Dissertation Proposal

Allison Theobold

October 5, 2018

Introduction & Motivation

With society's increased computational power and the volume and variety of data, the practice of environmental science is changing. These changes have made computational abilities, such as data management, data visualization, and data analysis vital to researchers. However, in this changing climate the curricula in environmental science remains stagnant. A 2012 survey of graduate students in environmental sciences is eye opening, over 80% of students reported that they had received no formal training in computing, even at the most basic level, and 74% stated that they had no skills in any programming language (Hernandez, Meyernik, Murphy-Mariscal, & Allen, 2012).

Select undergraduate programs across the nation have begun to incorporate CS courses for non-majors into their life science curriculum, yet graduate programs have yet to see such an inclusion of computing into the curriculum. Meanwhile, in the environmental sciences, Statistics preparation is considered fundamental. Statistics courses have been readily incorporated into graduate environmental science programs, with the implicit assumption that these courses would also introduce students to the computational skills necessary for their research.

Statistics educators across the nation have called for a re-evaluation of the Statistics curriculum to incorporate computing, stating that the curriculum as it stands does not prepare students for the modern practice of Statistics (Cobb, 2015; Lumley, 2001; Hampton et al., 2017; Nolan & Temple Lang, 2010). All arguments considered, there are barriers to incorporating computing that many institutions fall victim to (Hampton et al., 2017). The question is thus, if these students are not being taught these vital computational skills needed for their research, where are they acquiring them?

Most, if not all, of what environmental science researchers know about data management, visualization, and analysis has been learned piecemeal, through the lab, or not learned at all (T. Teal et al., 2015). This communal process of acquiring the computation knowledge necessary for environmental science research lends itself well to the theoretical perspective of distributed cognition. Distributed cognition, like any cognitive theory, investigates the organization of cognitive systems. Unlike traditional cognitive theory, distributed cognition (1) considers the cognitive system beyond the individual, (2) includes interactions between individuals, and (3) incorporates resources and materials into an individual's environment (Hollan, Hutchins, & Kirsh, 2000). While social constructivism attends to the knowledge created through interactions between individuals, it does not attend to the tools and resources available to the individual in the acquisition of knowledge.

The theory of distributed cognition allows for us to seek to understand the relationships between elements in an individual's cognitive process. The literature on computational knowledge acquisition in the environmental sciences notes how an individual distributes the cognitive processes of learning computing to their social environment. Additionally, the literature acknowledges the variety of resources available to students when seeking to acquire computational knowledge (Eglen, 2009; Hampton et al., 2017; Sedgewich & Wayne, 2008, 2015), but does not embrace the coordination required of an individual when balancing internal (social) and external (resource) materials.

This study acknowledges each of these environmental influences, aiming to understand how individuals coordinate their environments to acquire the computational skills necessary to perform their research. Additionally, I plan to investigate how an additional resource, workshops targeted towards key computational skills, influence individuals in attaining these necessary skills.

Research Questions

To address the aims outlined above, we propose three research questions:

- What computing skills are necessary for environmental science graduate students to successfully implement applications of statistical computing in their research?
- How are students filling the gap between the computing skills they know and the computing skills they need to know, in order to perform applications or statistics in their research?
- How can workshops help to alleviate this gap between statistical computing knowledge and expectations?

Methodology

Workshops

This research proposes to situate itself within the classroom design study methodology. This methodology is appropriate as it would be difficult to authentically observe these students acquiring the computational skills necessary for their research. Additionally, the current research on the computational skills necessary for environmental science research is inadequate on specifying what subject matter is relevant for these disciplines, as well as what instructional methods are best suited for teaching these skills.

The interpretive framework that I am selecting focuses a student's learning based on the supports of a workshop tasks and tools, the nature of computational norms within each student's field, and the quality of a workshop's discourse around a computational skill. This focus is reinforced by the body of educational literature which suggests that student's reasoning is not an independent process, but is instead shaped by the settings of their learning, and by the collective practices they participate in while learning (Hall, 2001; Hoyles, Noss, & Pozzi, 2001).

I will use the computational learning goals suggested by Hampton et al. (2017). These goals state that for data-intensive research in the environmental sciences, researchers should be able to

- work with messy or organized datasets, stored in varied data formats, in a reproducible workflow,
- employ a variety of statistical methods and simulation,
- make use of basic software skills, and
- create effective data visualizations (Hampton et al., 2017, p. 547).

