



Working with Big Data in RStudio

My experiences using RStudio Software for Research

Brock Tausan **April 27, 2021**

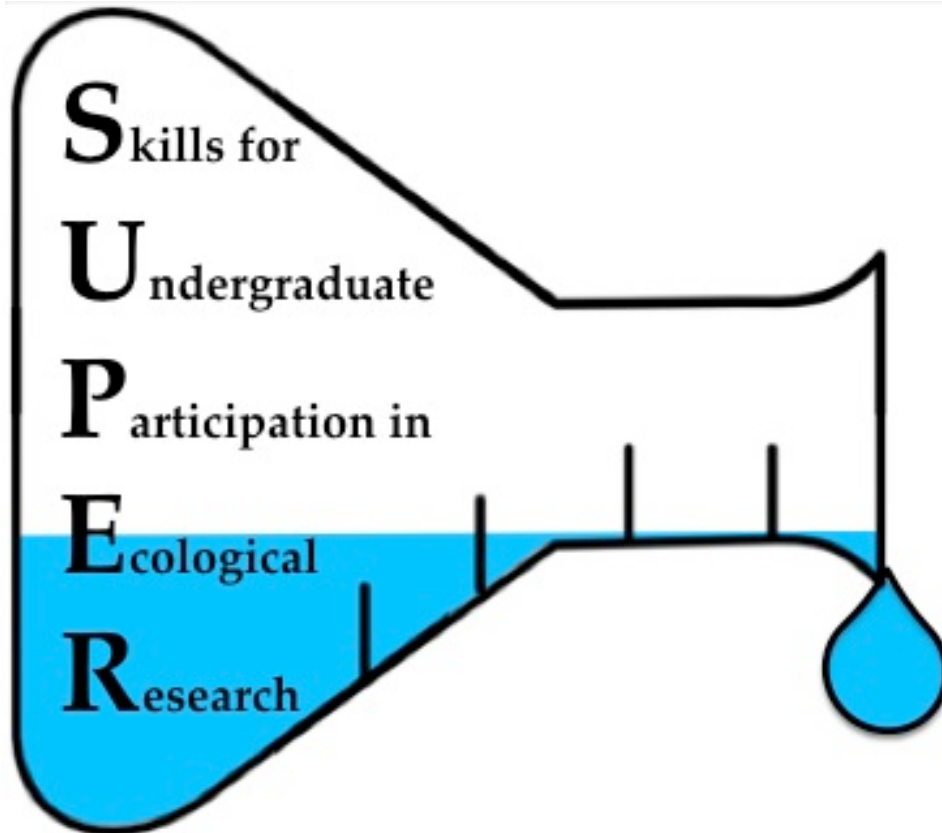
The Research

The North Fork of the Crow River watershed located northeast of Minneapolis, Minnesota was the area of study for my research project titled "Drivers of Chlorophyll-a in Lakes Across the Crow River Watershed of Minnesota".

Under the Skills for Undergraduate Participation in Ecological Research (SUPER) program, I conducted an analysis of the the Chlorophyll-a dynamics over this large region using in-situ water quality, hydrologic, and land use data. Data for this research project was accessed through the Lake Multi-Scaled Geospatial and Temporal Database (LAGOS-NE) using R.

In my research I found that Total Kjeldahl Nitrogen and Nitrite+Nitrate were the primary drivers of Chlorophyll-a, and could

be used to model Chlorophyll with and R-squared of 0.63.



