Supporting Data-Intensive Environmental Science Research: Data Science Skills for Scientific Practitioners of Statistics

Allison Theobold

Abstract

The importance of data science skills for modern environmental science research cannot be understated, but graduate students in these fields typically lack these integral skills. Yet, over the last 20 years statistics preparation in these fields has grown to be considered vital, and statistics coursework has been readily incorporated into graduate programs. As "data science" is the study of extracting value from data, the field shares a great deal of conceptual overlap with the field of Statistics. Thus, many environmental science degree programs expect students to acquire these data science skills in an applied statistics course. A gap exists, however, between the data science skills required for students' participation in the entire data analysis cycle as applied to independent research, and those taught in statistics service courses. Over the last ten years, environmental science and statistics educators have outlined the shape of the data science skills specific to research in their respective disciplines. Disappointingly, however, both sides of these conversations have ignored the area at the intersection of these fields, specifically the data science skills necessary for environmental

science *practitioners* of statistics.

This research focuses on describing the nature of environmental science graduate students' need for data science skills when engaging in the data analysis cycle, through the voice of the students. In this work, we present three qualitative studies, each investigating a different aspect of this need. First, we present a study describing environmental science students' experiences acquiring the computing skills necessary to implement statistics in their research. In-depth interviews revealed three themes in these students' paths toward computational knowledge acquisition: use of peer support, seeking out a "singular consultant," and learning through independent research. Motivated by the need for extracurricular opportunities for acquiring data science skills, next we describe research investigating the design and implementation of a suite of data science workshops for environmental science graduate students. These workshops fill a critical hole in the environmental science and statistics curricula, providing students with the skills necessary to retrieve, view, wrangle, visualize, and analyze their data. Finally, we conclude with research that works toward identifying key data science skills necessary for environmental science graduate students as they engage in the data analysis cycle.