Each arm of these computational abilities coincides with workshop currently implemented or in development to be administered through partnership with the Library. The current Introduction to `R` workshop gets researchers started on working with `R`, exploring the basics of: structure and organization of datasets, data summary techniques, and data visualization techniques. The current Intermediate `R` workshop guides researchers through the use of: conditional and relational statements for working with data, and looping and user-defined functions for repeated processes.

Based on the pilot study in Spring 2017 and the faculty interviews from Spring 2018, I created the starting points for computational abilities and computational expectations of environmental science researchers. In the interviews with graduate environmental science students in Stat 512, where they reasoned through applications of statistical computing, it became clear that these students had misconceptions of fundamental concepts for working with data in `R`. In the interviews with environmental science faculty, I questioned them about what computational skills they believe are necessary for students to successfully perform research in the environmental sciences. As expected, answers varied between fields of research, however, many faculty emphasized students' ability to write their own functions, use conditional statements to process their data, looping and/or vectorization to automate their processes, database storage, and data manipulation.

The theoretical framework that I base this study on is the necessity for teaching statistical computing concepts in the context of data. Teaching the concepts in context provides students with authentic experiences that help them to successfully apply these concepts to their future computational endeavors [dunlang, p. 98]. In this setting students are able to grapple with the frustration of solving computing applications, with the support of instructors, where their creativity is applauded.

Longitudinal Interviews

Interviews from the cohort will be situated within the theoretical perspective of distributed cognition. As I am seeking to understand what computational skills are necessary for research in environmental science fields, I am assuming students are acquiring these skills through use of their social environment (as seen in the pilot study) and utilization of resources, such as coursework and online or written tools. This study does not attend to the structure of these resources and how students interact with them, instead focusing on what the critical components of these environments are and under what situations they use them.

Data Collection

Workshops

The day prior to each workshop, the students, faculty, and staff that are registered are sent an email. This email requests that they complete the workshop pre-survey (Appendix), so that I am able to gauge everyone's backgrounds. During the workshop, applications of the computational concepts are given to participants for them to work through. Participants are requested to submit their answer to each of these questions during the workshop (through a Google Form). The day after each workshop, I send an email with any necessary information (e.g. `R` code from applications we did not get to), and a request for participants to complete a workshop post-survey (Appendix).

Longitudinal Students

Interviews of the recruited cohort of environmental science students are aimed to take place twice a semester, namely early and late in the semester. Approximately one week prior to each interview, participants will be requested to submit their most recent code for their research. During the interview, participants will be questioned about the computational methods identified in their code, as to where they learned these computational methods.

Computational Courses

Observations and content analysis of courses taken by a large proportion of environmental science graduate students, which teach aspects of computing, will be carried out (e.g. Stat 511 & 512, PSPP 516, WILD 401 & 501). Materials from these courses will be solicited from the instructor or pulled from course websites.

Data Analysis

Workshops

The pre-workshop surveys can be used when assessing the contingent aspects of each workshop. These surveys contain important background information that will aid in identifying what aspects of the workshop environment were potentially due to participants' backgrounds. Additionally, responses to "What do you hope to learn from this workshop?" can be used to inform the computational skills necessary for performing environmental science research. Responses to "Why did you choose to come to this workshop?" help to delineate participants seeking computational skills for their coursework from those seeking skills for their research. Lastly, “What resources have you used while learning to program in R?” helps to reinforce and identify the dominant resources participants are utilizing when acquiring computational skills.

The submissions for each application of computational concepts helps to outline any misunderstandings students are having when independently coding. These misunderstandings can be used to inform restructuring of the workshop, to facilitate more concrete understandings of these concepts prior to the applications. Finally, the post-workshop survey will provide guidance on both the environment of the workshop (number of facilitators, structure and location of workshop) and the structure of the workshop (material presented in an understandable way, confidence in implementation in research). Directly asking participant “What did you get out of this workshop?” highlights the aspects of the workshop that were successful for each participant, which can be used to note any aspects that could use improvement. Lastly, asking participants “What (content) changes would you recommend for this workshop?” reinforces the computational skills these participants heave deemed necessary for their research.

Longitudinal Students

This study seeks to investigate the evolving nature of users’ computational tasks, their history of computational tasks, and their relationship with other information sharing resources (peers, consultants, advisers, courses, online resources, etc.). The historical comparison of computational abilities over time provides rich information about evolving nature of these skills, and the resources students employ throughout their research experiences.