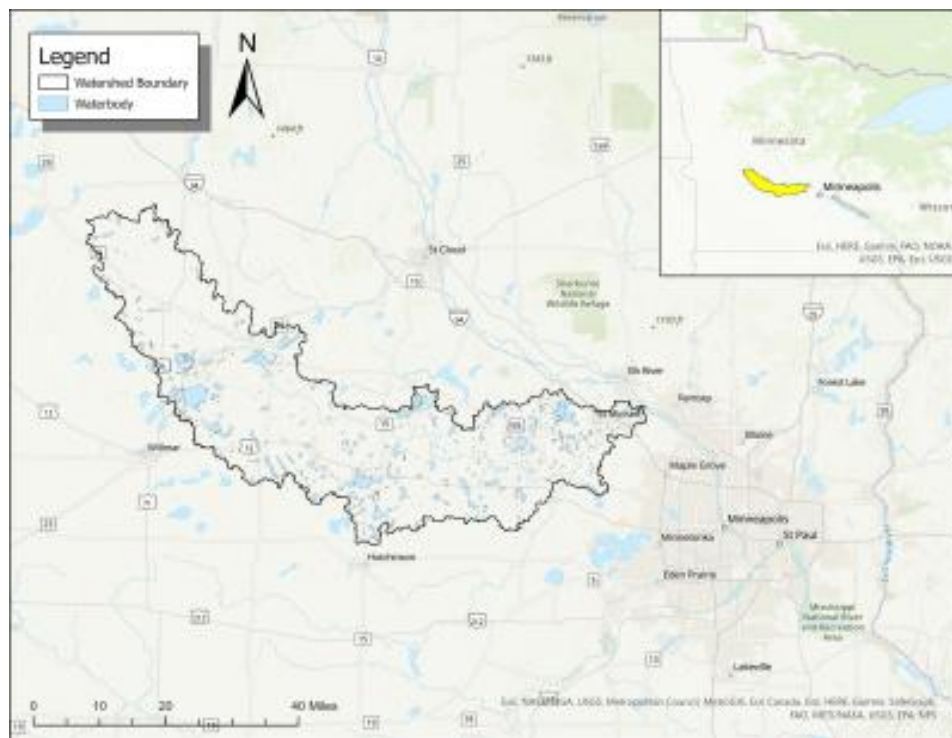
**NATURAL
RESOURCE
ECOLOGY
LABORATORY**



**Colorado
State**
University

**WARNER COLLEGE OF
NATURAL RESOURCES**

*Department of Ecosystem
Science and Sustainability*



Big Data

Large data sets such as LAGOS-NE offer major advantages when attempting to learn about a regions hydrologic and ecologic dynamics.

When working with a large amount of data, it is very rare for the data to be arranged in a format that is useable to you from the moment you obtain it. Here are some things that I learned in my own process.

- Make sure to pick an area that has an abundant amount of observations for your variable of interest. If your variable of interest has a small amount of observations, then it doesn't matter how much other data you have to explain it.
- Blank spaces in the data are impossible to avoid, but try to avoid having too many. If it is possible, take the daily, monthly, or annual mean values to get rid of blank observations. If you cannot do this, it might be best to leave out that variable.

- Make sure to pay attention to how data were collected and what units the variable uses. Metadata is essential knowledge down the road in a report.
- Use your literature review to inform your decisions on what variables to use. With so much data to choose from, it can be overwhelming to try to pick what to use based on your own judgement
- Do not underestimate the amount of time it will take to get to know your data. Exploratory data analysis takes time and creativity. Utilizing different strategies of data visualization is an essential tool I had to use to feel comfortable with my data



Using RStudio


RStudio statistical software is a great way to work with large amounts of data, and it has a tool for just about anything you can think of. There is a pretty steep learning curve when getting started, and being resourceful is essential to your own success. Each package in R has a similar but slightly different language, so it takes time to learn. Google and youtube are your best friends whenever you run into an issue. Even after my research project, I would say I'm barely past a beginner level understanding so don't expect to know it all right away. Here are some of my own recommendations for learning how to use R effectively.

