Data management and introduction to GIS and R for spatial analysis

Module information

Author: Meghan Graham MacLean (mgmaclean@umass.edu)

Department of Environmental Conservation, University of Massachusetts, Amherst, MA 01003

# Overview

Today is a quick refresh or introduction to GIS and R and RStudio for spatial analysis. Students will open and visualize a land cover dataset in QGIS and R Studio, as well as determine the most abundant land cover type in Hampshire County, MA\*. Ensure that students have both QGIS and RStudio installed prior to trying to complete this module.

## Connection to other modules

I connected this module with the NEON Data TIEE module presentation published in 2016 by Jim McNeil1 and Megan Jones2 that can be found [here](http://tiee.esa.org/vol/v13/issues/data_sets/mcneil/abstract.html) to relay the importance of metadata and data organization. I, however, did not use their mark-recapture data and instead used the opportunity to build on this lesson and introduce spatial data and data manipulation in R and GIS.

1 - Smithsonian-Mason School of Conservation, George Mason University, Smithsonian Conservation Biology Institute, Front Royal, VA, 22630

2 - National Ecological Observatory Network – Battelle, Boulder, CO 80301, Corresponding Author: Megan A. Jones (mjones01@battelleecology.org)

# Learning objectives

By the end of the module, students should be able to:

* gain familiarity with land cover
* gain familiarity with QGIS and RStudio
* understand land cover types and quantify area of each type
* create and modify metadata files

# Data and code

* National Land Cover Database in Massachusetts (or your state of choice) - MA\_NLCD.tif\*
* Shapefile of unlabelled points for the students to figure out - points.shp\*
* Code: Raster-explorer.Rmd

\*Note: I highly recommend personalizing this module to your area - especially if you are planning to do more with spatial data in your region. This makes the land cover maps much more accessible to students if they have some real life experience with the landscape. You can download data from [www.mrlc.gov](http://www.mrlc.gov) or the state GIS database (if available).

# Assessment and implementation strategies

I introduced that the class would be doing this module a week ahead of time and asked the students to ensure they had both QGIS and RStudio installed and updated prior to class. I also posted/gave to the students some introductions/help files to RStudio and QGIS (linked or stated below) for students nervous about using either platform. Then the module was started within a single 1.25 hour class period and then questions (written below) were assigned as group work to be submitted the following week. I taught the module with groups of 2-3 to discuss each of the questions, though it could be taught as either a group or individual activity. To form groups, I had students self-identify as either having more familiarity with GIS or more comfortable in RStudio (or neither), and then had them find partners or 2 others that complimented their skills. Students were then asked to use their expertise to teach their fellow group-mates. This seemed to work well for an upper-level undergraduate/graduate level class. Much of the initial class period was used to troubleshoot individual computers (even though the students were asked to have QGIS and R installed prior to class). So groups used outside class time to complete most of the module.

Data management and introduction to GIS and R for spatial analysis

# Student directions

# Overview

Today is a quick refresher or introduction to GIS and R and RStudio for spatial analysis. You will open and visualize a land cover dataset in QGIS first, as well as determine the most abundant land cover type in Hampshire County, MA. Then we will go through similar processes and answer the same questions in RStudio, so use this opportunity and each other's expertise to learn the program you feel less comfortable with!

**For the next class: Please have all of the answers to the below questions and plan to have each member of your group report your answers to the class (please equitably distribute the questions).**

# Introduction to Data Management

Data management is super important for both yourself to remember what you did when, but also for reproducibility of analyses and to those that may use your data at a later date. So, we will practice some of the basics of data management by both organizing our data in a logical fashion along with creating metadata for the data in this module. **Metadata** is the who, what, when, where, how, and why of data. All metadata should include:

* WHO created the data?
* WHAT are the contents of the data?
* WHEN were the data created?
* WHERE is it geographically?
* HOW were the data developed?
* WHY were the data developed?

So, let’s get started!

1. Pick a location on your computer hard drive (I might suggest in your documents) and create a folder for this module
2. Save the data and script for this module in this new folder
3. Start a .txt editor (e.g. text edit, notepad, notepad++) and create a metadata file with the following information and save it in that folder with the name “ReadMe.txt”:
	1. Who: the names of the people in your group
	2. What: the names of the datasets and code used in this module (you’ll add a description later!)
	3. When: today’s date

Now that you have this file, we’ll use GIS and RStudio to explore the data more and then add more information to our metadata file.

# Introduction to GIS

While this module will be taught using QGIS because it is free, open source, and can work on a variety of operating platforms (e.g., Windows, Mac OS X, Linux), if you are more familiar with ArcGIS, please feel free to use it instead. If you are new to GIS, I recommend using the latest stable version QGIS (3.10, and the time of writing this module) along with the tutorial, which can be found [here](https://docs.qgis.org/3.10/en/docs/training_manual/index.html) and complete Module 2: The Interface, to get you started.

Once you feel comfortable where buttons are in QGIS, import the two provided datasets:

* MA\_NLCD.tif
* points.shp

Once you have those in your workspace, **answer the following questions as a group**:

1. What is the resolution, or cell size, of MA\_NLCD.tif?
2. What is the projection of MA\_NLCD.tif?
3. What do you think is the most prevalent land cover type ([check out the NLCD website for help](https://www.mrlc.gov/data/legends/national-land-cover-database-2016-nlcd2016-legend) and the included legend)?
4. What do you think is the most prevalent forest type?
5. If you zoom in on the provided points – what do you think they are?
6. BONUS CHALLENGE: Can you extract what land cover most of the points fall in?

# Introduction to R and RStudio for spatial analysis

Open RStudio and the R Markdown file provided. If you are new to RStudio, I recommend using the Data Carpentry intro to R, which can be found [here](https://datacarpentry.org/R-genomics/01-intro-to-R.html), to get you started.

Once you are able to input data in RStudio, run each of the script sections in the provided R Markdown file and as a group try to understand what is happening in each section of the script.

Once you have completed the script, please try to answer the following questions:

1. What is the resolution, or cell size, of MA\_NLCD.tif?
2. What is the projection of MA\_NLCD.tif?
3. What is the most prevalent land cover type?
4. What is the most prevalent forest type?
5. What do you think the points represent?
6. BONUS CHALLENGE: can you compute how much area of forest there is in Massachusetts?

# Revisit your metadata

Update your ReadMe.txt (metadata file) with the information you learned about your spatial data! Add a description to each of the datasets (in the WHAT section) that describes what is in each of them. Then add:

1. Where: where geographically are each of the datasets?
2. How: how do you think each of the datasets were created?
3. Why: why were these created or why might they be useful?

Save and feel good about working in both QGIS and RStudio!

**Bonus to discuss in class next time**: which program did you like better? Why?