Minimum recommended R skills for spectroscopy modeling

Philipp Baumann // philipp.baumann@usys.ethz.ch

This document contains useful R resources for R topics that you should be familiar with prior to starting with soil and plant spectral modeling. There are plentiful of good R books, tutorials and other online resources such as blogs that cover these aspects. Thus, feel free to explore other resources to improve your R skills to manage and analyze your data efficiently and in a reproducible manner. You can have a very steep learning curve in R. In particular, being familiar with the basics of the R language and setting up a proper R working environment can help you to avoid or understand common errors and pitfalls. **Bold text** in this pdf document indicates that there is a link for online resources you can click on.

R resources

Install the newest version of R

The **Comprehensive R Archive Network** (CRAN) is a network of servers that mirror code and documentation of the R language and contains the R package repository that currently features over 10000 available packages. You can download the latest binary version of R for your operating system.

Start working with the R environment

RStudio is an integrated developing environment (IDE) for R. It is open source and is available for Linux, Windows and Mac. RStudio facilitates writing and executing R code and provides, among many other features, interactive help, code highlighting and completion. Download the lastest version of *RStudio Desktop (Open Source License)*.

Before you start working with R in the RStudio environment, I strongly recommend you to set up RStudio in a way that facilitates working continuously on your data analysis and modeling projects. You should always work from R scripts (text files with the extension *.R*) to document your data analysis. Creating RStudio projects helps you to keep track of your data files, scripts and outputs. Please follow the project workflow provided in **this chapter** of the online book **R for Data Science**. Besides, this book is useful for you to learn you best practices for reading, transforming, visualizing and modeling of data.

Basic foundations for running R code are well documented in **this chapter** of R for Data Science. This chapter covers how to work with R objects and how to do assignments to create objects with associated values. This helps you to avoid common pitfalls when working with objects in R environments.

General Introduction to R

The document **An Introduction to R** is provided by the R core team. There are many other comprehensive manuals and tutorials that introduce R as well. Some of them are listed on CRAN (see **here** and go to `Documentation > Contributed`).

Introduction to R: *Online resources*

The free DataCamp course **Introduction to R** provides you with tutorials to understand and practice working with basic R syntax and concepts of data structures such as vectors, matrices, data frames and lists. Its interactive online sessions are equipped with a virtual R environment you can directly practice within browser window. An alternative introduction tutorial to R is the **O'RELLY Code School**.

To get a deeper understanding of basic R data structures I recommend you to read the Chapter 2 *Data structures* of the book **Advanced R** (p. 13–31, see **here** for an online link to this chapter).

Data manipulation for exploratory analysis

The DataCamp course **Exploratory Data Analysis in R: Case Study** offers a free case study including tutorial videos for an exploratory data analysis based on United Nations voting dataset. This course covers data rearrangement, summary, filtering and sorting based on the famous and powerful dplyr package. Check topics in session 1 Data cleaning and summarizing with dplyr. The second session of this course covers data visualization using the ggplot2 package. You might need to make a free account on DataCamp in order to work on the second session of the course.

This package vignette of the dplyr package delivers you the principles of the dplyr tools and comprehensive examples.

Good coding style

There is not one single style guide to format well readable R code. Nevertheless, it is very helpful to have a concise and meaningful notation and naming practice as well as syntax style and code commenting guidelines. A good example of well organizing and formatting R code is provided in **this chapter** of Advanced R.

Cheat sheets

RStudio provides well-illustrated and comprehensive cheat sheets that make it easy to remember R functions for specific data analysis tasks. The cheat sheet pdf documents are available **here**.

You might find those particularly useful:

- *Data Import Cheat Sheet*
- *Data Transformation Cheat Sheet*
- *RStudio IDE Cheat Sheet*
- *Data Visualization Cheat Sheet*
- *Contributed Cheatsheets: Base R*

List of R topics and tasks you should master prior to attending the workshop

In order to make this training course efficient and useful for you, I expect that you have a good understanding of and be experienced with the following aspects of R free software environment for statistical computing and graphics:

- Be familiar to work with basic R data structures: (atomic) vectors, lists, matrices, data frames.
- Good knowledge of subsetting operators (`[[`, `[` and `$`) and behavior for basic R data structures. Know how to combine subsetting and assignment (also consider the **"Subsetting" chapter** of the Advanced R book.

- Practice with basic R functions for working with basic R data structures and statistics.
- Data transformation and manipulation using the dplyr package. Familiarity with the pipe (`%>%`) operator.
- Data visualization using the base R plotting system using a generic function such as `plot()`) as well as using the ggplot2 package.
- *Recommended*: Applied knowledge of common methods in multivariate statistics (e.g., multiple linear regression and Principal Component Analysis (PCA)). However, we will also cover these topics in the course.