- Utilize the internet for problem solving
- Do all of the R primers offered by RStudio Cloud
- Give yourself breaks ~ it can be hard to stare at lines of code for too long
- Don't limit yourself to the base packages, use packages such as tidyverse that can make your life much easier
- Try to be creative and try different things. There is not a single set way to get things accomplished so experiment with different lines of code

- Keep your code chunks organized, and comment on the code itself so that if you revisit the code you can quickly remember your thought process

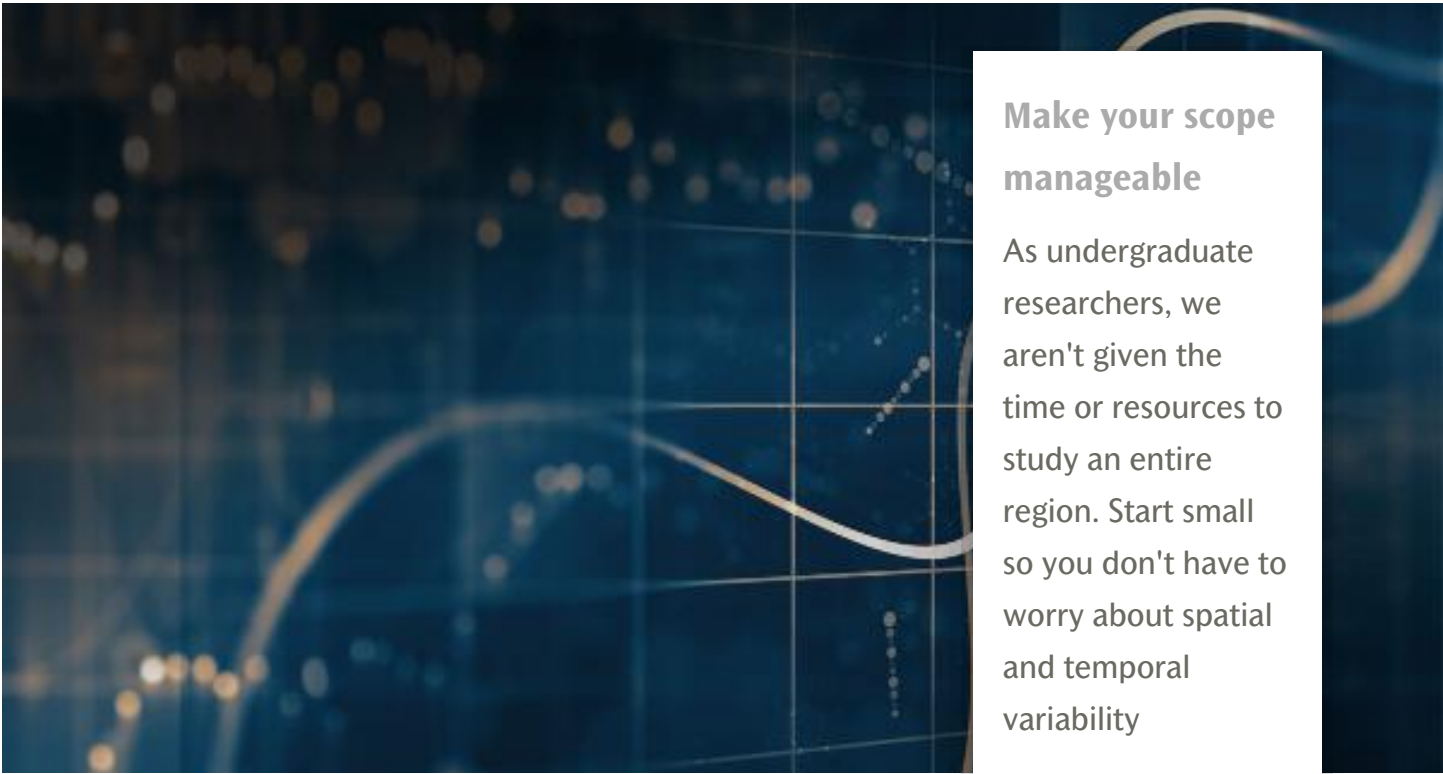


Key Points



Build your study around good data

If you have the option, you should wait to commit to a site until you have identified that there is an adequate amount of data available to you. Try selecting several sites and conduct an exploratory data analysis.