At every interview, I will have the ability to track each student’s research code. The first submission of research code will be inspected to identify the computational skills demonstrated by each student. These will be identified as skills pertaining to each of the four learning goals outlined in the workshop learning goals (working with data of a variety of formats, employing a variety of statistical methods, basic software skills, and effective data visualizations). Each subsequent submission of research code will be inspected to identify any new computational skills employed. During the interview, participants will be asked where they learned each of these computational skills. The interviews will be transcribed and open-coded to feature the dominant resources these students are employing in their acquisition of the computational skills necessary for their research.

Computational Courses

Course observations will take place at the consent of the instructor, and are intended to identify the culture of the classroom and the classroom discourse surrounding the learning of computational skills. The analysis of each course’s materials will be used to compare the computational understandings taught across different course instructors (Stat 511), and to outline what computational concepts are taught and the context in which they are taught.

Study Aims

The aims of this study is to develop publicly available resources for data-intensive research in the environmental sciences. Developing introductory and intermediate workshops with low barriers to entry, is of the utmost importance for the computational expectations of today’s environmental science researchers. The dearth of resources available for learning skills for statistical computing assume that researchers posses a basic understanding software skills, which many researcher do not. The workshops created for this study aid in constructing these vital understandings of software skills, so that a broader array of resources are accessible to researchers.

Outlining the key computational skills necessary for successful implementation of statistics to research in environmental science fields, helps to guide Statistics and environmental science faculty in better integrating these skills into the curriculum. Materials for the workshops, which address these key computational skills can be used by faculty, to pool resources for teaching computing, so that the materials are available in formats that can be quickly adapted and customized.

Timeline of Research Study

Semester	Project(s)
Spring 2017	• environmental science graduate students in Statistics 512 were interviewed
Spring 2018	• Introduction to R & Intermediate R were advertised broadly to graduate environmental science students, through partnership with Library, with pre- and post-surveys administered • environmental science faculty were interviewed about computational expectations • environmental science graduate students in Statistics 511 were recruited for longitudinal study
Fall 2018	• Introduction to R & Intermediate R revised and advertised broadly to graduate environmental science students, through partnership with Library • First interviews with longitudinal environmental science graduate cohort • Recruit first year environmental science graduate students through workshop participation • Perform PSPP 516 classroom observations and course material analysis • Submit manuscript on computation knowledge acquisition strategies to SERJ
Spring 2019	• Revise and administer Introduction to R & Intermediate R workshops • Create and administer Data Visualization and Data Wrangling workshops, through partnership with Library • Continuing interviews with longitudinal environmental science graduate cohort • Perform WILD 401 and 501 observations and course material analysis • Begin draft of manuscript documenting gaps between computational abilities and expectations of graduate students in environmental science
Summer 2019	• Complete manuscript of gaps between computational abilities and expectations • Analyze research code provided by longitudinal cohort • Analyze Stat 511 and 512 materials from previous instructors
Fall 2019	• Revise and administer Introduction to R, Intermediate R, Data Visualization, and Data Wrangling workshops, through partnership with Library • Continuing interviews with longitudinal environmental science graduate cohort • Perform Stat 511 and 512 observations and course materials analysis • Draft paper on computational skills necessary for research in the environmental sciences
Spring 2020	• Finalize paper on computational skills necessary • Draft manuscript on impact of computing workshops • Defend dissertation

References

Cobb, G. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, *69*(4), 266–282.

Eglen, S. (2009). A quick guide to teaching r programming to computational biology students. *PLoS Computational Biology*, *5*(8), 1–4.

Hall, R. (2001). Schedules of practical work for the analysis of case studies of learning and development. *Journal of the Learning Sciences*, *10*(1-2), 201–222.

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., … Aukema, J. E. (2017). Skills and knowledge for data-intensive environmental research. *BioScience*, *67*(6), 546–557.

Hernandez, R. R., Meyernik, M. S., Murphy-Mariscal, M. L., & Allen, M. F. (2012). Advanced technologies and data management practives in environmenal science: Lessons from academia. *BioScience*, *62*(12), 1067–1076.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174–196.

Hoyles, C., Noss, R., & Pozzi, S. (2001). Proportional reasoning in nursing practice. *Journal for Research in Mathematics Education*, *32*(1), 4–27.

Nolan, D., & Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, *64*(2), 97–107.

Sedgewich, R., & Wayne, K. (2008). *Introduction to programming in java*. Addison Wesley.

Sedgewich, R., & Wayne, K. (2015). *Introduction to programming in python*. Addison Wesley.

Teal, T., Cranston, K., Lapp, H., White, E., Wilson, G., Ram, K., & Pawlik, A. (2015). Data carpentry: Workshops to increase data literacy for researchers. *International Journal of Digital Curation*, *10*(1), 135–143.