)
Ndays_Tmin<0	number of days with Tmin < 0
Ndays_Tmin>20	number of days with Tmin > 20
Ndays_Tmin>25	number of days with Tmin > 25
NdaysFog	mean number of days with fog per month
NdaysDew	mean number of days with dew per month
NdaysFrost	mean number of days with frost per month
Martonne	Martonne climatic index (De Martonne 1925) (M=Pannual/T+10)