FISH 6002: Data Collection, Management, and Display

Dr. Brett Favaro

Fall 2017

E-mail: brett.favaro@mi.mun.ca Web: mifisheriesscience.github.io/courses/6002Data/
Office Hours: Thursdays 0900-1200 Class Hours: M 1500-1700, T 0900-1000
Office: Marine Institute W2009 Class Room: W3033/35

Modern fisheries scientists work in a complex data environment. This course will introduce students to the basics of R statistical software - including programming best-practices, optimizing workflows, and producing tidy data. A focus on data display and visualization will be present throughout this course, reflecting the importance of good graphing habits in science.

Learning Outcomes

By the end of this course, students will develop a working competency of R Statistical Software, and will be introduced to the software environment that fisheries scientists must master. Great importance will be placed on fostering the ability to self-teach and stay abreast of developments in data collection, management, display, and programming.

Students will also develop the following specific competencies:

Data collection

- Understand the wide diversity of data that one may collect within fisheries science
- Understand "Tidy data"
  - How to collect it
  - How keeping data tidy helps with analysis
  - How to clean up untidy data in a reproducible, transparent way (especially using dplyr)
- Familiarity with best practices in recording data in the field, and preventing loss between collection and analysis
- Understand how to acquire and use data from databases

Data management

- Understand metadata, archiving, and how to build an efficient project workflow
- Able to create and implement a data management plan
- Understand how to manipulate data and prepare it for analysis
- Focus: *dplyr*

- Understand the difference between open, community-supported software, and conventional licensed software
- Able to produce reproducible, well-documented R code
- Understand the concept of open data

Data display

- Able to communicate data visually, selecting the appropriate figure to represent data
- Familiarity with both R base plots and the ggplot2 ecosystem
- Demonstrated ability to produce effective graphs for:
  - Scientific posters
  - Scientific publications
  - Powerpoint-based talks
- Ability to produce figures that meet journal standards for publication

Expectations and Aspirations

Look, I get it - very few people go into fisheries or biological sciences because they want to program all day long. But the reality is that computer competency is a core skill - just like reading and writing - for the modern scientist, and you are no different. But if you build these skills early on, then you will spend less time overall struggling to make your way throught the digital landscape of science.

My goal for you is to emerge with two skillsets. First: the ability to **efficiently collect, store, and prepare data for analysis and display** Second: To be able to **create beautiful and effective visual depictions of data** using R and the ggplot2 package.

My expectation is that if this is all new to you, that you keep up with the course and seek help proactively when needed. If you're more advanced, my hope is that you share your knowledge with the rest of the class, so that we can all become more effective at these critical skills needed to succeed in science.

Course Structure

The course will meet twice weekly - one 2-hr block and one 1-hr block. Speaking generally, we will spend about 1 hour of lecture introducing the theory behind the week's activities. The remaining two hours will be spent on how to actually use this knowledge in the conduct of research.

Reference Books

The Internet is awash in information on the subject matter covered by this course. The following books and papers are excellent references. You don't need to buy them all, but you should certainly look at all of them at some point during your research career.

The Tufte books are timeless references for data display theory. The Wickham books and articles are constantly evolving as the software changes, but represent great starting places.

Essentially, our theory will follow Tufte, and our practice will derive from Wickham. In some cases, full text of the books may be available online for free.

Tufte, Edward R. (1986). *The visual display of quantitative information* , p. 200. ISBN: 978-0961392147.

Tufte, Edward R. (1990). *Envisioning information*. New York: Graphics Press, p. 126.

Tufte, Edward R. (1997). *Visual explanations: images and quantiles, evidence, and narrative* , p. 157. ISBN: 02768739. DOI: 10.1109/TPC.1998.678564.

Tufte, Edward R. (2006). *Beautiful evidence*. New York: Graphics Press LLC, p. 213. ISBN: 0961392177.

Wickham, Hadley (2014). "Tidy data". In: *Journal of Statistical Software* 59.1, pp. 1–23. DOI: 10.18637/jss.v059.i10.

Wickham, Hadley (2016). *ggplot2: elegant graphics for data analysis (use R!) 2nd ed. 2016 edition*. Springer, p. 260. ISBN: 978-1491910399. http://ggplot2.org/book/.

Wickham, Hadley and Garrett Grolemund (2017). *R for data science: visualize, model, transform, tidy, and import data*. O'Reilly Media, p. 518. ISBN: 978-1491910399. http://r4ds.had.co.nz/index.html.

Course Policies

Social Media

Students are welcome to tweet about the course using the hashtag #MIData - but the Chatham House Rule is in effect. That means you **may not reveal the identity of the person speaking** in your tweets without their express permission. We want to encourage people to actively participate and make mistakes without fear of their mishaps being broadcast acrosst he world.

Code of Conduct

You have the right to expect a supportive, safe environment in this course. This course will be governed by our Fisheries Science Code of Conduct, which all participants are expected to respect.

Digital Competency

Students are expected to have basic computer competency. You should be able to operate Microsoft Word, Powerpoint, and Excel, or equivalent (e.g. OpenOffice or Google Docs). You should be able to download and install software onto your computer. Please install R Statistical Software and RStudio prior to begining the course.

If you lack these skills, please consult training materials on your own time. **Please bring a laptop to every class.**

E-mail Policy

E-mail is not a primary tool for communication in this class. If you have questions about course content, your order of operation should be:

1. Check the syllabus
2. Ask in class, or discuss with colleagues
3. Ask on Slack (this way, everyone can benefit from an answer)
4. Request a meeting with me

If emailing me a meeting request, use the subject line "FISH 6002: Meeting request." Please indicate three potential meeting times (I prefer afternoon meetings) and explain in 1-3 lines what you want to meet about.

E-mail is impersonal, burdeonsome, and adds to confusion.

Class Participation

There will be a LOT going on in this class. Most assignments are designed to be completed mostly in-class time. The class is highly collaborative, meaning you need to be present to do it.

Accomodations will be made for serious illness or other extenuating circumstances. However, it is the student's responsibility to stay caught up with course materials - and missing in-class activities will result in a decreased participation grade.

So please, don't make it part of your plan to miss class!

Academic Honesty

This course is governed by MUN's regulations on academic misconduct.

Course Schedule

Week 1: Data and Software in Fisheries - Sept 11/12

Week 2: Introduction to R - Sept 18/19

Week 3: Intro to Tidy Data - Sept 25/26

Week 4: Visual Display of Data 1 - Oct 2/3

Week 5: Visual Display of Data 2 - Oct 11/12

Week 6: Working with Messy Data - Oct 16/17

Week 7: Collecting and Managing Tidy Data - Oct 23/24

Week 8: Visual Display of Data 3 - Oct 30/31

Week 9: Figures for Science Communication - Nov 6/7

Week 10: Maps - Nov 14/17

Week 11: Markdown, Lab Notebooks, and Advanced RStudio Workflow - Nov 20/21

Week 12: Interactive Plots - Nov 27/28

Assignments and Grading

See assignment guide for details.

10% Participation

20% Explo**R**ation - partner presentation

20% Minor Assignments

50% Major Assignment