School of Geography & Earth Sciences
McMaster University

GEOG 4GA3: Applied Spatial Statistics

INSTALLATION OF R AND R STUDIO

1. Background

Statistical analysis is the study of the properties of a dataset. There are different aspects of statistical analysis, and they often require that we work with data that are messy. According to Grolemund and Wickham (2017), computer-assisted data analysis includes the steps outlined in Figure 1.

First, the data are imported to a suitable software package. This can include data from primary sources (suppose that you collected coordinates using a GPS) or from secondary sources (the Census of Canada). Data will likely be text tables, or an Excel file, among other possible formats. Before data can be analyzed, they need to be tidied. This means that the data need to be arranged in such a way that they match the process that you are interested in. For instance, a travel survey can be organized so that each row is a traveler, or as an alternative so that each row is a trip.

Once that data are tidy, Exploratory Data Analysis (EDA) and/or its geographical extension Exploratory Spatial Data Analysis (ESDA) can be conducted. This involves transforming the raw data into information. Examples of transformations include calculating the mean and the standard deviation. Visualization is also part of this exploratory exercise. In EDA this could be creating a histogram or a scatterplot. Mapping is a key visualization technique in spatial statistics.

Modeling is a process that further extracts information from the data, typically by looking at relationships between multiple variables.

All of the tasks mentioned above, and many more, can be handled easily in a variety of software packages. For this course, you will use the statistical computing language R.

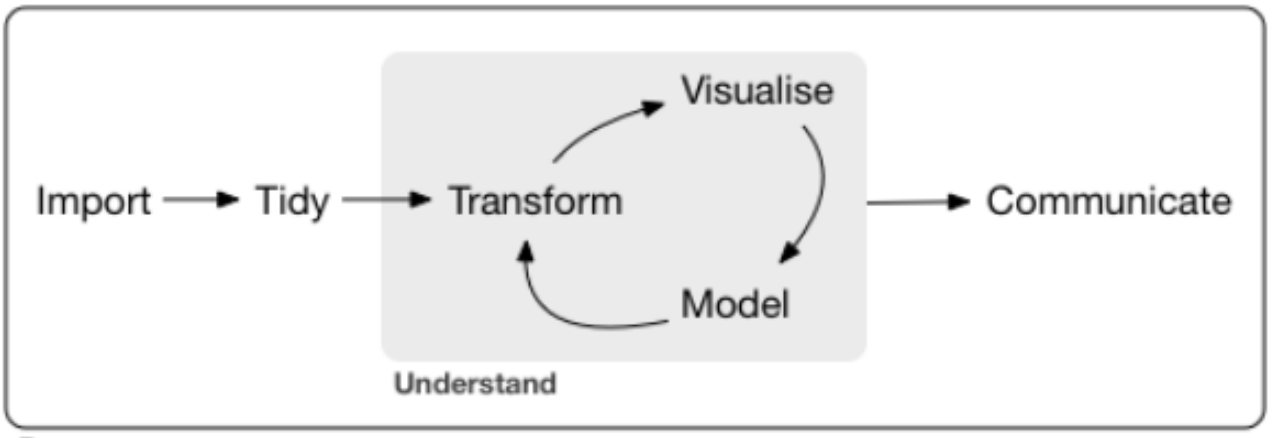

Figure 1. The process of doing data analysis (from Grolemund and Wickham, 2017)

2. R: The open statistical computing project

What is R?

R is an open-source language for statistical computing. It was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, as a way to provide their students an accessible, no-cost tool for their courses. R is now maintained by R Development Core Team, and developed by hundreds of contributors around the globe. R is an attractive alternative to other software packages for data analysis (e.g., Microsoft Excel) due to its open-source character (i.e., it is free), its flexibility, and huge user community, which means if there's something you want to do (for instance, linear regression), it is very likely that someone has already developed a package for it in R.

A good way to think about R is as a core package, to which libraries can be attached to increase its functionality. R can be downloaded for free at:

https://cran.rstudio.com/

R comes with a built-in console (a user graphical interface), but better alternatives to the basic interface, including R Studio, which can also be downloaded for free here:

https://www.rstudio.com/products/rstudio/download/

R requires you to work using the command line, which is going to be unfamiliar to many of you accustomed to user-friendly graphical interfaces. Do not fear. People worked for a long time using the command line, or even more cumbersome, punched cards in early computers. Graphical user interfaces are convenient, but they have a major drawback, namely their inflexibility. A program that functions based on graphical user interfaces allows you to do only what is hard-coded in the user interface. Command line, as we will see, is somewhat more involved, but provides much more flexibility in operation.

Go ahead. Install R and RStudio in your computer. If you are working in the GIS lab, you will find that these have already been installed there.

Before discussing tables further, we will quickly tour R Studio.

3. R Studio window

The R Studio window provides a complete interface to interact with the language R. It consists of a window with several panes. Some panes include in addition several tabs. There are the usual drop-down menus for common operations. See Figure 2 below.