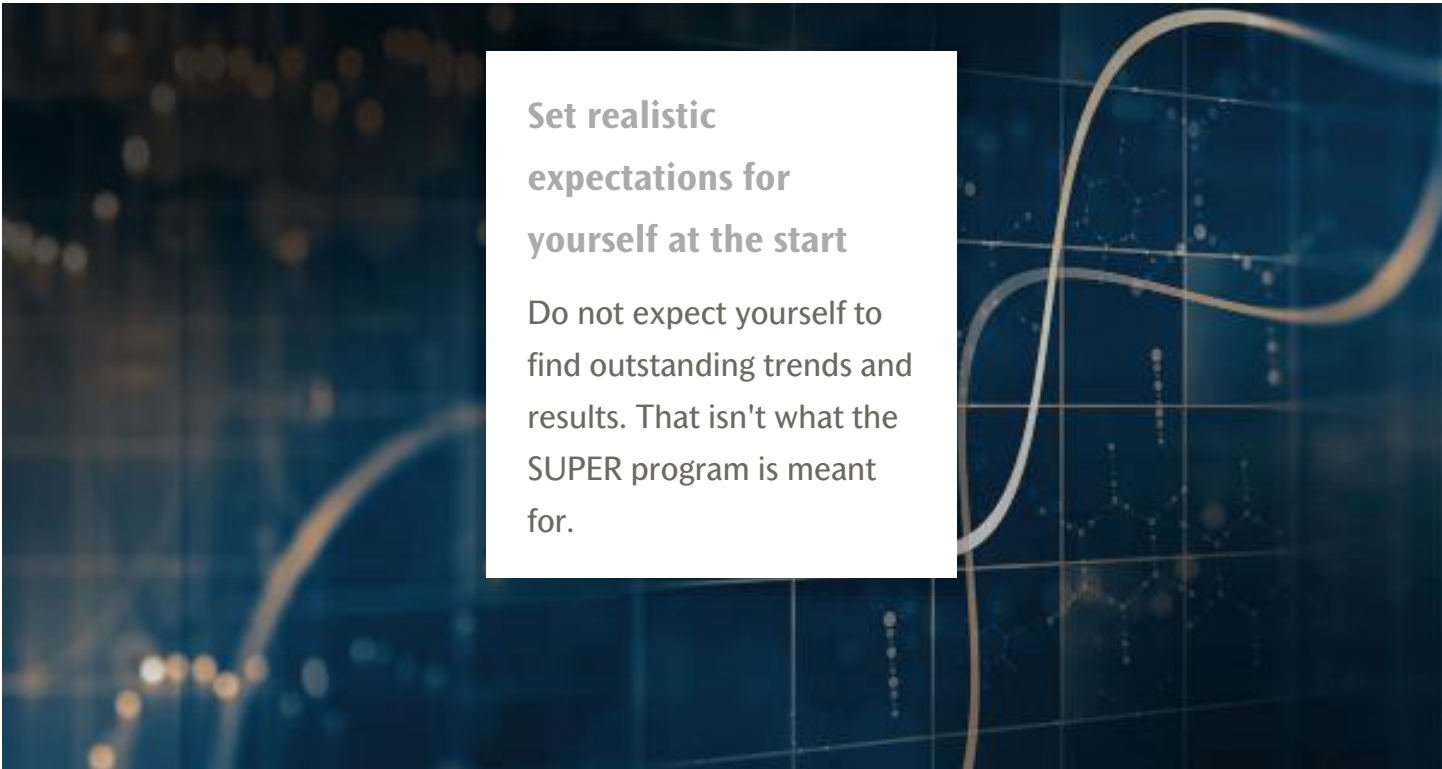
Make your scope manageable

As undergraduate researchers, we aren't given the time or resources to study an entire region. Start small so you don't have to worry about spatial and temporal variability



Ask lots of questions

Your teachers and mentors have a lot to offer, and if they don't have an answer then you can both work through it together. Don't ever feel like you are alone in research.



**Set realistic
expectations for
yourself at the start**

Do not expect yourself to find outstanding trends and results. That isn't what the SUPER program is meant for.

This project was completed as part of the Ecosystem Science and Sustainability SUPER Program (Skills for Undergraduate Participation in Ecological Research)

Instructors: Dr. Stacy Lynn and Anna Clare Monlezun

Communication Piece ESS 221

Powered by ArcGIS StoryMaps