The editor pane allows you to open and work with text and other files, where you can write instructions that can be passed on to the program. Writing something in the editor does not execute the instructions, it merely records them for possible future use.

The console pane is where instructions are passed on to the program. When an instruction is typed (or copied and pasted) there, R will understand that it needs to do something. The instructions must be written in a way that R understands, otherwise errors will occur.

The environment is where all data that is currently in memory is reported. The History tab acts like a log: it keeps track of all instructions that have been executed in the console.

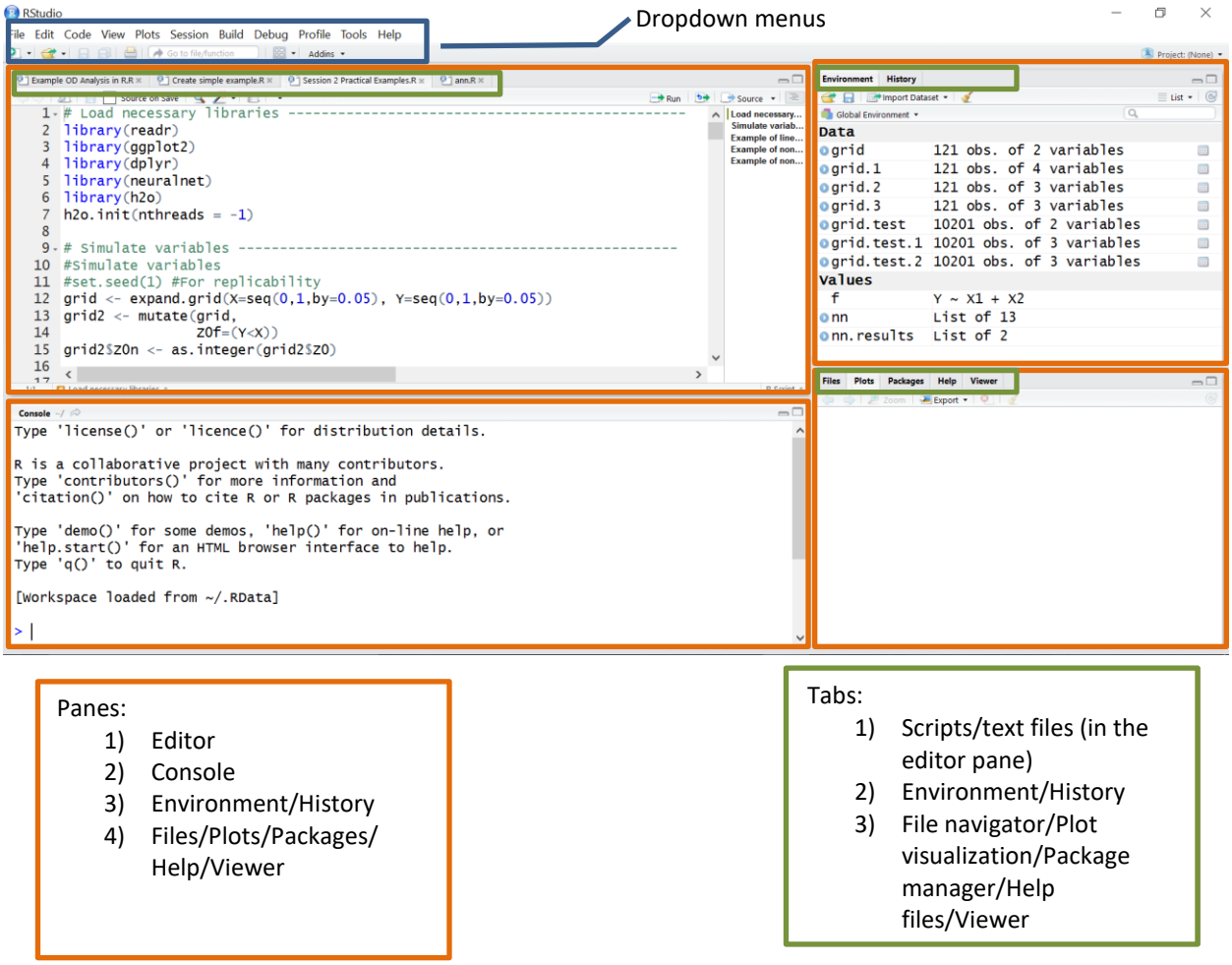

Figure 2. R Studio Window.

The last pane includes a number of useful tabs. The File tab allows you to navigate your computer, change the working directory, see what files are where, and so on. The Plot tab is where plots are rendered, when instructions require R to do so. The Packages tab allows you to manage packages, which as mentioned above, are pieces of code that can augment the functionality of R. The Help tab is where you can consult the documentation for functions/packages/see examples, and so on. The Viewer tab is for displaying local web content, for instance, to preview a Notebook (more on Notebooks soon).

Once you have installed R and R Studio, download the file "01 Basic Operations and Data Structures.Rmd", and use R Studio to open it.

References

Grolemund G and Wickham H (2017) R for Data Science. O'Reilly. Also see: http://r4ds.had.co.nz/index.html